# Expansion of the genome editing toolbox of *Clostridium autoethanogenum*

By François Marc-Antoine Seys, MSc

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy.

December 2021

University of Nottingham
UK | CHINA | MALAYSIA

# Declaration

Unless otherwise acknowledged, the work presented in this thesis is my own. No part has been submitted for another degree at any institute of learning.

François Marc-Antoine Seys.

December 2021

# Abstract

Current environmental crises – from rising global temperatures to biodiversity collapse – have created an urgent need for new, circular and sustainable ways of producing and recycling materials and energy independently of fossil resources. Gas fermentation, by consuming greenhouse gases to produce fuels and chemicals, offers human society an attractive route to sustainability. This thesis describes the development of genome editing tools to facilitate the engineering of the gas fermenting *Clostridium autoethanogenum*.

Initial studies illustrated, for the first time, the concept of genomic Bookmarks, which streamline complementation studies in any organism compatible with Cas9-mediated homology-directed mutagenesis. An array of nine of these 24 nt sequences were inserted in the genome of *C. autoethanogenum* in place of the *pyrE* gene. Each was subsequently targeted in a second round of Cas9-mediated homology-directed mutagenesis to restore the *pyrE* locus, generating a perfect complemented strain.

Then, a mysterious drop in the efficiency of gene transfer was subjected to an in-depth analysis. The absence of mutations uncovered by whole genome sequencing suggested that epigenetic changes likely affect the conjugation efficiency of donor and/or recipient cells. By inoculating the donor strain directly from the transformation plate and significantly reducing the number of donor cells, improved conjugation efficiencies were reliably restored.

Finally, Target-AID, a state-of-the-art genome editing tool that creates premature STOP codons by substituting selected C nucleotides with T, was exemplified in *C. autoethanogenum*. While its initial application to the *pyrE* locus was inconclusive, its application to knock out three genes encoding distinct alcohol dehydrogenases was successful. Up to two targeted mutations could be carried out simultaneously, although three should be possible. In the process, the targeting space of Target-AID was expanded to encompass NG and NAA protospacer adjacent motives (PAM) and showed that an array of multiple sgRNA expression cassettes was a better multiplexing strategy than a single expression cassette processed by intervening tRNAs or direct repeats (DR) from the native *Streptococcus pyogenes* clustered regularly interspersed repeat (CRISPR) system.

The research presented in this thesis provides valuable tools and insights which will help engineering *C. autoethanogenum* and many other organisms. Among other biotechnological applications, this research could thus facilitate the emergence of industrial processes with a lower environmental footprint.

# Acknowledgement

This work was funded by the BBSRC/EPSRC Synthetic Biology Research Centre and would not have been possible without the advice and support of my supervisors, Prof. Nigel P. Minton, Dr. Christopher M. Humphreys and Dr. Edward L. Bolt. In particular, I would like to thank Nigel and Christopher for their patience and the many opportunities they helped me seize throughout the duration of this thesis. I would also like to thank Nigel, Prof. Chris Denning, the staff of the University of Nottingham, the Biodiscovery Institute and particularly of the Clostridium Research Group for moving mountains to allow us back safely into the lab throughout the Covid-19 pandemic.

Besides the core research work presented in this report, I had the chance of participate in many extra-curricular activities ranging from the participation to conferences to student competitions and popularization. These would not have happened without the help and encouragements of Louise Dyne, Jacqueline Minton, Loretta Waddon, Paul Brett and Tracy Hassall-Jones.

On top of my formal collaborations with Claudio Tomi Andrino, Dr. Thomas Millat and Dr. Peter Rowe for specific parts of my research, I would also like to express my gratitude towards the rest of my colleagues in no particular order for their daily interactions and exchange of ideas, which were sorely missed during the last stages of this thesis as the Covid-19 pandemic unfolded: Claudio, Thomas, Amaury Montarnal, Margaux Poulalier-Delavelle, Dr. James Millard, Dr. Andrew Dempster, Dr. Bunmi Banire Omorotionmwan, Alex Agius, Dr. Christian Gude, Marcel Te Vrugt, Elizabeth Redfern, Dr. Craig Woods, Dr. Terry Bilverstone, Dr. Jonathan Humphreys, Dr. Robert Mansfield, Ruud Hendriks, Dr. Maria Zygouropoulou, Dr. Raquel Rodrigues, Dr. Daphne Groothuis.

I would like to thank my fiancée, Athina Liaskoni, for her strength and unwavering solidarity. You have been an inspiration! I am also grateful for the stimulating conversations with my family and my close friend Tim Blokker throughout these formative years.

This thesis is dedicated to the 38 deaths, the one disappearance, and the tens of thousands of people affected by the dramatic floods of summer 2021 in Belgium, which were attributed to the preventable rise in global average temperature.

## Covid-19 impact statement

On the 23rd of March 2020, roughly 2 years and 9 months into my research project, the Prime Minister of the United Kingdom announced the first national lockdown to curb the spread of SARS-Cov-2 virus, responsible for the Covid-19 pandemic. The lockdown ended on the 23rd of June, allowing people to come back to work on the condition that they follow stringent public health measures. For my laboratory, this meant that no wet-lab experiment could be done for the three months of lockdown, all expenses became strictly controlled, and we had to operate at less than half-capacity to respect social distancing guidelines. A shift schedule was put in place and maintained until the end of my project: we could only be at work for seven-and-a-half hours per day, from 6h30 to 14h or from 14h30 to 22h. Work was constantly interrupted to sanitize pieces of equipment or ensure social distancing guidelines were respected. This shift schedule was rotated each week, which wrecked havocs on my sleep pattern and the one of my colleagues. Similar measures were adopted throughout the planet, disrupting global and local supply chains and creating delays in the delivery of critical consumables. New Covid-19 testing facilities and Covid-19 research initiatives exploded the demand for laboratory consumables and molecular biology services, and they became unreliable as a result.

These exceptional circumstances had an undeniable impact on my thesis. Counting everything together, I easily lost at least 6 months of experiments in the final year of my funding while enduring a great amount of stress. My research plan had to be changed drastically, taking each remaining week into account to decide which experiments should be prioritized, which ones could be grouped together and done in parallel, and which ones should simply be abandoned. Like everybody else, these disruptions also affected my personal life, and I am grateful to my fiancée, my friends, and my family for their support during these trying times.

# Table of Content

# List of Figures

# List of Tables

# Abbreviations

| Acronym | Meaning |
|---|---|
| **2,3BDH** | NADH dependent butanediol dehydrogenase |
| **aa** | Amino acid |
| **Ack** | Acetate kinase |
| **ACS** | Acetyl-CoA synthase |
| **ADH** | Alcohol dehydrogenase |
| **Adh** | Alcohol dehydorgenase |
| **A-EJ** | Alternative End Joining |
| **AID** | Activation-induced deaminase |
| **Ald** | Aldehyde dehydrogenase |
| **Aldc** | Acetolactate decarboxylase |
| **Als** | Acetolactate synthase |
| **Alt-NHEJ** | Alternative Non-homologous End Joining |
| **AOR** | Aldehyde oxydoreductase |
| **AP** | Apurinic/Apyrimidic |
| **ATP synthase** | ATP synthase |
| **BECCS** | Bioenergy with Carbon Capture and Storage |
| **BER** | Base excision repair |
| **BIR** | Break-induced replication |
| **BM** | Bookmark |
| **BMa** | Bookmark array |
| **Cam** | Chloramphenicol |
| **Cas** | CRISPR associated protein |
| **CASS** | CRISPR-Cas system |
| **CCU** | Carbon Capture and Utilisation |
| **CCUS** | Carbon Capture, Utilisation and Storage |
| **CDS** | Coding Sequence |
| **CFU** | Colony-forming units |
| **Chi** | Cross-over hotspot instigator |
| **CODH** | CO dehydrogenase |
| **CRISPR** | Clustered, regularly interspersed repeats |
| **crRNA** | CRISPR RNA |
| **DAC** | Direct Air Capture |
| **D-loop** | Displacement loop |
| **DNA** | Deoxyribonucleic acid |
| **DNA-PKcs** | DNA-dependent protein kinase catalytic subunit |
| **DSB** | Double-stranded break |
| **dsDNA** | Double-stranded DNA |
| **Em** | Erythromycin |
| **EMBOSS** | European Molecular Biology Open Software Suite |
| **FDH** | Formate dehydrogenase |
| **FOA** | 5-Fluoroorotic acid |
| **FTHFS** | Formyl-THF synthase |
| **gRNA** | Guide RNA |
| **GUIDE-seq** | Genome-wide, unbiased identification of DSBs enabled by sequencing |
| **HA** | homology arms |

| Acronym | Meaning |
|---|---|
| **HDGR** | Homology-directed gap repair |
| **HDR** | Homology-directed repair |
| **HDV** | hepatitis delta virus |
| **HGT** | horizontal gene transfer |
| **HH** | Hammerhead ribozyme |
| **HJ** | Holliday Junction |
| **HytA-E** | Electron bifurcating hydrogenase |
| **IDT** | Integrated DNA Technologies |
| **indel** | Insertion-deletion |
| **IPTG** | Isopropyl β-thiogalactopyranoside |
| **KLD** | Kinase, Ligase, DpnI |
| **LB** | Luria-Bertani broth/ lysogeny broth |
| **Lct** | Electron bifurcating lactate dehydrogenase |
| **Ldh** | Lactate dehydrogenase |
| **LVA-tag** | Leucine-Valine-Alanine Tag |
| **MET** | Methyltransferase |
| **MMEJ** | Micro-homology End joining |
| **MMR** | Mismatch repair |
| **MTHFC** | Methenyl-THF cyclohydrolase |
| **MTHFD** | Methylene-THF dehydrogenase |
| **MTHFR** | Methylene-THF reductase |
| **MTHFR(b)** | Ethylene-THF reductase |
| **nCas9** | Cas9 nickase |
| **NCBI** | National Center for Biotechnology Information |
| **NEB** | New England Biolabs |
| **Nfn** | NADH-dependent reduced ferredoxin:NADP + oxidoreductase |
| **NGS** | Next generation sequencing |
| **NHEJ** | Non-homologous End joining |
| **OD** | Optical density |
| **ODN** | Oligodeoxynucleotides |
| **PAM** | Protospacer adjacent motif |
| **PBS** | Phosphate-buffered saline |
| **PCR** | Polymerase Chain Reaction |
| **pegRNA** | Prime editing gRNA |
| **PFOR** | Pyruvate ferredoxin oxidoreductase |
| **Pol λ** | DNA polymerase lambda |
| **Pol μ** | DNA polymerase mu |
| **Pta** | Phosphate acetyltransferase |
| **R-loop** | RNA:DNA hybrid loops |
| **RM** | Restriction-Modification |
| **RNA** | Ribonucleic acid |
| **RNAP** | RNA polymerase |
| **RNAse** | RNA nuclease |
| **Rnf** | Rhodobacter Nitrogen Fixation complex |
| **RSKO** | Restriction site knock-out |
| **SD** | Standard deviation |
| **SDSA** | synthesis-dependent strand annealing |
| **sgDNA** | DNA coding for sgRNA |

| Acronym | Meaning |
|---|---|
| **sgRNA** | Single-guide RNA |
| **SOS** | Save our souls |
| **SpCas9** | Cas9 protein from Streptococcus pyogenes |
| **SRA** | Sequence Read Archive |
| **SSA** | Single-strand annealing |
| **SSB** | Single-stranded break |
| **ssDNA** | Single-stranded DNA |
| **SNP** | Single-nucleotide polymorphism |
| **Ta** | Annealing temperature |
| **TA** | Target-AID |
| **TAE** | Tris base, acetic acid, and EDTA |
| **TdT** | terminal deoxynucleotidyl transferase |
| **TLS** | Translesion DNA synthesis |
| **tracrRNA** | trans-activating CRISPR RNA |
| **trTA** | Truncated Target-AID |
| **trTarget-AID** | Truncated Target-AID |
| **UDG** | Uracil DNA glycosylase |
| **UGI** | Uracil glycosylase inhibitor |
| **UTR** | Untranslated region |
| **v/v, %** | Concentration calculated as a volume ratio and expressed as a percentage. |
| **WGS** | Whole-Genome Sequencing |
| **WT** | Wild-Type |
| **YTF** | Yeast tryptone fructose |

“ The nation behaves well if it treats the natural resources as assets which it must turn over to the next generation increased, and not impaired, in value. „

- Theodore Roosevelt, 1858-1919, 26th President of the United States (Republican, NY), Speech, Colorado Live Stock Association, Denver, Colorado, August 29, 1910.

# Chapter I: Introduction

## I.1. General context

Synthetic biology is a fascinating and quickly evolving discipline. It is easy to lose oneself into its inner workings, but we must remember that it is only a tool; a tool which must be wielded for a worthy purpose, within a realistic strategy. The primary objective of this thesis is clearly defined: to expand the genetic toolbox available for the engineering of *Costridium autoethanogenum*, a gas fermenting bacterium. However, its true purpose is much larger: to facilitate the emergence of a low-carbon, circular economy, which will secure a fair and prosperous future for humanity.

### I.1.1. We are living through a man-made climate breakdown

The summer of 2021 saw the death of more than 200 people in devastating floods in western Europe [1]. Floods also displaced more than a million people across China, causing at least 141 deaths or disappearances [2,3]. Simultaneously, all-time records of high temperature were broken across the world (cf. Figure 1), notably in northern America [4], causing widespread wildfires – despite the presence of an El Niña (*i.e.,* cold) global temperature pattern. More than 62,300 square miles of taiga have burned throughout Russia alone, while nearly 9000 square miles of forest burned in the United States, 13,000 in Canada, and a further 1508 square miles were collectively scorched among Turkey, Greece and Italy [5,6].

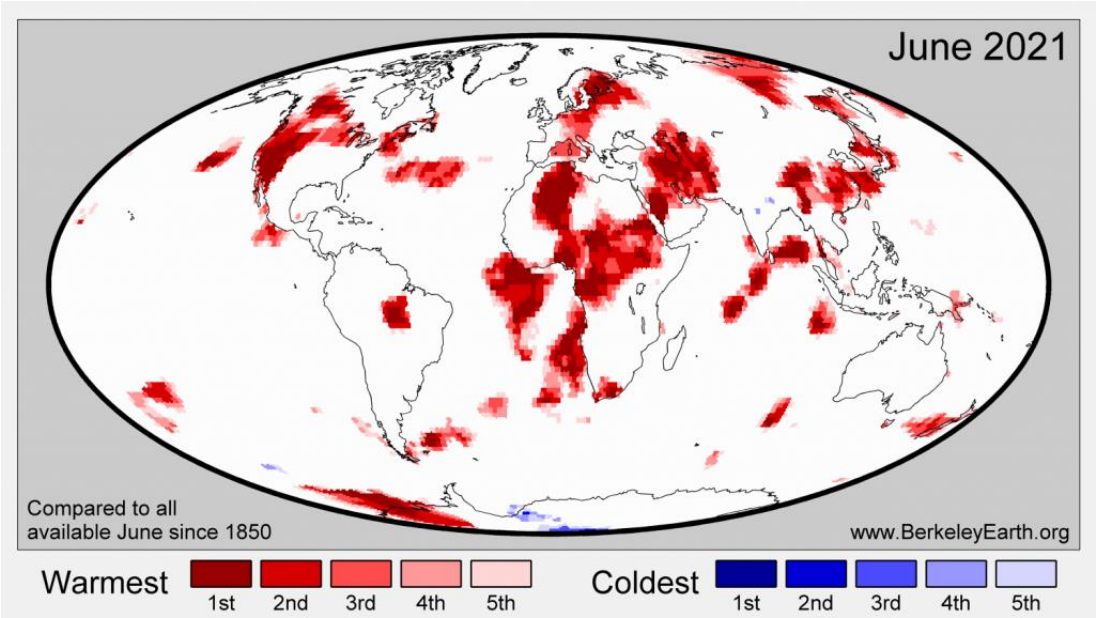



**Figure 1: Average temperature ranking of June 2021 compared to June temperature records since 1850.** Each coloured pixel represent a location on Earth where a local temperature record (warm or cold) has been broken. Notice the overwhelming absence of cold temperature records. Courtesy of Berkeley Earth [7].

According to the Intergovernmental Panel on Climate Change, such extreme weather events have become – and will become – more likely and more important as the global average temperature rises (cf. Figure 2) [8]. Indeed, 2021 is set to be among the seven warmest years ever recorded – the six other warmest years having all occurred over the six previous years [9,10]. If allowed to continue, this increase in global temperature threatens to push billions of people outside their ecological niche [11], causing widespread chaos and suffering [8,12].

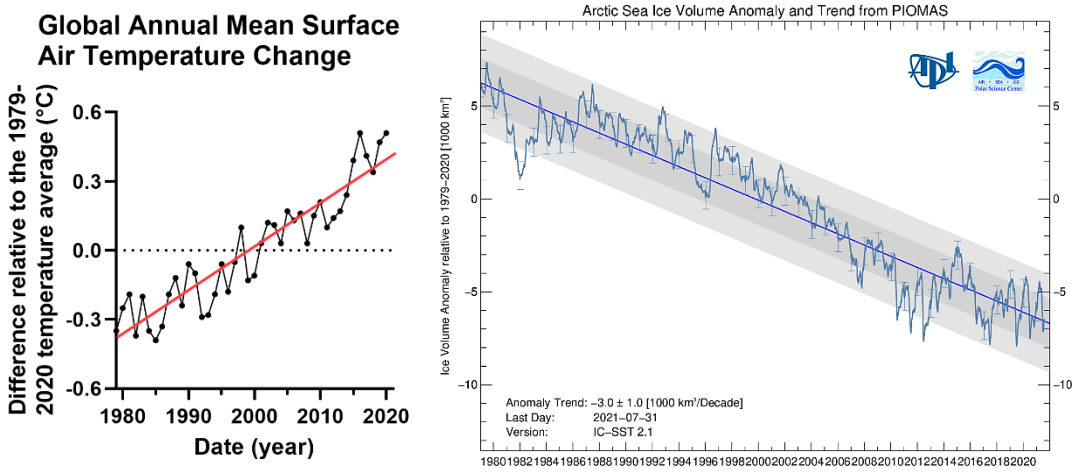



**Figure 2: Evolution of global mean average anomaly of (a) surface air temperature and (b) arctic sea Ice volume anomaly since 1979, compared to the 1979-2020 average.** Temperature data courtesy of NASA (GISS) [13]; arctic sea ice graph courtesy of Polar Research Center [14,15].

Rising global temperatures are caused by an increase in atmospheric concentration of man-made greenhouse gases (GHG) – notably $CO_2$, $CH_4$, and $N_2O$ – which results from the combustion of fossil resources such as coal, gas and oil [8]. Other anthropogenic gases do not directly absorb terrestrial infrareds, but can still increase global temperature through atmospheric chemistry. Carbon monoxide (CO), for example, reacts with atmospheric OH, which can increase the concentration of ozone ($O_3$) and indirectly increase the stability of $CH_4$. Consequently, it is thought that 100 tons of CO have the same indirect global warming potential as 5 tons of $CH_4$ [16], or about 100 to 1000 tons of $CO_2$ – depending on the models and timescales considered [17,18]. Furthermore, each molecule of atmospheric CO is eventually converted into $CO_2$; this specific contribution of CO to global warming is however usually accounted for as direct $CO_2$ emission to prevent double counting [18].

Despite extensive efforts for international cooperation and rapid technological progress since the First World Climate Conference in 1979 [19], global GHG emissions have continued to rise unabated year after year [20]. GHG concentrations still increased in 2020, when the global economy was reeling from the consequences of the Covid-19 pandemic [8,9].

According to the NGO Climate Action Tracker, under the current trajectory, the global temperature is predicted to rise from +1.2°C today to +2.9°C by the end of the century (compared to pre-industrial global temperature average) (cf. Figure 3) [21]. It is thus difficult to overstate the urgency of the need to reduce our GHG emissions.

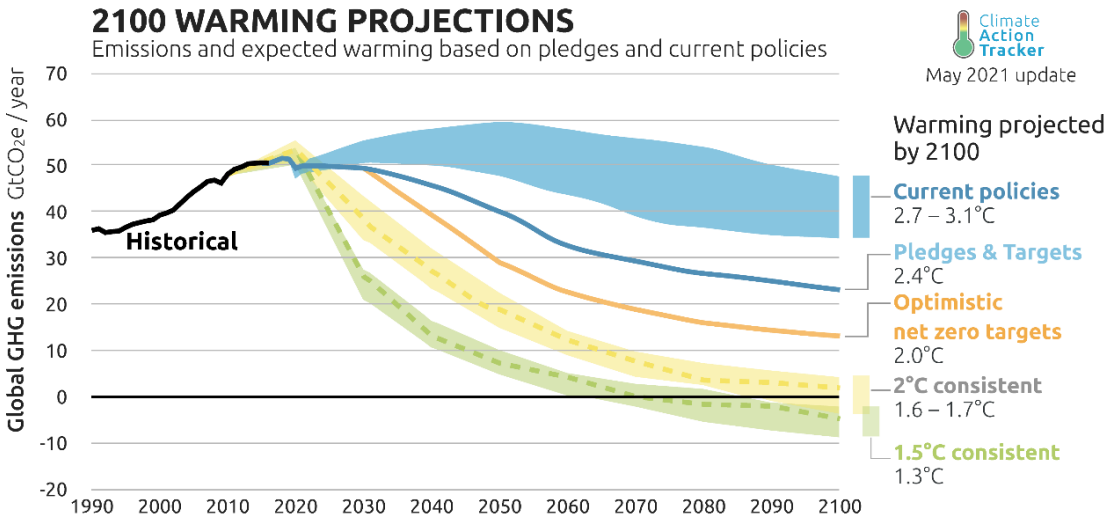



**Figure 3: Historical and projected global greenhouse gas emissions over time, with associated degrees of warming. Courtesy of Climate Action Tracker, 2021** [21]**.**

### I.1.2. Climate breakdown is just one facet of the global sustainability crisis

In 2009, Rockström and al., introduced the concept of planetary boundaries [22]. These are nine quantitative environmental parameters which, once past a certain threshold, might threaten the stability of the Earth system as modern humans have known it for the last 12,000 years [23]. While sometimes contested [24–29], it is a reasonably convenient framework to illustrate the scale and the complexity of the impact of humanity on the planet [30]. Within this framework, climate breakdown is just one of these planetary boundaries which have crossed the "safe" threshold and are now entering the zone of uncertainty. Other parameters, such as the ongoing biodiversity collapse [29,31–34] and the disturbance of the phosphorus and nitrogen flows [28,35–37], are well beyond the safe operating space of humanity. The other suggested planetary boundaries are stratospheric ozone depletion, atmospheric aerosol loading, ocean acidification, freshwater use, land-system change and "novel entities".

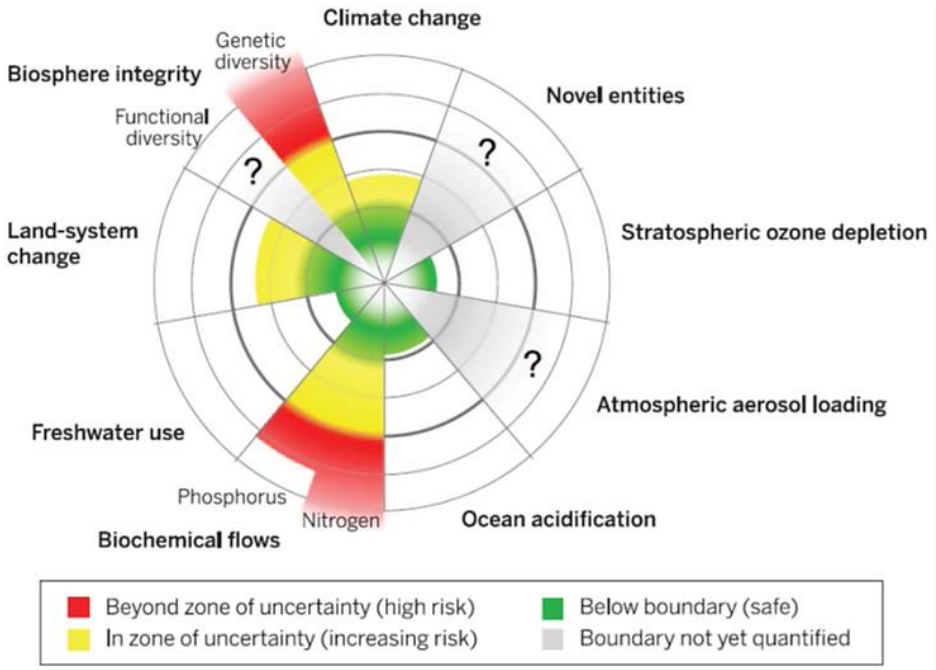



**Figure 4: Representation of planetary boundaries from Steffen et al., 2015** [23]**.**

One of the main advantages of the concept of planetary boundaries is that it formalises the links between these different aspects of environmental sustainability, and warns of the danger of focusing on one while neglecting the others [38–42]. Accordingly, reducing GHG emissions at the price of a further decline of biodiversity would not be a desirable outcome for humanity. A solution to any environmental problem needs to be sustainable across multiple dimensions.

### I.1.3. Dependence on fossil resources is a threat to energy security

Within the scope of this thesis, I am particularly interested in one more aspect of sustainability which is more political in nature: energy security. Indeed, the same fossil resources which warm the planet are also exposing many countries to unacceptable levels of economic and geopolitical risks [43–46], without even considering their environmental impact.

According to the International Energy Agency, fossil resources make up 81% of global primary energy supply [47]. They are also the primary ingredient of materials ranging from plastics to fertilizers and fine chemicals [48]. However, they are unequally distributed across the world; this creates conflicts [49,50] and provides a handful of actors which control the supply of fossil resources with disproportionate political and economic influence [51–53].

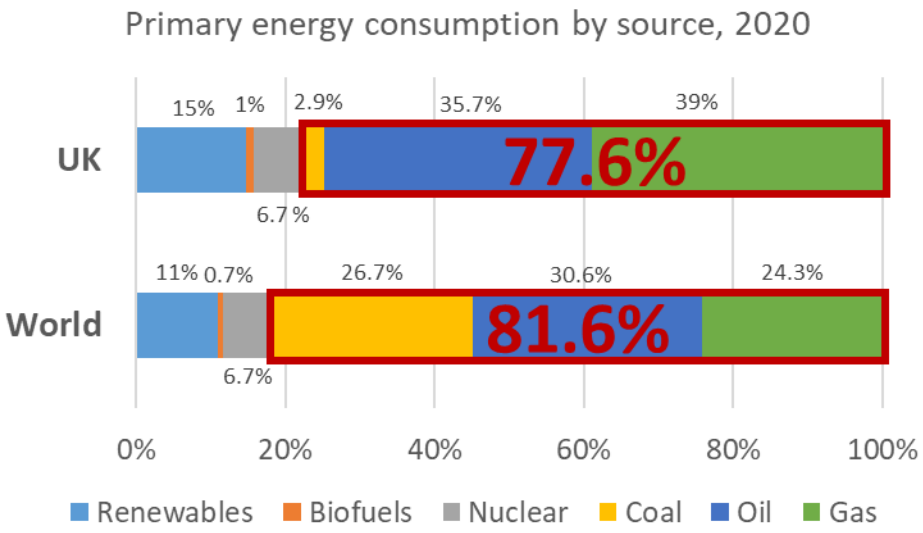



**Figure 5: Primary energy consumption by source in 2020.** Primary energy is shown based on the 'substitution' method which takes account of inefficiencies in energy production from fossil fuels. The share of fossil fuels is highlighted in red. Data from Our World in Data, BP p.l.c., Agora Energiewende and Ember, 2021 [54–56].

Although the UK is slightly less dependent on fossil fuels than the world's average, they still constitute roughly 78 % of its energy supply and consumption (cf. Figure 5) [54,57] and 41% of its electricity generation [54]. In 2019, net energy imports in the UK still accounted for 35% of energy consumption [57]. On top of the necessity of reducing GHG emissions, there is thus an enormous security incentive to find locally available alternatives to fossil resources, both as a source of energy and as a feedstock for the chemical industry.

### I.1.4. Industrial biotechnology as the basis of a decentralized, low-carbon, circular economy

Industrial biotechnology, or the craft of enlisting microbiology to produce valuable products, has the ability to alleviate many of our GHG emissions, environmental sustainability and energy security concerns. The incredible diversity of enzymes and microorganisms can indeed be exploited to deliver goods and services largely independently from fossil resources and with greatly reduced environmental impact [58]. When this natural diversity is not sufficient, synthetic biology can be used to tailor these organisms to better fit our needs. Biorefineries can be built almost anywhere and – contrarily to fossil resources – only depend on the availability of feedstocks which are widely accessible locally such as corn, agricultural waste, food waste, or even municipal solid waste and industrial off-gas. This stimulates local economies and improves their resilience and energy security [59,60]. It has already resulted in the production of fuels, plastics, drugs, food (including alternatives to meat), detergents, [61]…

Most commercial bioprocesses to-date exploit heterotrophic micro-organisms which get their energy and carbon from a variety of organic materials. Unfortunately, any large-scale exploitation of biomass often has adverse environmental or social consequences – even if it does alleviate dependence on oil. Although some of these accusations are contested [62,63], corn-based bioethanol, for example, has been criticized [64] for negative externalities such as deforestation [65], insufficient [66] – or even negative [67] – $CO_2$ emission reductions, and increasing the volatility of food prices [68]. It is thus important to proceed to rigorous life-cycle analysis before the commercialisation of bio-based (or petrochemical) products [66,69–72].

This thesis focuses on one particular bioprocess called gas fermentation, which is based on autotrophic organisms feeding off inorganic wastes. Not only does this strictly limit the possibility of any negative externality, it actively mitigates GHG emissions by using them as feedstock and consuming them. More specifically, a range of genetic tools was developed to better study and domesticate some of these gas fermenting microbes. I hope that this work will eventually contribute to further development of this promising bioprocess, and will ultimately lead to a fairer, safer, and more sustainable future.

## I.2. Gas fermentation

Gas fermentation, and in particular synthesis gas (or syngas) fermentation, is a bioprocess in which a C1 gas (e.g. CO, $CO_2$ or $CH_4$) is converted into valuable chemicals such as ethanol [73–75] (cf. Figure 6). Syngas is a mixture of mainly CO, $CO_2$ and $H_2$ which is obtained through gasification of biomass, coal and petrochemicals, or even municipal waste. Syngas is also a by-product of existing industrial processes such as steelmaking. This versatility makes it a valuable medium to upcycle a large range of wastes, forming the basis of a low-carbon circular economy. However, the nature of the feedstock – as well as the parameters of the gasification process – affects the final composition of the syngas and its compatibility with downstream applications [76–78]. Microbial fermentation is a particularly interesting way of valorising syngas because it is less sensitive to contaminations and fluctuations of syngas composition than thermochemical processes such as Fischer-Tropsch [75,77,79]. Microbial fermentation also operates at low pressure and low temperature compared to conventional thermochemical processes, and can boast of higher conversion efficiencies and higher product specificity [80].

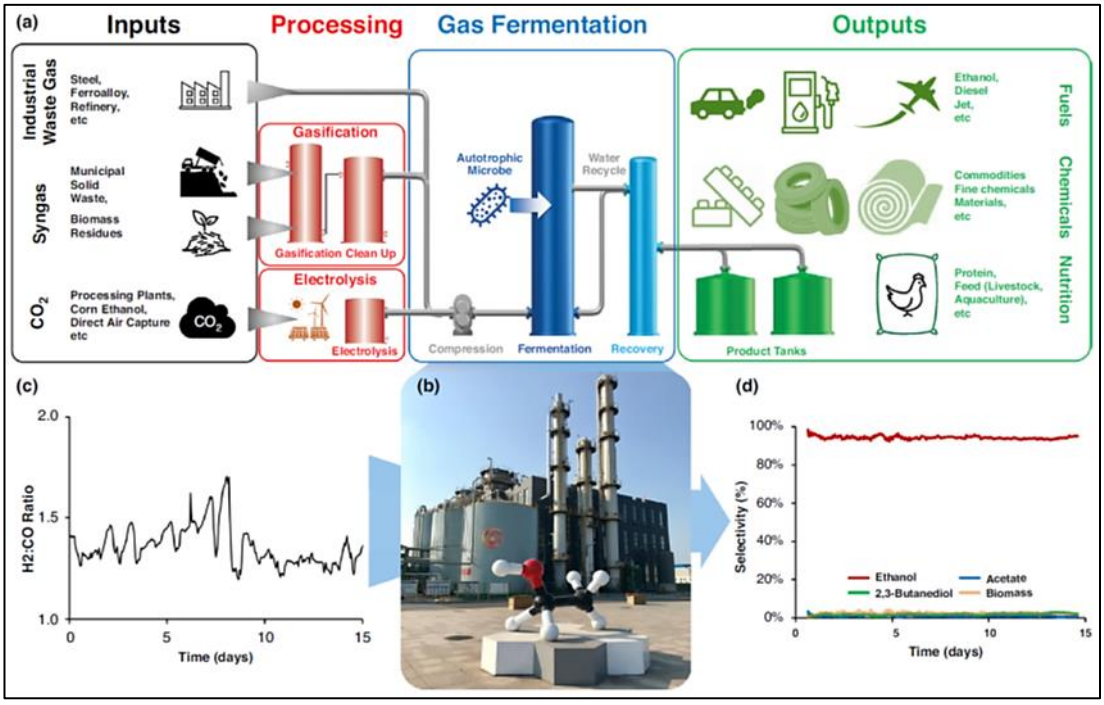



**Figure 6: Overview of the LanzaTech gas fermentation process. (a)** Conceptual schematic of a gas fermentation process, with four potential feedstocks, two types of feedstock processing, the gas fermentation process itself, and finally potential products (currently, only derived from ethanol or biomass). **(b)** Commercial-scale gas fermentation plant (Beijing Shougang LanzaTech New Energy Science & Technology Co., Ltd.). **(c)** Example of fluctuations of feed gas profile from gasification of unsorted, non-recyclable municipal solid waste (Sekisui, LanzaTech) and **(d)** corresponding stable fermentation output (Sekisui, LanzaTech). Reproduced from Köpke and Simpson, 2020 [75].

Nonetheless, several hurdles still hamper large-scale gas fermentation processes [81,82]: low gas transfer coefficients into the liquid phase, vulnerability to some contaminants such as hydrogen cyanide, low productivity rate, limited range of end-product molecules, limited basic understanding of the biology of the gas-fermenting biocatalysts; and, most relevant to this thesis, a lack of effective synthetic biology tools [60,75,83–85].

### I.2.1. Carbon accounting with Carbon Capture, Storage and Utilisation

By giving commercial value to carbon dioxide, gas fermentation facilitates the commercialisation of carbon dioxide removal technologies such as Direct Air Capture (DAC) and Bioenergy with Carbon Capture and Storage (BECCS). These controversial technologies have now become necessary to reach the objectives of the Paris Agreement [12,86,87]. However, storing captured carbon within the products of gas fermentation would only constitute a form of Carbon Capture and Utilisation (CCU) [74]; consequently, gas fermentation is not a negative emission technology unless it is specifically used to produce carbon-polymers which will never be burned. In all other cases, gas fermentation only allows to extract more value out of a molecule of $CO_2$ before it is eventually released into the atmosphere. Gas fermentation products can thus only claim to reduce GHG emissions if they replace their fossil-derived equivalents [75]. Accordingly, the development of negative emission and CCU technologies such as gas fermentation is not an alternative to phasing out fossil resources from the global economy, which should remain the absolute priority [87–89].

### I.2.2. Acetogenesis and *Clostridium autoethanogenum*

Although there are many types of gas fermenting organisms and associated autotrophic pathways [74,90,91], syngas fermentation is reliant on a specific class of strictly anaerobic chemoautotrophic microbes, called acetogens, which can utilize the reductive potential of CO and/or $H_2$ and assimilate the inorganic carbon of $CO_2$ and CO into organic compounds such as biomass and acetyl-CoA [92]. *Clostridium autoethanogenum* is a Gram-positive acetogen of particular interest for its ability to produce large quantities of ethanol, a drop-in fuel, in addition acetate and small amounts of lactate and 2,3-butanediol [93–95]. Since 2018, this model organism has been exploited to produce ethanol at commercial scale by the gas fermentation company LanzaTech Inc. (Skokie, USA) [75] (cf. Figure 6). For these reasons, *C. autoethanogenum* is the sole focus of this thesis.

### *I.2.2.1) The Wood-Ljungdhal pathway*

Acetogens all rely on variations of the reductive acetyl-CoA pathway (or Wood-Ljungdhal pathway) [96], illustrated for *C. autoethanogenum* in Figure 7 alongside the rest of its metabolism. A summary of the enzymology of *C. autoethanogenum* is available in Table 1.

Depending on whether $H_2$ & $CO_2$ or only CO are available, the pathway starts with the oxidation of $H_2$ with a facultative electron-bifurcating hydrogenase/formate dehydrogenase complex (HytA-E/FDH) (cf. Energy conservation commentary 1); or with the oxidation of CO with a CO dehydrogenase/acetyl coenzyme A synthase complex (CODH/ACS). If both are available, CO oxidation is favoured and even inhibits HytA-E/FDH [97,98], but it results in the emission (and thus loss) of $CO_2$ [99].

**Energy conservation commentary 1:**

Electron bifurcation (and its reverse but equivalent mechanism, electron confurcation) is a recently established energy conservation mechanism (in addition substrate-level and electron-transport-linked phosphorylations) in which the reducing power of one electron from an electron pair is increased at the detriment of the reducing power of the second electron [100,101]. This allows to couple an exergonic reaction to an endergonic reaction, which limits the loss of free energy as heat. In the example of HytA-E/FDH, the reversible endergonic reduction of $2H^+$ into $H_2$ ($E_0$' = -414 mV, E'= -350 mV) with ferredoxin ($Fd^{2-}$) ($E_0$' = -400 mV, E'= -500 mV) is made possible by the concomitant exergonic reduction of $H_2$ with NADPH ($E_0$' = -320 mV, E'=-370 mV) [97]. Similarly, the same enzyme complex is able to catalyse the reversible endergonic reduction of $CO_2$ ($E_0$'=-430 mV) with ferredoxin ($Fd^{2-}$) into formate as a result of the simultaneous exergonic reduction of $CO_2$ with by NADPH into formate [101,102].

This mechanism has been proposed to play a key role in preserving the redox equilibrium of the cell [97]: if the pool of ferredoxin becomes too reduced (E'=-520 mV) during growth on CO, HytA-E/FDH would oxidize ferredoxin and NADPH to produce $H_2$ – which would immediately be processed into formate without leaving the enzymatic complex – and thus bring back the reduction potential of the ferredoxin pool to -500 mV.

After reduction of $CO_2$ into formate by HytA-E/FDH, one ATP is hydrolysed by formyltetrahydrofolate synthase (FTHFS) to form formyltetrahydrofolate from formate and tetrahydrofolate (THF) (cf. Energy conservation commentary 2).

**Energy conservation commentary 2:**

This ATP consumption is only compensated during acetate formation with one substrate-level phosphorylation, making necessary the existence of other energy conservation mechanisms. Accordingly, the Rnf complex couples the oxidoreduction of ferredoxin and $NADP^+$ with translocation of protons across the cell membrane; this constitutes a proton motive force which can then be exploited by ATP synthase to phosphorylate ADP. In addition to these two enzymes, the Nfn complex (NADH-dependent ferredoxin-$NADP^+$ oxidoreductase) balances the concentration of NADPH with the concentration of $Fd^{2-}$ + NADH through electron bifurcation and confurcation [102,103].

Methenyl-THF cyclohydrolase (MTHFC) then converts formyl-THF to methenyl-THF (cf. Energy conservation commentary 3), which is immediately converted to methylene-THF by methylene-THF dehydrogenase (MTHFD) through the consumption of one NADPH.

**Energy conservation commentary 3:**

It is still unclear whether the enzyme methylenetetrahydrofolate reductase (MTHFR) is electron bifurcating or not (both options are represented in Figure 7) [102]. Accordingly, the reduction of methenyl-THF into methyl-THF either oxidises one NADH molecule ($E_0$' = -320 mV, E'= -280 mV) or it oxidises two but also reduces one ferredoxin.

In the last step of the methyl-branch, a methyltransferase transfers the methyl group of methyl-THF to the cobalt group of a corrinoid-FeS-protein. There, in the carbonyl-branch of the WLP, the key enzyme of the WLP – the enzymatic complex CODH/ACS – condensates one molecule of CO with the methyl group of the corrinoid-FeS-protein, and the resulting acetyl group is bonded to Coenzyme A (CoA) to finally form acetyl-CoA.

From there, acetyl-CoA can continue its catabolic journey towards the final electron acceptors (acetate, ethanol, 2,3-butanediol, and lactate) or form the basis of biomass production.

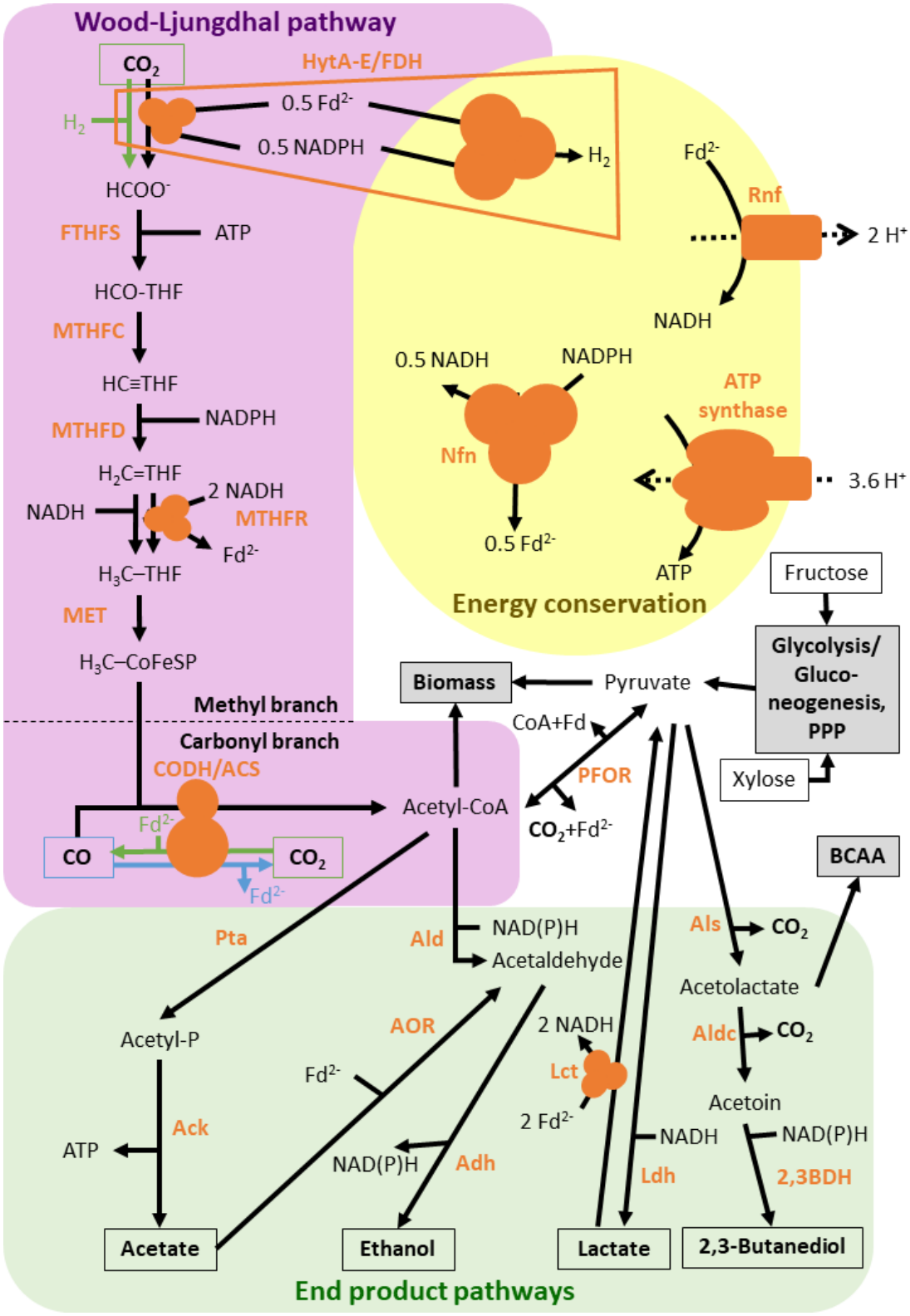



**Figure 7: Representation of the metabolism of *C. autoethanogenum*.** Three main pathways have been highlighted: the Wood-Ljungdhal pathway, which encompasses the reduction of carbon dioxide ($CO_2$) into acetyl-CoA with the concomitant oxidation of hydrogen ($H_2$, green) and/or carbon monoxide (CO, blue); the end product pathways which sees the further conversion of acetyl-CoA into acetate, ethanol, lactate or 2,3-butanediol; and the energy conservation pathway which maintains the redox equilibrium of the cell and produces ATP from a transmembrane proton gradient. Enzymes are represented and labelled in orange. = electron-bifurcating enzyme; BCAA= Branched amino acids synthesis pathway; PPP= Pentose Phosphate Pathway.

#### *I.2.2.1) Heterotrophic pathways*

*C. autoethanogenum* can also consume fructose through the glycolysis pathway described by Embden-Meyerhof-Parnas, or xylose through the pentose phosphate pathway [95]. The reducing equivalents produced in these pathways can be used by the WLP to generate extra ATP and acetyl-CoA, notably through the conversion of pyruvate to acetyl-CoA catalysed by the enzyme pyruvate:ferredoxin oxidoreductase (PFOR) which also reduces one ferredoxin [104–106]. Additionally, *C. autoethanogenum* can obtain ATP from amino acids (aa) catabolism, arginine in particular (not represented) [93].

#### *I.2.2.2) End product pathways*

##### I.2.2.2)(1) Acetate

Acetate and ethanol production deplete most of the acetyl-CoA stock. Acetate is derived from acetyl-CoA in two steps: first by phosphate acetyltransferase (Pta) to produce Acetyl-phosphate, then acetate kinase (Ack). As mentioned previously, this pathway generates one ATP, which is the highest energy yield of all end products [107].

##### I.2.2.2)(2) Ethanol

Counter-intuitively, the pathway to ethanol which results in the most energy for the cell doesn't start with acetyl-CoA but acetate. The enzyme aldehyde ferredoxin oxidoreductase (AOR) oxidizes $Fd^{2-}$ to reduce acetate into acetaldehyde, which is then oxidized into ethanol by alcohol dehydrogenase (Adh) and generates one NAD(P)H in the process. By pushing more acetyl-CoA through the acetate pathway, ethanol production through AOR generates more ATP than through direct production of acetaldehyde from acetyl-CoA with an aldehyde dehydrogenase (Ald) which also consumes one NAD(P)H [108].

##### I.2.2.2)(3) Lactate

Lactate and 2,3-butanediol are both derived from pyruvate, which can be obtained from acetyl-CoA with PFOR at the price of one $Fd^{2-}$. Lactate can be produced from a conventional lactate dehydrogenase (Ldh) which consumes one NADH, or consumed by several electron-bifurcating lactate dehydrogenases (Lct) [109] which also reduce $NAD^+$ with $Fd^{2-}$.

##### I.2.2.2)(4) 2,3-butanediol

Acetolactate synthase (Als) produces acetolactate from two pyruvate molecules. From there, other acetolactate synthases can redirect the acetolactate into the anabolism of branched-chain amino acids (valine, leucine and isoleucine) [94]. Carboxylation of the remaining acetolactate to acetoin is catalysed by acetolactate decarboxylase (Ald), and 2,3-butanediol is finally obtained by reduction of acetoin with NAD(P)H with NADH dependent butanediol dehydrogenase (Bdh) or an NADPH dependent primary-secondary alcohol dehydrogenase (CLAU_0532) [110].

**Table 1: Enzymology of *C. autoethanogenum* with associated pathways, reactions, and genomic loci.**

| Pathway | Acronym | Reaction | Genes |
|---|---|---|---|
| WLP | **CODH** | $CO+Fd^{2-}+H_2O \rightleftharpoons CO_2+Fd+2H^+$ | CLAU_1578, 2924, 2924 |
| | **FDH** | $CO_2+0.5\ Fd^{2-}+0.5NADPH+0.5H^+ \rightleftharpoons HCOO^-+0.5\ Fd+0.5NADP^+$ | CLAU_2713 |
| | **FDH (b)** | $CO_2+H_2 \rightleftharpoons HCOO^-+H^+$ | CLAU_2713 |
| | **FTHFS** | $HCOO^-+ATP+THF \rightleftharpoons HCO\text{-}THF+ADP+Pi$ | CLAU_0275 |
| | **MTHFC** | $HCO\text{-}THF+H^+ \rightleftharpoons HC{\equiv}THF+H_2O$ | CLAU_1575 |
| | **MTHFD** | $HC{\equiv}THF+NADPH \rightleftharpoons H_2C{=}THF+NADP^+$ | CLAU_1574 |
| | **MTHFR** | $H_2C{=}THF+NADH+H^+ \rightleftharpoons H_3C\text{-}THF+NAD^+$ | CLAU_1572-1573 |
| | **MTHFR (b)** | $H_2C{=}THF+2NADH+Fd \rightleftharpoons H_3C\text{-}THF+Fd^{2-}+2NAD^+$ | CLAU_1572-1573 |
| | **MET** | $H_3C\text{-}THF+CoFeSP+H^+ \rightleftharpoons H_3C\text{-}CoFeSP+THF$ | |
| | **ACS** | $H_3C\text{-}CoFeSP+CO+CoA \rightleftharpoons acetyl\text{-}CoA+CoFeSP+H^+$ | CLAU_1569 |
| Energy conserva-tion | **HytA-E** | $1.5H^++0.5NADPH+0.5Fd^2 \rightleftharpoons H_2+0.5NADP^++0.5Fd^-$ | CLAU_2718-2722 |
| | **Rnf** | $Fd^{2-}+NAD^++3H^+(in) \rightleftharpoons Fd+NADH+2H^+(out)$ | CLAU_3144-3149 |
| | **ATP synthase** | $ADP+Pi+3.66H^+(out) \rightleftharpoons ATP+2.66H^+(in)+H_2O$ | CLAU_2288-2292 |
| | **Nfn** | $NADPH+0.5\ Fd+0.5\ NAD^+ \rightleftharpoons NADP^++0.5\ Fd^{2-}+0.5\ NADH+0.5H^+$ | CLAU_1539 |
| End products | **PFOR** | $CO_2+acetyl\text{-}CoA+H^++Fd^{2-} \rightleftharpoons pyruvate+COA+Fd$ | CLAU_2947,0896 |
| Acetate | **Pta** | $Acetyl\text{-}CoA+Pi \rightleftharpoons CoA+acetyl\text{-}P$ | CLAU_3274 |
| | **Ack** | $ADP+acetyl\text{-}P \rightleftharpoons ATP+acetate$ | CLAU_3275 |
| Ethanol | **AOR** | $Acetate+Fd^{2-}+H+ \rightleftharpoons acetaldehyde+H_2O+Fd$ | CLAU_0089, 0081, 0099, 3655, 3656 |
| | **Ald** | $Acetyl\text{-}CoA+NAD(P)H+H^+ \rightleftharpoons acetaldehyde+CoA+NAD(P)^+$ | CLAU_1772, 1783, 3204 |
| | **Adh** | $Acetaldehyde+NAD(P)H+H^+ \rightleftharpoons Ethanol+NAD(P)^+$ | CLAU_0532, 0534, 1794, 3861, 3655, 3656, … |
| Lactate | **Ldh** | $Pyruvate+NADH+H^+ \rightleftharpoons lactate+NAD^+$ | CLAU_1108 |
| | **Lct** | $Lactate+2NAD^++2Fd^{2-} \rightleftharpoons Pyruvate+2NADH+2Fd$ | CLAU_0111-0113, 0235-0237, 3386-3388 |
| 2,3-butane-diol | **Als** | $2Pyruvate+H^+ \rightleftharpoons CO_2+acetolactate$ | CLAU_1694 |
| | **Aldc** | $H^+ + acetolactate \rightleftharpoons Acetoin+ CO_2$ | CLAU_2851 |
| | **2,3BDH** | $Acetoin+NADH+H^+ \rightleftharpoons NAD^++2,3\text{-}Butanediol$ | CLAU_0370, 0532 |

## I.3. Basic principles of genome editing

If knowledge of the metabolism of acetogens is important to manipulate them, so is an awareness of their genome biology. Most genome editing strategies work by hijacking one of the host's DNA repair pathways. As such, it is critical to first understand these DNA repair pathways to gain perspective on the advantages and limitations of each mutagenesis strategy. This section reviews these basic concepts so that they can be referred to throughout the thesis.

Once the appropriate mutagenesis strategy has been selected based on its associated DNA repair pathway, special attention must be paid to the genomic context of the desired mutation to avoid unintended consequences. Polar mutations are one of such problematic mutations, perhaps the most common in prokaryotes, and they are also introduced in this section.

### I.3.1. SOS response

Because many DNA repair pathways can also be mutagenic, they are tightly regulated. Eukaryotes detect and repair DNA damage through a set of checkpoints which are coordinated with the cell cycle [111]. Prokaryotes use a different pathway called SOS response, which is based on the detection of single-stranded DNA (ssDNA) [112]. Only the prokaryotic SOS response will be reviewed here.

The SOS response is based on two proteins: LexA and RecA. LexA is a self-cleaving transcriptional repressor which normally binds an operator called the SOS box to suppress the SOS response. RecA is a protein which binds ssDNA (not to be confused with SSB, the ssDNA-binding protein encoded by *ssb* in *E. coli*) and which stimulates self-cleaving of LexA when it is activated (RecA*) by a nucleoside triphosphate. Self-cleavage of LexA relaxes the expression of over 50 genes involved in DNA repair, initiating the SOS response (cf. Figure 8). Importantly, not all genes are de-repressed at once or at the same level, and some genes involved in DNA repair are not part of the SOS regulon. DNA repair pathways with low error potential are induced after only a small decrease of the LexA pool, while more mutagenic enzymes will require a drastic and sustained decrease of LexA in the cytoplasm before being expressed.

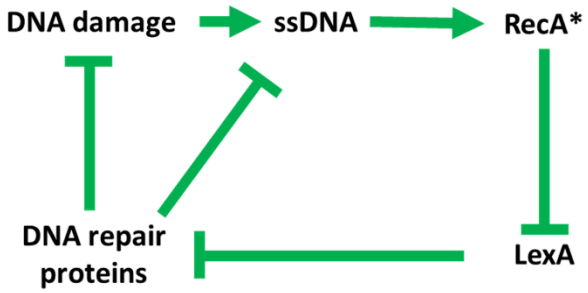



**Figure 8: Logical network of the SOS response.** DNA damage leads to an increase in ssDNA, which activates RecA into RecA*, which promotes LexA self-cleavage, limiting LexA's ability to repress expression of the DNA repair enzymes involved in the SOS response. The now expressed DNA repair enzymes can either replicate the ssDNA directly or fix the DNA damage, which eventually reduces the amount of ssDNA in the cell. In turn, this reduces the amount of RecA* and allow LexA to shut down the SOS response. A triangular arrowhead represents stimulation, and a perpendicular arrowhead represents inhibition.

Most notably, the UvrD helicase is one of the first to be expressed, followed by RecA and other proteins involved in homologous recombination; DNA polymerase II, with low processivity but a performant proofreading activity [113], is also induced along with the cell division inhibitor SulA. At last, error prone DNA polymerases IV and V are expressed as a last attempt to rescue the cell at the price of numerous mutations. LexA expression is also upregulated during the SOS response, insuring that the SOS genes are repressed as soon as the DNA damage has been repaired.

Many kinds of DNA damage produce ssDNA, either by directly producing single-stranded and double-stranded breaks (SSBs and DSBs), or through interfering with the replication process. Indeed, when DNA polymerase III fails to replicate a damaged base and stalls, a new replisome can be primed downstream of the lesion to allow the replication fork to continue undisturbed. SSB proteins will initially bind and protect the stretch of ssDNA left in between the stalled DNA polymerase and the downstream priming locus, but they will progressively be replaced by activated RecA* ssDNA-binding proteins. As discussed previously, RecA* will induce the SOS response; they will also actively participate in DNA damage tolerance pathways such as translesion DNA synthesis (TLS) and homology-directed gap repair (HDGR) (cf. Figure 9). These pathways do not actually repair a mutation, they just allow the cell to finish replicating its DNA.

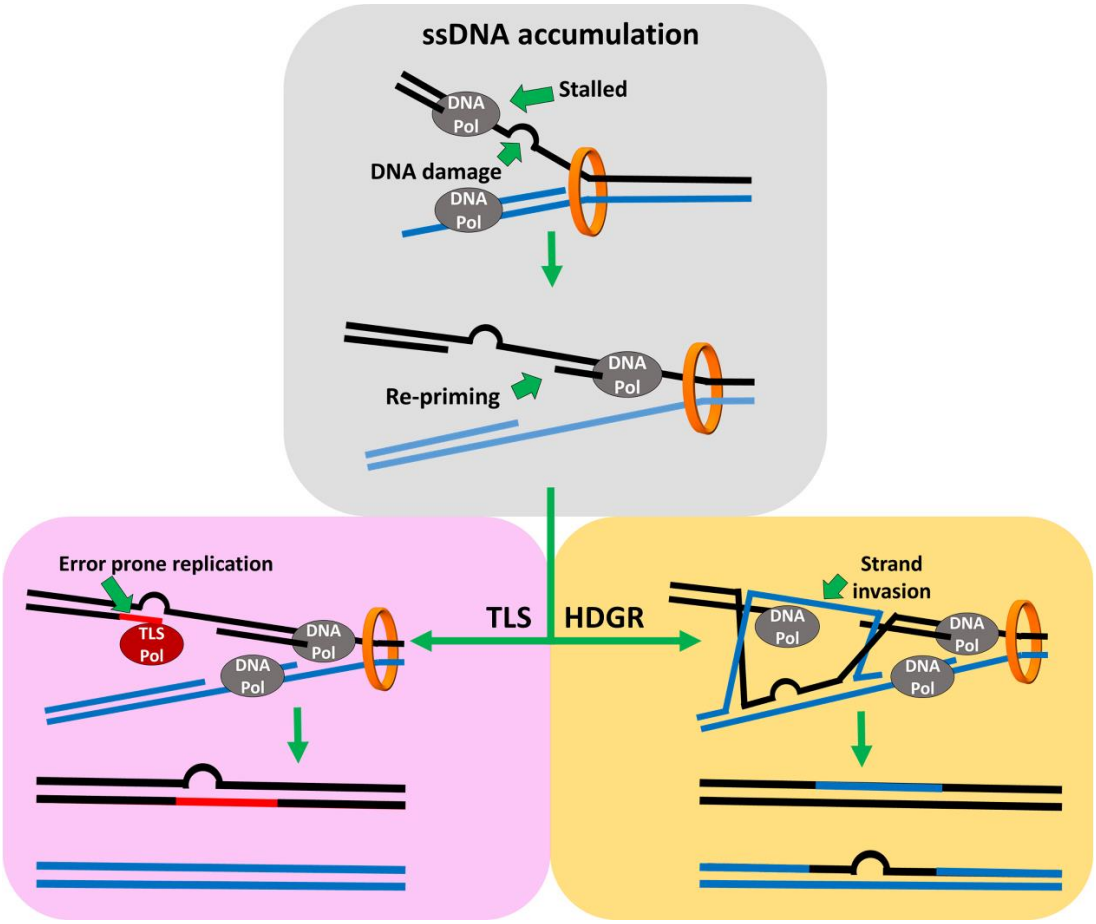



**Figure 9: Damage tolerance pathways.** DNA damage can stall the main DNA polymerase, leading to the accumulation of stretches of ssDNA. Synthesis of the complementary strand in these loci can be done through translesion DNA synthesis (TLS) which uses error prone DNA polymerases to fill in the gap with random nucleotides, or through Homology-directed Gap Repair (HDGR), which uses the homologous recombination mechanism of strand invasion to acquire a WT DNA template from the sister chromatid and transfers the DNA damage to it after resolution of the Holliday junctions. Not represented here are SSB and recA proteins.

In TLS, DNA polymerases II, IV or V (also known as TLS polymerases) synthesize new DNA to fill in the ssDNA gap at the price of a high error rate. The resulting DNA segment is then ligated to downstream DNA previously replicated by DNA polymerase III. In HDGR, homologous recombination with the sister chromatid fills in the ssDNA gap, and the damaged base is transferred to the sister chromatid with the resolution of the Holliday Junction. In each case, the DNA damage has not been fixed, but its potentially lethal impact on DNA replication has been averted [112,114–118].

The SOS response can induce several error-prone repair pathways which are usually not desired during a mutagenesis experiment. Indeed, the SOS response could introduce random mutations anywhere in the genome, or on the mobile genetic element used to carry out the targeted mutagenesis. It can contribute to the occurrence of false positives (colonies that

survive the selective pressure meant to isolate the desired mutant from the WT even though they do not have the desired mutation) and makes necessary to control for off-target mutations when characterising mutants (e.g. complementation study).

### I.3.2. Mismatch repair

Mismatch repair (MMR) is usually based on two proteins: MutS and MutL [119–122]. MutS screens the genome after DNA replication and detects mismatches; if it finds one, it recruits the MutL endonuclease which nicks the daughter strand – i.e., the strand which was synthesised from the template strand during DNA replication (Figure 10.a).

In *E. coli*, the daughter strand is identified by its absence of methylation. During DNA synthesis, *E. coli* methylates the adenines within all d(GATC) sites; however, this methylation process in not instantaneous, which briefly allows the MMR complex to discriminate between the template strand and the daughter strand. Such a GATC locus is described as being hemimethylated: out of the two strands of DNA, only the template strand is methylated. In contrast with the MMR system of eukaryotes and most studied prokaryotes, *E. coli* MutL seems to only play a role of enabler and matchmaker, with the endonucleolytic activity taken over by another protein called MutH. MutH is sequence- and methylation-specific, cutting 5' of an hemimethylated d(GATC) site.

In organisms which do not possess a MutH homolog (which includes *C. autoethanogenum* ), the daughter strand is thought to be identified through nicks in the DNA backbone [119–122]. Such discontinuities can indeed be a consequence of the replication of a new DNA strand, especially on the lagging strand of the replication fork but also on the leading strand [123]. In the MMR of these organisms, an additional nick is thus produced by the MutL homolog next to the mismatch, on the discontinuous strand.

Once the daughter strand has been nicked, a DNA helicase unwinds the DNA from the nick towards the mismatch. This unwinding is bidirectional: it can occur in a 5'→ 3' or 5'→3' direction. SSB proteins then protect the template strand against nuclease attack, while various exonucleases trim down the daughter strand until the mismatch sequence has been eliminated. A DNA polymerase and DNA ligase then respectively synthesize a new sequence complementary to the template strand and ligate it to the rest of the daughter strand [Figure 10(b)].

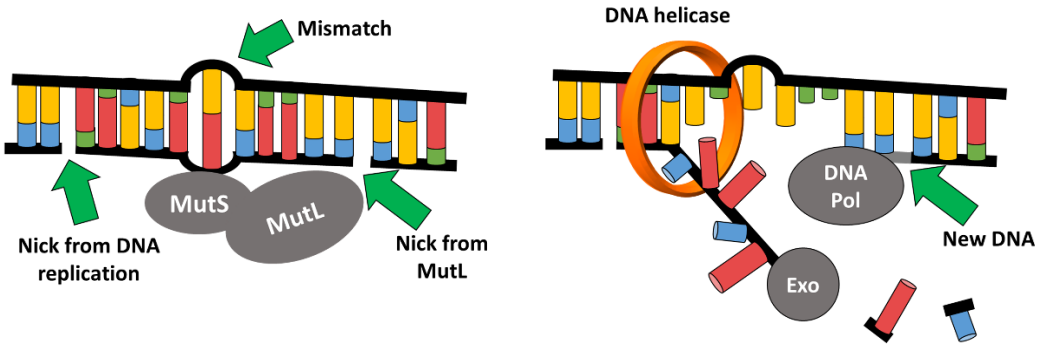



**Figure 10: Summary of mismatch repair (MMR) in eukaryotes and many prokaryotes. (a)** MutS and MutL detect a mismatch and cut the strand with is already nicked (often the daughter strand). **(b)** The nicked strand is unwound by a DNA helicase, stabilised by SSB proteins (not represented), and digested by various exonucleases (Exo) until the mismatch has been removed. Subsequently, DNA polymerase (DNA pol) synthesises a new complementary strand which is finally ligated to the rest of the daughter strand by a DNA ligase (not represented).

This DNA repair pathway is being exploited in base editing mutagenesis as implemented in Chapter V:.

### I.3.3. Base excision repair

Base excision repair (BER) is a mechanism used to remove modified bases such as uracil or 3-methyladenine [124]. First, a DNA glycosylase flips a modified base into its active site without unwinding the DNA helix, then it breaks the glycosidic bonds to remove the modified base without cleaving the DNA backbone (Figure 11.a). This creates an apurinic or apyrimidic site (AP site) depending on the base which has been removed. An AP endonuclease then nicks the backbone adjacent to the AP site, and a deoxyribose phosphodiesterase removes the leftover deoxyribose phosphate backbone [Figure 11(b)]. In *E. coli*, this exonuclease activity can be taken up by the proofreading domain of DNA polymerase I, which will then proceed to synthesise a new nucleotide complementary to the intact strand. At last, the final nick is bridged by a DNA ligase.

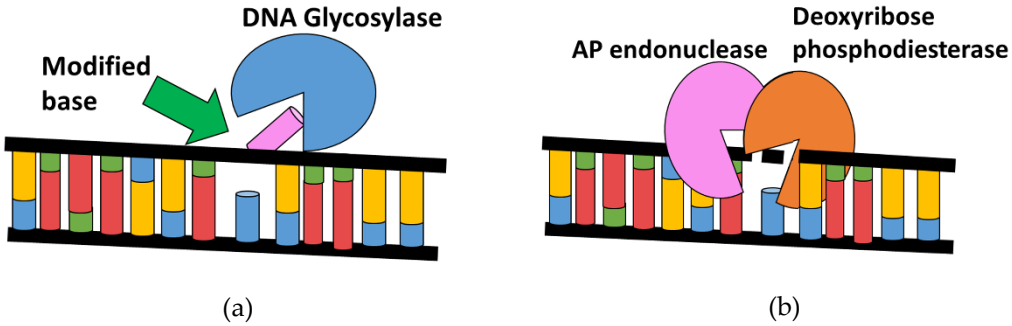



**Figure 11: Summary of base excision repair (BER). (a)** A specialised DNA glycosylase removes a modified base and creates an AP site. **(b)** AP endonuclease and deoxyribose phosphodiesterase respectively cut the DNA backbone and remove the leftover deoxyribose phosphate. In the last step (not represented), DNA polymerase and DNA ligase respectively synthesise the missing nucleotide and ligate it to its adjacent base.

Specialised DNA glycosylases have been identified for different base modifications (AlkA and Tag for 3-methyladenine, Ung for uracil).

### *I.3.3.1) BER of deaminated cytosines*

A particularly prevalent case of BER is the removal of deaminated cytosines [124]. Cytosine bases spontaneously lose their amine group over time, becoming uracil bases (Figure 12). Because uracil is not a canonical DNA base, and because it has the potential to generate a mutation by annealing with adenine, it is detected and removed from DNA by a uracil DNA glycosylase (UDG), which starts the BER pathway as previously explained. Because cytosine deaminations are so prevalent, UDG are ubiquitous enzymes. This DNA repair pathway is the principal obstacle to base editing as implemented in Chapter V:.

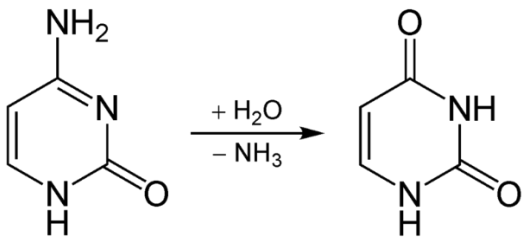



**Figure 12: Cytosine deamination into uracil.**

### I.3.4. Non-homologous End Joining

Non-homologous end joining (NHEJ) is the most common double-strand break (DSB) DNA repair mechanism in eukaryotes, but it is rare in prokaryotes [125]. It is characterized by its ability to fuse two double-stranded DNA helixes without any homology, facilitating horizontal gene transfer (HGT) and large chromosomic reorganisation events [126–128]. However, it is also often leaving scars in the process of joining the ends of each DNA strand, which leads to insertion-deletion (indel) mutations. This property is often exploited to knock out genes in mutagenesis experiments: indel mutations have indeed a high chance of altering the frame of translation of any gene, scrambling the aa sequence and quickly leading to the random occurrence of a STOP codon [129–132].

Both prokaryotic and eukaryotic systems are based on the Ku proteins, which bind and align together both broken dsDNA helixes. In eukaryotes, juxtaposition of the DNA ends is catalysed by the DNA-dependent protein kinase catalytic subunit (DNA-PKcs), and ligation of the DSB is made by DNA ligase IV with the assistance of proteins XRCC4, XLF and PAXX [133,134]. If the ends of each broken strand are so damaged that they cannot support ligation, they need to be removed by the Artemis endonuclease and replaced by DNA polymerases mu (pol μ) and lambda (pol λ) as well as terminal deoxynucleotidyl transferase (TdT). Pol μ and λ are error-prone eukaryotic TLS DNA polymerases while TdT can add nucleotides without template [126,134]. As such, NHEJ is not necessarily mutagenic: it only leaves a scar if ends resection is necessary or if information is already lost during the damaging of the DNA [126,128,133,135,136].

The prokaryotic NHEJ pathway such as found in *Mycobacterium tuberculosis* [125] shows the same fundamental steps (Figure 13). However, it is absent in many prokaryotes and is simpler, only requiring DNA ligase D in addition to the Ku protein homolog. On top of its ligase domain, Ligase D often exhibits a DNA polymerase domain and a nuclease domain which enables strand resection [125–127,132,137–139].

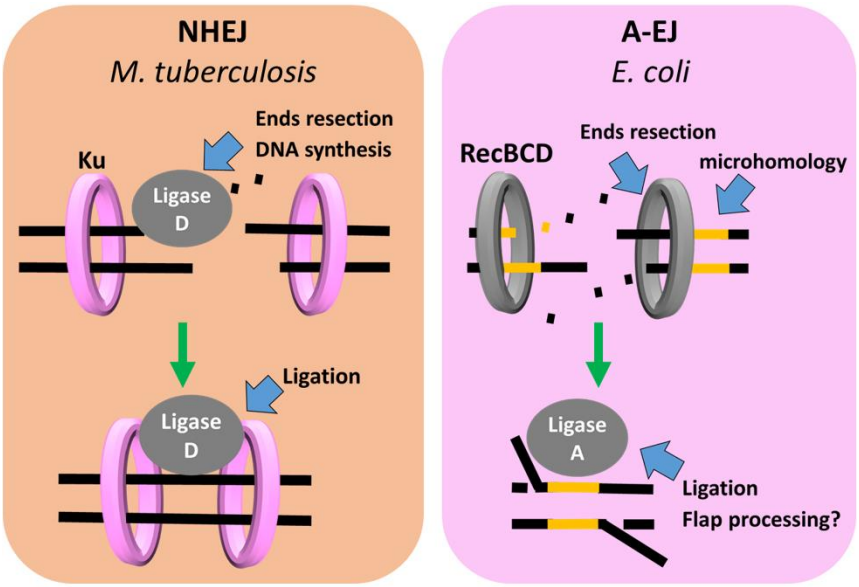



**Figure 13: Comparison of the conventional NHEJ pathway as found in *M. tuberculosis* with the alternative A-EJ pathway as found in *E. coli*.** In ***M. tuberculosis***, specialized Ku proteins bind and align the broken dsDNA strands, while ligase D removes the damaged bases, synthesize new DNA to fill in gaps, and ligates both DNA duplexes together. In ***E. coli***, RecBCD processes the DNA duplexes to expose long ssDNA stretches and reveal microhomologies between the ssDNA of each duplexes, and ligase A proceeds to ligate both duplexes together after the non-homologous flaps have been removed.

### *I.3.4.1) Alternative NHEJ pathway*

Alternative minor NHEJ pathways which do not make use of Ku proteins were also identified in eukaryotes (Alt-NHEJ/MMEJ/A-NHEJ, SSA)[128,135,136] and prokaryotes (A-EJ) [131]. They make use of extensive ends resection and exploit micro-homologies to anneal the dsDNA fragments (Figure 13). They are much more likely to produce mutations and large deletions. They were discovered relatively recently, and, as such, some mutagenesis event previously attributed to the canonical NHEJ system might actually have been caused by these alternative pathways [135]. Most importantly, it means that many prokaryotes such as *E. coli* which were previously assumed to be incapable of NHEJ are actually somewhat proficient in repairing DSB without homologous DNA template, which opens the door to new mutagenesis strategies [129,130,132] and hints at limitations for existing ones.

In *E. coli*, ends resection and ligation during A-EJ are thought to carried out by the RecBCD complex and ligase A, respectively, but this pathway is still poorly described [131,140,141].

### I.3.5. Homologous recombination

Homologous recombination (HR) is a conservative DSB repair mechanism which is tightly associated with DNA replication, as it requires access to an intact DNA duplex which has identical sequence to the locus of the DSB. For this reason, it is limited to the S and G2 phases of the cell cycle in eukaryotes but it is thought to be the most common DSB repair mechanism in prokaryotes [140,142].

Several variations of this canonical pathway have been described (synthesis-dependent strand annealing (SDSA), break-induced replication (BIR), HDGR), all exploiting the same basic principles [128,140,143–146]. They have been collectively regrouped under the name of homology-directed repair (HDR) which designates all DNA repair mechanisms which rely on homology with intact DNA molecules to repair DNA. In the interest of brevity, only the canonical prokaryotic HR pathway [147] will be described here.

In *E. coli*, HR is initiated by two proteins: the ssDNA binding protein RecA and the helicase-endonuclease RecBCD. The RuvABC helicase/resolvase and RecG and PriA helicases are also involved in later steps (cf. Figure 15).

Upon DSB, the RecBCD holoenzyme binds blunt or nearly blunt ends of linear DNA duplexes and unwinds their helix. Its endonuclease domains preferentially degrade either strand depending on local metabolite concentrations (mainly the ratio of $Mg^{2+}$ to ATP) [141,148]. When RecBCD encounters a Chi site (cross-over hotspot instigator, 5'-GCTGGTGG-3') while moving from its 3' end, it cuts the DNA duplex to isolate 3'-ssDNA. RecBCD then actively loads RecA proteins onto the 3'ssDNA, priming it for strand invasion. If the ssDNA does not recombine, RecBCD will eventually degrade that strand too [148–150].

If a homologous DNA molecule is available (sister chromatid, additional vector/chromosome, or even foreign DNA) the RecA filament will guide the 3'-ssDNA to its homologous sequence, invading the homologous DNA duplex and creating a 3-way DNA junction or displacement loop (D-loop). RuvAB and/or RecG then bind the three-way junction and move it along the homology region, migrating the displaced strand from one DNA duplex to another. The three-way junction becomes stable when it is joined by the non-invading strand and becomes a 4-way DNA junction called Holliday junction (HJ) (cf. Figure 14) [151]. Branch migration continues in either direction, until ruvC joins the ruvAB-HJ complex. When RuvC recognizes a specific short degenerate sequence (5'-(A/T)TT(G/C)-3'), it resolves the HJ by cutting two strands of opposite polarity, which can then be ligated [152–154].

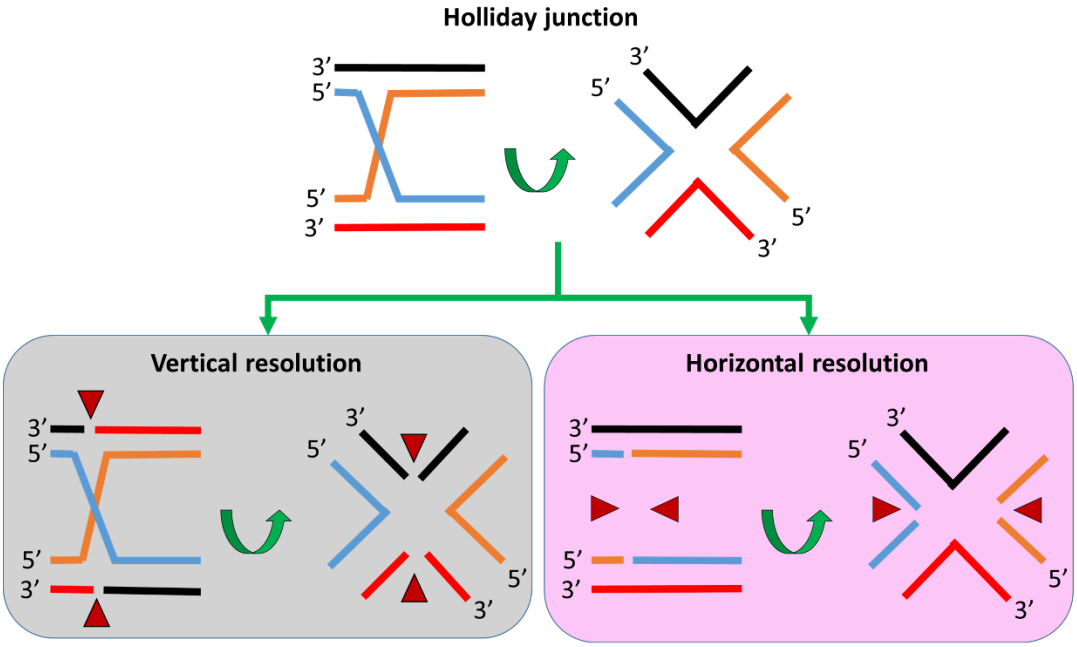



**Figure 14: Representation of four-way DNA junctions (Holliday junctions).** They can be resolved either horizontally or vertically, cutting two homologous strands of opposite polarity and resulting in different configuration of DNA duplexes after ligation.

At the same time, strand invasion keeps displacing the homologous strand, which elongates the D-loop and provides a DNA template for the replication of the non-invading strand. The PriA helicase joins RecG at the D-loop and recruits the DnaB helicase to initiate a replication fork, with the extension of the invading strand as the leading strand and the replication of the non-invading strand as the lagging strand [155–157]. DNA replication then continues until it joins another replication fork, potentially one initiated in the same way by the other end of the DSB [151].

As such, if an HR event occur between the genome and the extremities of the same DNA template, then the template has a chance of being seamlessly integrated into the genome even if the intervening sequence is not homologous to the chromosome (cf. Figure 16) [146,157].

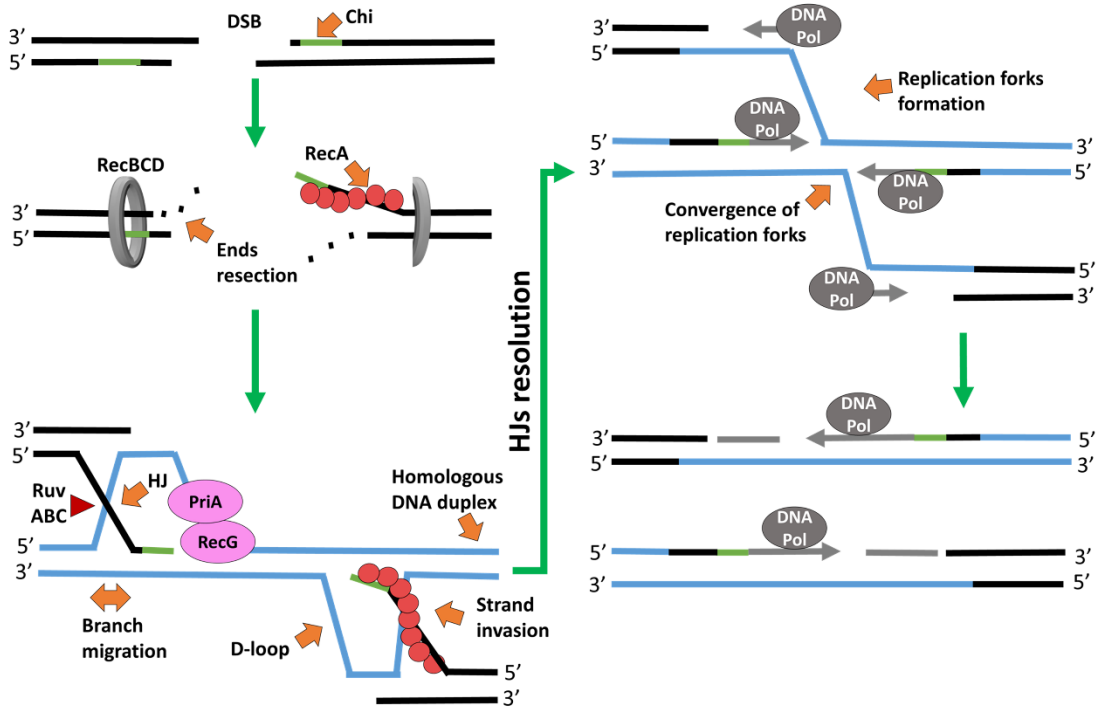



**Figure 15: RecBCD model of homologous recombination in *E. coli* when [ATP]<[Mg2+].** RecBCD digests a DSB and loads a 3'-end with RecA upon recognition of a Chi site. RecA facilitates strands invasion of a homologous DNA duplex, leading to the formation of a four-way DNA junction (HJ) which is stabilised by RuvAB and RecG. RecG helps recruiting PriA which can recruit DnaB and the rest of the replisome to create a new replication fork. Upon convergence of two replication forks, both DNA duplexes finally separate.

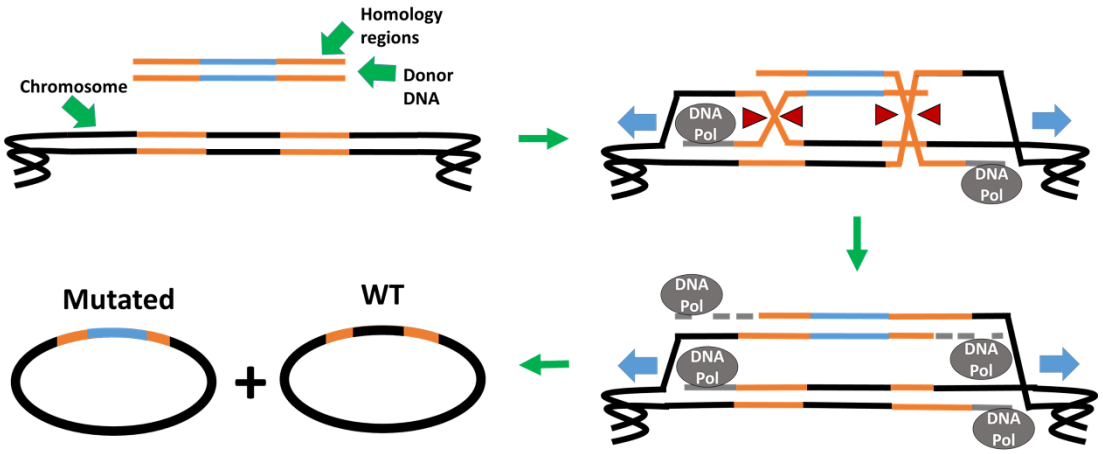



**Figure 16: "Ends-out" model of homologous recombination-mediated targeted mutagenesis in prokaryotes.** A double-stranded donor DNA with flanking regions each homologous to a locus in the chromosome. Strand invasion of the chromosome by each 3' end of the donor DNA creates two Holliday junctions and two D-loops extending in opposite direction. Upon appropriate resolution of the Holliday junctions and successful priming of a new replication fork in each D-loop, replication of the whole chromosome can proceed. Upon segregation of the chromatids, one chromosome has integrated the donor DNA while the other remains unchanged (WT) [146,157]. The "Ends-in" model is equivalent to the standard RecBCD model presented in Figure 15, with a DSB in the chromosome initiating strands invasion at each homologous end of the donor DNA, and the resulting replication forks converging in the middle of the donor DNA.

HDR is extensively exploited in genome editing strategies as it does not leave any scar and is very flexible – allowing to knock out genes as easily as knocking them in. In many organisms, it is sufficient to provide foreign DNA with flanking sequences homologous to the desired integration locus to induce successful HDR events. Associated with the appropriate selective pressure to isolate the mutant cells, it is a simple and powerful method [142,158].

### I.3.6. Polar mutations

Polar mutations are defined as mutations which have an impact on the translation of proteins downstream to the location of the mutation within the same translational or transcriptional system [159]. To understand this concept, it is useful to remember that DNA – like RNA or proteins – is a polar molecule, which means that it can be objectively represented with a start and an end (respectively: 5' and 3' for DNA or RNA, and N-terminal and C-terminal for proteins). Similarly, DNA transcription into RNA and RNA translation into proteins are both polar processes: they start upstream, at the 5'-end of DNA or RNA, and end downstream, at the 3'-end. As such, DNA mutations which affect the transcription or translation of an upstream gene might also impact the transcription of a downstream gene. This is especially likely if several genes are organised within an operon. By definition, an operon is a set of genes which share the same promoter [160]; as such, the genes of an operon are all transcribed on the same RNA molecule, but they do not necessarily share the same ribosome binding site, nonsense codon or reading frame.

Polar mutations in operons can be divided into two categories: polar mutations in functional genes and polar mutations in structural genes [161]. The formers occur in sequences which regulate transcription, such as operators and promotors, or even within the coding sequence of transcriptional repressors or activators which regulate their own operon. It is rather straightforward that a mutation occurring in these regions which are specifically tasked to regulate the transcription of an operon might disrupt the transcription of the entire operon. The latter category of polar mutations is more complex. Structural polar mutations are caused by nonsense mutations, also called premature STOP codons – more specifically mutations which produce a `TAA`, `TGA` or `TAG` triplet within the reading frame of a coding sequence. Obviously, a nonsense mutation is expected to affect the translation of the coding sequence in which it occurs, but for a long time it was not clear how these mutations could affect the translation of neighbouring genes in an operon.

Indeed, in some operons such as the *lacZ* operon, it was observed very quickly that nonsense mutations which occurred earlier in the sequence of *lacZ* decreased the expression of the

next two genes of the operon (*lacY* and *lacA*) more effectively than nonsense mutations which occurred later in the gene [162]. When it was discovered that this phenomenon disappeared in strains where *suA* – which codes for the termination of transcription factor rho – was knocked out, a model of structural polar mutations could be built [163,164]. In our current understanding of polar mutations in structural genes of operons, when a nonsense mutation occurs in a gene, it exposes rho-dependant transcriptional terminators which would have been otherwise protected by ribosomes (Figure 17). In bacteria, transcription and translation are indeed coupled, with many ribosomes covering the nascent mRNA molecule as soon as it transcribed. As such, rho factors are unable to access rho-dependant terminators for as long as translation is occurring. If translation were to slow down too much or to stop completely, rho factors would bind to their terminators within the untranslated part of the RNA and interact with RNA polymerase to interrupt transcription of the downstream genes [165].

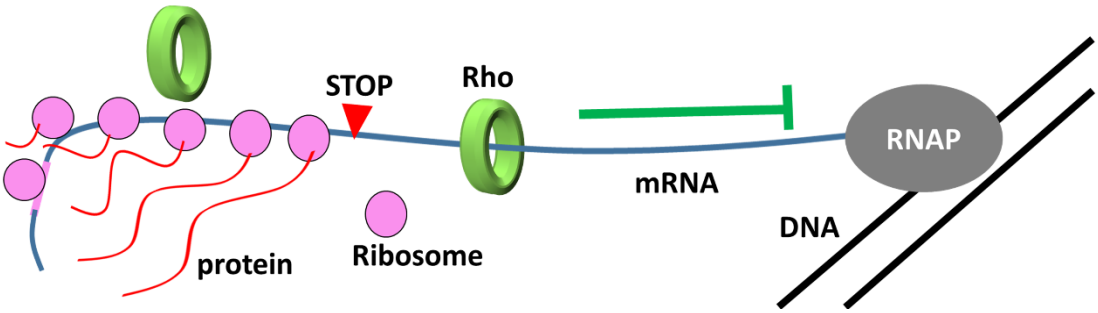



**Figure 17: Model of the mechanism of polarity in the expression of the *E. coli lac* operon. A premature STOP codon exposes stretches of DNA to interactions with the termination factor Rho, which inhibits transcription of the downstream genes. RNAP= RNA polymerase.**

However, this model of polarity in structural genes is by no means exhaustive: for example, the polarity of nonsense mutations in the phage lambda operon has been shown to be caused instead by secondary structures which prevent ribosomes from binding the ribosome binding sites of downstream genes (Figure 18). In the WT operon, these structures are unfolded during translation of the previous gene by ribosomes; as such, if translation ends prematurely, the rest of the mRNA molecule remains folded onto itself, the next ribosome binding site stays inaccessible, and translation of the downstream genes cannot proceed [164]. In some cases, mutations within the coding or the non-coding regions of the operon might also affect the activity of endoribonucleases such as RNAse P, which might in turn change the stability of the mRNA and its accessibility to ribosomes [166,167].

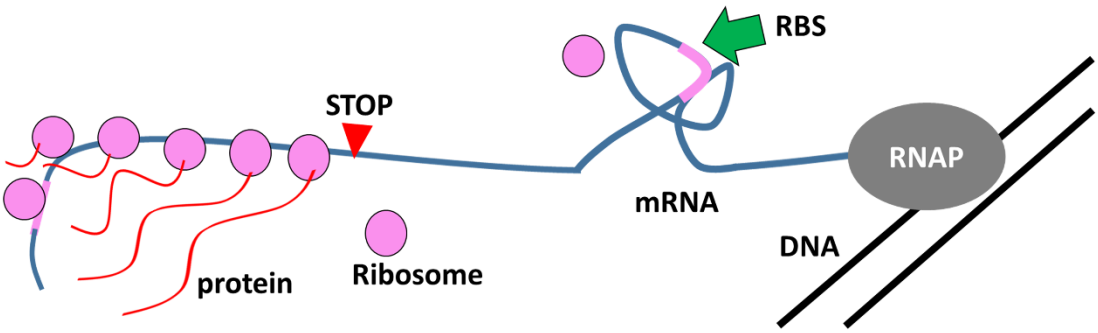



**Figure 18: Model of the mechanism of polarity in the phage lambda operon. A premature STOP codon prevents ribosomes from unfolding the secondary structure of the mRNA, which inhibits translation of the downstream genes. RNAP= RNA polymerase.**

In summary, polar effects are mainly caused either by mutations in functional regions such as promoters, or by nonsense mutations within the genes of an operon. In this second case, the subsequent interruption of translation exposes the RNA molecule to interaction with the transcription termination factor rho, or allows the rest of the RNA molecule to stay folded in a way which prevents ribosomes from initiating translation of the downstream genes of the operon. As such, when designing a genome editing strategy, one should be particularly wary of mutations to be inserted in a WT operon. When such mutations have to be induced, one should avoid introducing premature stop codons (for example with a frameshift mutation or a nonsense mutation) in an upstream gene, for fear of affecting the downstream genes.

## I.4. Genome editing strategies in acetogens

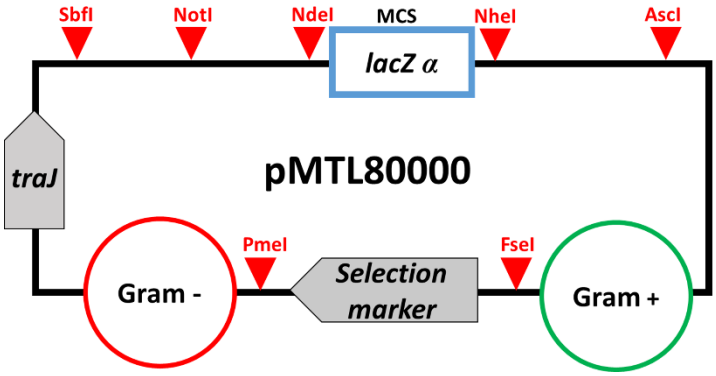
The goal of this thesis is to expand the genome editing toolbox in acetogens. Accordingly, having a good picture of the existing tools available, their advantages and their limitations is necessary both to understand what is currently achievable in this chassis and what could be improved upon.

### I.4.1. pMTL vector series

Most genome editing tool ultimately need to be expressed inside a cell from a DNA molecule, which often consists of a plasmid. For a given construct to be assembled and eventually expressed into *Clostridium*, this plasmid needs to be maintained in *E. coli* as well as in *Clostridium* – and it must thus possess two replication origins. It also requires a selectable genetic marker, and, at last, it must be built in a modular manner to facilitate assembly of any given construct. These vectors have been rigorously standardized and are well characterised, which mean that few aspects of them are still worth optimizing.

#### *a) pMTL80000 vector series*

The pMTL80000 vector series (Figure 19) is used as shuttle vector to clone constructs into *Clostridium*. Each vector of this series is composed of two replication origins (Gram+ and Gram-), one antibiotic resistance marker, one cloning site, and potentially an origin of conjugation transfer function (*traJ*) that allows conjugation of the vector from *E. coli* to *C. autoethanogenum.* Each part is delimited by different restriction sites, allowing the user to conveniently swap parts between two vectors using a simple restriction-ligation protocol [168]. Altogether, a basic pMTL8000 vector backbone is about 4kb long.

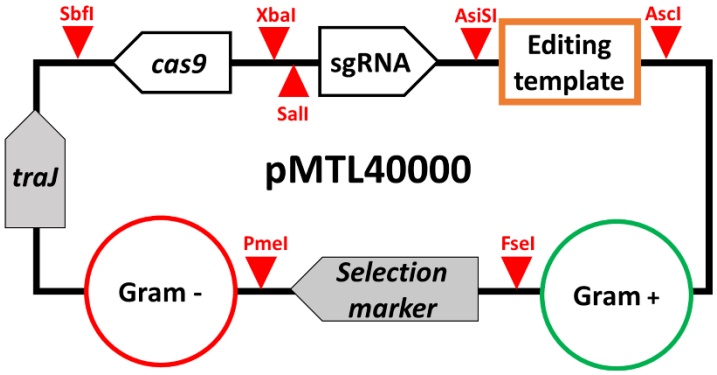



| Gram+ replicon | Selection Marker | Gram- replicon | Cloning site |
|---|---|---|---|
| 0. spacer | | | 0. spacer |
| | 1. catP | 1. p15a | 1. MCS |
| 2. pBP1 | 2. ermB | 2. p15a + tra | 2. $P_{thl}$ + MCS |
| 3. pCB102 | 3. aad9 | | 3. $P_{fdx}$ + MCS |
| 4. pCD6 | 4. tetA(P) | 4. ColE1 | 4. CatP |
| 5. pIM13 | | 5. ColE1+tra | |

**Figure 19: Structure and nomenclature of the pMTL80000 vector series.** Each vector is named by a number whose first digit is 8, and the last four digits are determined by the identity of each part of the vector. Adapted from [168].

### *b) pMTL40000 vector series*

The pMTL40000 vector (Figure 20) series is very similar to the pMTL80000 vector series. Its only difference is that the cloning site is occupied by a CRISPR/Cas9 system – *i.e.* a Cas9 expression cassette followed by a sgRNA expression cassette and an editing template. Once again, each locus is delimited by different restriction sites. Because of the large size of Cas9 and the size requirement for the donor DNA template, a pMTL40000 vector is generally longer than 10 kb.




| Gram + replicon | Selection Marker | Gram – replicon | sgRNA promoter | Cas9 promoter |
|---|---|---|---|---|
| 0. spacer | 1. catP | 1. p15a | 1. $P_{araE}$ | 1. $P_{thl}$ |
| 2. pBP1 | 2. ermB | 2. p15a + tra | 2. $P_{tcdB}$ | 2. $P_{fdx}$ |
| 3. pCB102 | 3. aad9 | 3. R6K | 3. $P_{fdx}$ | 3. $P_{ptb}$ |
| 4. pCD6 | 4. tetA(P) | 4. ColE1 | 4. $P_{j23119}$ | 4. $P_{tet}$ |
| 5. pIM13 | 5. kan | 5. ColE1+tra | | |
| 6. pIP404 | | 6. pSC101 | | |
| 7. pUB110 | | 7. pBBR1 | | |
| 8. p19 | | 8. pBBR1+tra | | |

**Figure 20: Structure and nomenclature of the pMTL40000 vector series.** Each vector is named by a number whose first digit is 4, and the last four digits are determined by the identity of each part of the vector. Adapted from [169].

### I.4.2. Clostron

The ClosTron tool is a genome editing tool based on a mobile group II intron from *Lactoccocus lactis*. This intron, *Ll.ltrB*, propagates into a genome in specific loci. This locus specificity is mediated by sequence homology with RNA intermediates, which can be re-programmed to target any sequence with *Ll.ltrB*. [170] It became possible to select for insertions of the group II intron into the genome by hijacking the intron with a selective marker disrupted by a self-splicing group I intron. When the re-programmed *Ll.ltrB* inserts itself into the genome with the selective marker cassette, the group I intron splice itself out – which allows expression of the selective marker. [171–173] In this fashion, any gene on the chromosome can be disrupted, and cargo up to 1kb can be inserted in the genome.

This tool – although effective – is far from ideal. Indeed, it leaves an antibiotic resistance gene at the locus of the gene disruption, requires marker recycling for multiple disruptions, can cause polar effects from the insertion of the intron, and is inappropriate for knocking in new genes. Furthermore, it can only target one locus at a time.

### I.4.3. Allele coupled exchange

Allele Coupled Exchange (ACE) [174] is a genome editing technology that relies on the HR machinery of the bacterial host to insert the sequence of a donor DNA into the chromosome. After transformation of the ACE plasmid into the cell, ACE breaks down the mutagenesis process into two steps in order to select for the mutant phenotype (cf. Figure 21): the first step selects for integration of the whole plasmid into the chromosome (editing template + backbone), and the second step for excision of the plasmid backbone from the chromosome, leaving the editing template within the chromosome.

The first step exploits a simple selective cassette such as an antibiotic resistance gene being expressed on a replication-deficient plasmid (*a.k.a.* pseudo-suicide vectors). As such, clones which have integrated the plasmid inside their chromosome will replicate the antibiotic resistance cassette more consistently, allowing them to grow faster than colonies which keep replicating their vector separately from the chromosome. This step is mediated by a long homology arm (>1 kb) to increase the likelihood that it occurs first [Figure 21(a) to (b)].

Subsequent excision of the plasmid out of the chromosome is achieved by selecting for a phenotype only achievable upon a second recombination event, with a second, smaller homology arm (~0.3 kb) at the other end of the homology cassette. For example, the editing cassette could harbour a promoterless antibiotic resistance gene which would only be

expressed upon recombination downstream of a chromosomal promoter; another example would be to knock-out or restore the function of a WT gene which is selectable [Figure 21(b) to (c)].

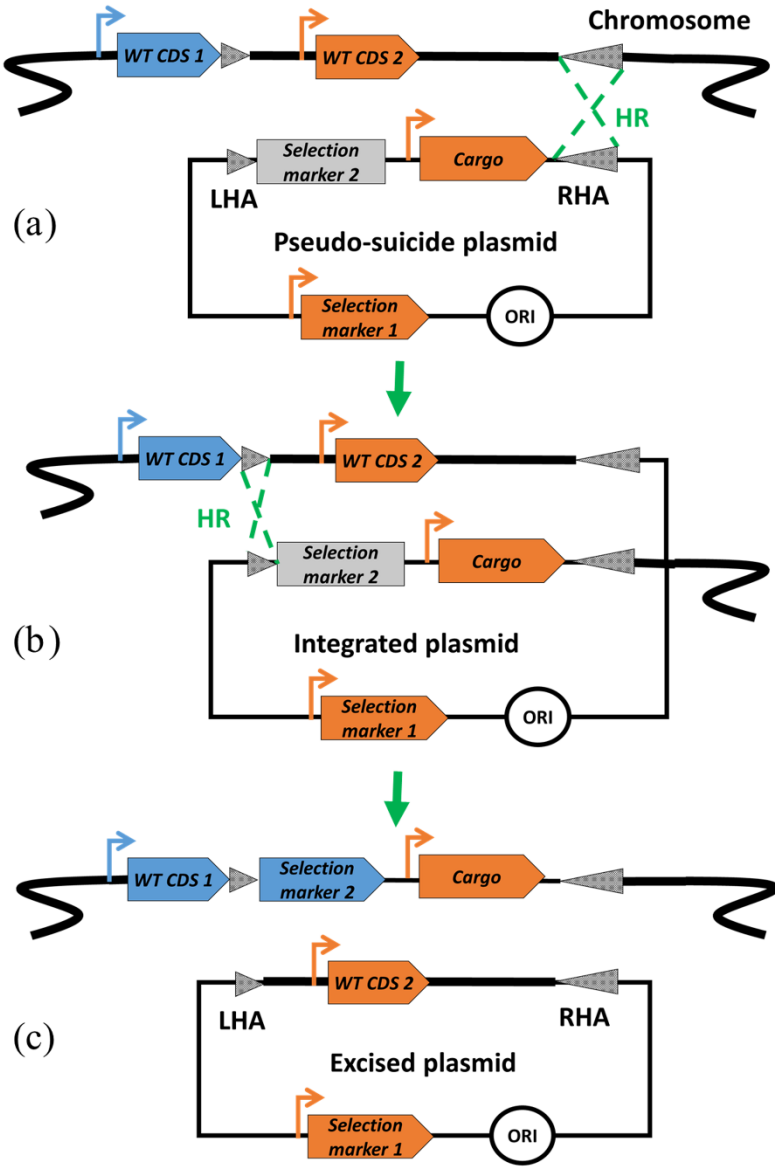



**Figure 21: Proposed mechanism of ACE mutagenesis. (a)** The first recombination event with the long homology arm (RHA) leads to **(b)** integration of the pseudo-suicide vector into the chromosome. Single-crossover mutants are selected by increased growth rate when selecting for the phenotype induced by the selection marker 1. At last, the second recombination event excises the plasmid from the chromosome **(c)** and activates the selection marker 2 (here, by placing it under the control of a constitutive WT promoter), which is used to select for the double-crossover mutants.

The final double-crossover mutant has thus integrated the editing template and lost the vector backbone. The only restriction lies in the locus of the recombination, which must be selectable during the second recombination step. If this second step is achieved through hijacking of a WT promoter by a selectable cassette, mutagenesis leaves a scar in the shape

of the selectable cassette. However, if this is done by targeting a WT selectable gene such as *pyrE*, a second ACE procedure can restore the WT *pyrE* allele and result in a scarless integration of heterologous DNA downstream of the *pyrE* locus. Indeed, *pyrE* is doubly selectable: *pyrE* strains are uracil prototroph but sensitive to 5-Fluoroorotic acid (FOA), while *ΔpyrE* strains are uracil auxotroph but resistant to FOA.

This last feature is exploited in a more streamlined formulation of this method [175] which enables knocking genes in and out anywhere in the genome without leaving a scar. It requires a custom *ΔpyrE* strain previously generated by standard ACE. A heterologous *pyrE* is expressed from the pseudo-suicide ACE vector backbone, which allows for the selection of the plasmid excision by loss of the ACE vector (*i.e.,* by loss of *pyrE*, which leads to resistance to FOA). Subsequently, the WT *ΔpyrE* locus can be restored by standard ACE procedure to restore the WT background.

One of the most frequent uses of ACE is to complement knocked-out genes at the *pyrE* locus while the chromosomal *pyrE* is being restored (Figure 22) [176], which makes the complementation strain more reliable (as it does not depend on expression on a vector). However, it is still a complex and long procedure which requires several selection and re-streaking steps. It requires a *ΔpyrE* strain, and the screening of colonies based on growth rate at the selection of the first recombination event can be misleading. Finally, it is not possible to target several loci at a time with this technique.

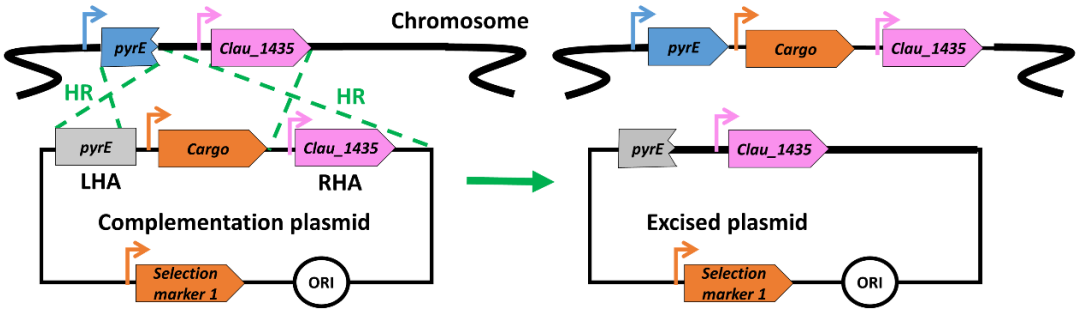



**Figure 22: ACE complementation in *C. autoethanogenum.*** After knocking out a gene in a *ΔpyrE* strain, the knocked out gene is returned to the knocked-out strain through double crossover at the *pyrE* locus. As during the standard ACE, the first recombination event is selected through the selection marker 1 on the plasmid backbone, but this time the second recombination event is selected through the restoration of the pyrE gene and growth on minimal medium (without uracil). The double-crossover mutant harbours the complemented gene in between pyrE and the downstream CDS *Clau_1435*.

## I.5. CRISPR-Cas genome editing

Over the last few years, genetic engineering techniques involving Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and RNA-guided CRISPR-associated nucleases (Cas) such as Cas9 have dramatically improved the effectiveness of genome editing [177,178]. This method exploits HR or NHEJ to produce mutants and selects the mutant phenotype by cutting the DNA of the WT allele.

### I.5.1. Mechanisms of Cas9-mediated HDR

More specifically, the Cas9 protein from *Streptococcus pyogenes* (SpCas9) induces a blunt-ended DSB between the 3$^{rd}$ and 4$^{th}$ base upstream of a protospacer associated sequence (PAM). This PAM consists of only three bases for SpCas9 (NGG), which allows targeting virtually any gene. Specificity to a particular locus is only mediated by a 20 bp homology region at the 5'-end of a non-coding, chimeric single-guide RNAs (sgRNA) [179]. This homology region can be easily customized to cause a DSB anywhere in the genome. This DSB can then be repaired by the NHEJ, leaving a scar that can disrupt gene expression and protects against further Cas9-mediated DSBs. Alternatively, if a donor DNA homologous to the locus of the DSB is available to the HR machinery, then a scar-free recombination will occur with great yield [146,180]. Indeed, if the recombination event mutates the genomic locus targeted by the Cas9-sgRNA complex, the cell will survive; however, if it fails to be repaired – or if it is repaired without mutation – then the resulting DSB will be lethal to the cell. CRISPR-Cas9 genome editing tools have thus no need of additional selectable marker for the genomic integration of the donor DNA: they induce their own selective pressure.

This tool is so effective that in many organisms, several loci can be targeted at a time (*a.k.a.* multiplex genome editing). In a striking example, 15 DNA parts were inserted *in vivo* into 3 different loci in one single transformation step, leaving no scar or marker in the yeast's genome [181]. It has also been successfully applied in prokaryotes [182], but it seems to be more toxic for them and it yields better results within an inducible system [183,184] or with an engineered Cas9 nickase (nCas9) [185,186] which only cuts a single DNA strand.

Cas9 mutagenesis is a single-step procedure that leads to scarless and markerless mutations anywhere in the genome. Its multiplexing is straightforward (it only requires expression of additional sgRNAs) and works on any genetic background (no need for a *ΔpyrE* strain). It is thus a very promising candidate for our genetic engineering ambitions. However, even using Cas9, no more than one locus is usually targeted at a time in *Clostridium* [84]. Indeed, the

*Clostridium* genus does not possess a NHEJ pathway, and it also suffers from a poor HR efficiency compared to more amenable organisms such as *Saccharomyces cerevisiae* [174,178,186]. This greatly reduces our ability to genetically engineer *C.autoethanogenum* using CRISPR-Cas9.

Because CRISPR/Cas systems are central to this thesis, the next section will give a broader insight into the history and mechanism of known CRISPR/Cas systems. The following sub-sections are heavily based on the review written by Barangou et al. in 2014 [187].

#### *I.5.1.1) Discovery of CRISPR-Cas systems*

The CRISPR locus has been described for the first time in *E. coli* in 1987 by Ishino et al. as "*Five highly homologous sequences of 29 nucleotides […] arranged as direct repeats with 32 nucleotides as spacing*." [188,189]. This description of the locus was later abbreviated as TREPs (Tandem Repeats), SRSR (Short Regularly Spaced Repeats) and finally CRISPR (Clustered Regularly Interspersed Palindromic Repeats) when structurally similar loci with associated proteins were found to be widespread in both archaeal and bacterial genomes [190,191]. An example of such a locus has been given in Figure 23. The structural homology of these loci across so many phyla hinted at an ancient origin and an important role in gene biology. This discovery led to increased scrutiny in the function of the locus and associated proteins. Eventually, after several independent reporting of viral sequences constituting the spacers between the palindromic repeats of the CRISPR locus [192–194], it was hypothesized that CRISPR-Cas systems were involved in immunity against mobile genetic elements through a mechanism similar to the eukaryotic RNA interference system [195].

tgatgggtttgaaaatgggagct
**gGGagTtctaCCGCaGagGCGGGGGAact**ccaagtgatatccatcatcgcatccagtgcgcc
**CGGTTTATCCCCGCTGatGCGGGGAACaC**cagcgtcaggcgtgaaatctcaccgtcgttgc
**CGGTTTATCCCtGCTGGCGCGGGGAACTC**tcggttcaggcgttgcaaacctggctaccggg
**CGGTTTATCCCCGCTaaCGCGGGGAACTC**gtagtccatcattccacctatgtctgaactcc
**CGGTTTATCCCCGCTGGCGCGGGGAACTC**ccgggggataatgtttacggtcatgcgccccc
**CGGTTTATCCCCGCTGGCGCGGGGAACTC**tgggcggcttgccttgcagccagctccagcag
**CGGTTTATCCCCGCTGGCGCGGGGAACTC**aagctggctggcaatctctttcggggtgagtc
**CGGTTTATCCtCGCTGGCGCGGGGAACTC**tagtttccgtatctccggatttataaagctga
**CGGTTTATCCCCGCTGGCGCGGGGAACTC**gcaggcggcgacgcgcagggtatgcgcgattcg
**CGGTTTATCCCCGCTGGCGCGGGGAACTC**gcgaccgctcagaaattccagacccgatccaaa
**CGGTTTATCCCCGCTGGCGCGGGGAACTC**tcaacattatcaattacaaccgacagggagcc
**CGGTTTATCCCCGCTGGCGCGGGGAACTC**agcgtgttcggcatcacctttggcttcggctg
**CGGTTTATCCCCGCTGGCGCGGGGAACTC**tgcgtgagcgtatcgccgcgcgtctgcgaaag
**CGGTTTATCCCCGCTGGCGCGGGGAACTC**tctaaaagtatacatttgttcttaaagcatttt
ttcccataaaaacaacccaccaaccttaatgtaacatttccttattattaaagatcagctaa
ttctttgtttcta

**Figure 23: CRISPR locus of *E. coli* K12 DH10B (CP000948.1).** Red [ ] and dark-red [ ] colours highlight the stop codons of neighbouring genes (*iap* and *ygbf*). The clustered repeats are represented with bold capital letters, and the palindromic sequences within those repeats are highlighted in yellow [ ]. Mismatches within the clustered repeats are not capitalized.

#### *I.5.1.2) CRISPR/Cas systems are prokaryotic heritable immune systems*

CRISPR loci all share a common structure. They start with a leader sequence which generally includes a transcriptional promoter, followed by the aforementioned clustered, regularly interspersed palindromic repeats of around 21-40 bp [190,191]. In between those repeats are spacers of 15-60 bp [191,195] that often match the sequence of mobile genetic elements [192–194]. In a mobile genetic element, the sequence that matches one of the spacers of a given CRISPR locus is called a protospacer. Finally, CRISPR loci end with a transcriptional terminator. The transcript of the CRISPR locus, between the leader sequence and the terminator, is called pre-CRISPR RNA or pre-crRNA.

New spacers are integrated from mobile genetic elements into the CRISPR locus in a process called spacer acquisition. After transcription of the CRISPR locus, the resulting pre-crRNA, composed of all the spacers that have been previously acquired, is processed into several mature CRISPR RNAs (crRNA), one for each spacer. The process of pre-crRNA transcription and maturation is called crRNA biogenesis. At last, specific endonucleases use the crRNAs to target the cognate protospacer and cleave it, in a process known as interference. As a result, the host of a functional CRISPR locus gains immunity against all mobile genetic elements that contain a protospacer matching one of the host's spacers. Since the CRISPR locus is part of

the host's genome, all the progeny of the host will also benefit from the same immunity. The proteins involved in spacer acquisition, crRNA biogenesis and interference are collectively named CRISPR associated proteins (Cas); their coding sequence is usually located near the CRISPR locus.

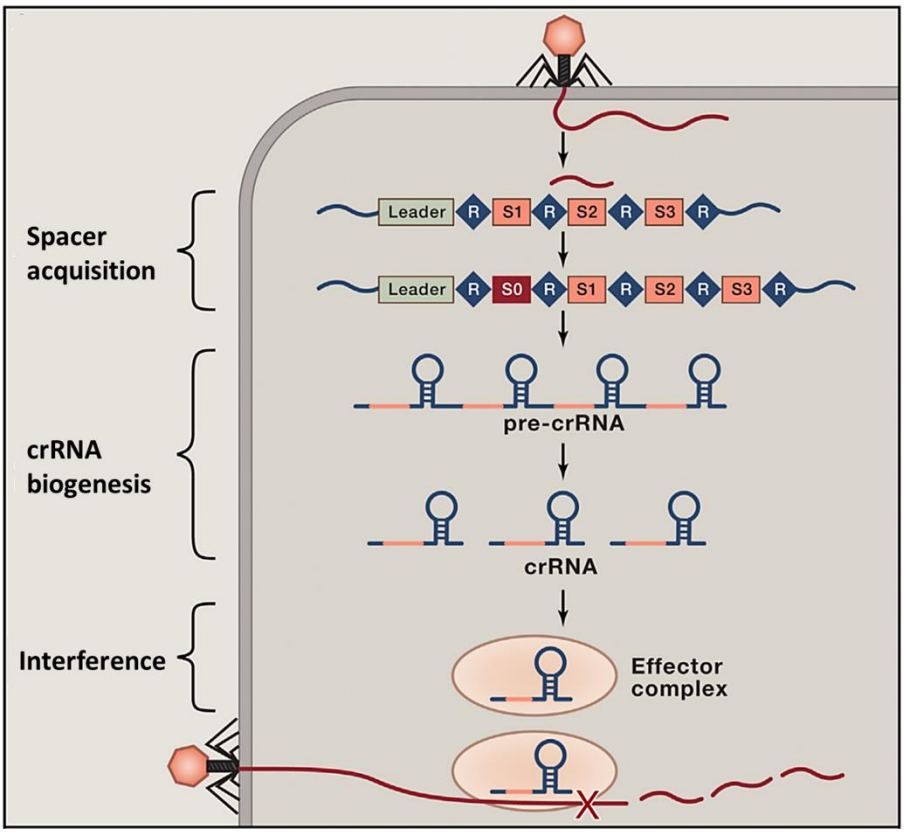



**Figure 24: Overview of the CRISPR-mediated immune system.** During spacer acquisition, a short sequence of foreign DNA is integrated into the CRISPR locus. Afterwards, functional CRISPR RNA is be generated from the transcript of the CRISPR locus during crRNA biogenesis. At last, the crRNA is used by an effector complex (e.g., Cas9-tracrRNA) to detect and cleave a foreign DNA sequence homologous to that of the crRNA spacer. R= short palindromic repeat; S=Spacer. Modified from Wright et al., 2016 [196].

Several types of CRISPR systems have been identified, each characterized by a specific set of Cas proteins that carry out CRISPR interference in their own way (cf. Table 2).

**Table 2: Brief overview of the different types of CRISPR/Cas systems and their associated effectors at each stage of CRISPR immunity** [196,197]**.**

| | *Type I* | *Type II* | *Type III* | *Type IV* | *Type V* | *Type VI* |
|---|---|---|---|---|---|---|
| ***Spacer Acquisition*** | Cas1,2,4 | Cas1,2,4 | Cas1,2 | DinG | Cas1,2,4 | Cas1,2 |
| ***crRNA biogenesis*** | Cas6 | Cas9, tracrRNA, RNase III; Cas12a | Cas6 | | | |
| ***Interference*** | Cas3,5,6,8 | Cas9, tracrRNA, crRNA; Cas12a,crRNA | Cas10, Csm2,3,4,5 | Csf1,2,3 | Csf2 | C2c2 |

Cas9 is the most commonly used Cas protein to mediate genome editing, because of the simplicity of its mechanism of interference. It only requires the expression of one protein (Cas9) and two RNA molecules (tracrRNA, crRNA) in its native state. In 2012, Jinek et al. engineered a chimeric single-guide RNA to fuse the crRNA and the tracrRNA in a single transcript which conserves (or even improves) the function of the tracrRNA:crRNA duplex, making the system even more convenient [179]. Since then, many genome editing initiatives have been reported which exploited various engineered versions of Cas9 with altered properties [198–203]; different Cas interference proteins such as Cas12a (previously Cpf1) – a Cas interference protein which does not even require a tracrRNA and which recognizes a T-rich PAM [204–207]; or endogenous CRISPR-Cas systems to avoid expressing an heterologous Cas protein [208,209].

### I.5.2. gRNA design

The most basic requirement to design a functional gRNA for Cas9-mediated mutagenesis is that the sequence that it targets should be 20 nt-long and preceded by a PAM (NGG). However, two additional criteria are considered by most protospacer prediction algorithms to design gRNAs: the on-target efficiency, or the probability that a gRNA will target a given protospacer sequence, and the off-target efficiency, or the probability that the same gRNA will not target another protospacer somewhere else in the genome. To be a good gRNA target, a protospacer must have a good on-target score and a good off-target score. It must be disclosed, however, that these scores are not entirely reliable: a gRNA with a terrible on-target score might still be functional [210], and a gRNA with a perfect off-target score could still cause off-target mutations [211]. For this reason, these scores were rarely given much weight in gRNAs design. Nonetheless, for the interested reader, the next two subsections briefly summarize how such scores are calculated, which helps understanding their usefulness and their limitations.

#### *I.5.2.1) On-target score*

On-target scores are based on the statistical analysis of large-scale mutagenesis experiments which exploit libraries of more than a thousand gRNAs to knock-out a few reporter genes by leveraging NHEJ DNA repair mechanisms, often in mammalian cells. The effectiveness of individual gRNAs is assessed by measuring their relative enrichment in a pool of cells which exhibit the mutant phenotype [212], by measuring their relative depletion in a pool of cells which exhibit the WT phenotype [213], or by directly sequencing the target sites [214,215]. Such large-scale analyses allow to test a number of hypothesis on optimal gRNA design.

The following parameters have been shown to contribute to the ability of gRNA to knock out genes through NHEJ in these experiments: the GC-content, the distance between the protospacer and the Transcription Start Site (TSS), the location of the protospacer in an untranslated region (UTR), an intron or an exon, the particular composition of the sequence of the PAM and the spacer, the melting temperature of the gRNA, or even the DNA strand targeted by the gRNA [212,213,216]. The resulting datasets can also be used to train algorithms without guiding them with preliminary hypothesis, and are being used to evaluate the impact of epigenetic features such as the presence of chromatine on gRNA effectiveness [217].

Some of the most straightforward gRNA design principles which have emerged across several such experiments are to maximise gRNA stability by having a guanine directly upstream of

the PAM [210,212], to avoid stretches of four identical nucleotides in the protospacer (especially guanines) [218], and to avoid extreme GC contents (<30% or >75%) [212].

Nonetheless, the scores based on these gRNA libraries have a certain number of inherent biases: they are often made in mammalian cells, they rely on eukaryotic genetic and epigenetic features, they often only register mutations which produce a measurable phenotype, they depend on the NHEJ machinery and they are affected by the particular experimental context of the mutagenesis [216,219–221]. The phenotypic bias can be addressed by proceeding to high-throughput sequencing of the whole pool of cells without preliminary selection, although the resulting scores might not predict phenotypical changes as effectively [214–216].

#### *I.5.2.2) Off-target score*

Off-target score calculations are often based on studies which test the effectiveness of gRNAs which have one or several mismatches with their target site. The target sites are sequenced after transfection of the mismatched gRNAs and the Cas9 vectors, and the proportion of mutated to WT target sites estimates the effectiveness of each gRNAs relative to a gRNA without any mismatch [213,222–224]. Once again, these studies rely on NHEJ and are often made in mammalian cells. As previous studies [179,225,226], they showed that the first 8 to 14 nt from the PAM are particularly sensitive to single nucleotide mismatches [182,222]. This phenomenon has led to identify the first 12 nt from the PAM as the “seed” or “core” sequence, particularly relevant to some off-target scoring algorithms [227]. Some of the criteria which were shown to affect gRNA specificity in these studies are: the nature and number of the mismatched bases, their clustering (contiguous or interspaced), their position in the protospacer, whether there are mismatches in the PAM and the likelihood of bulges in the RNA-DNA duplex [213,216,222]. By applying a multivariate regression algorithm on these datasets, it became possible to estimate the likely effectiveness of a given gRNA on an arbitrary target site [216].

Other studies use a different approach to measure off-target mutations caused by Cas9. Several next-generation sequencing methodologies are being exploited to directly measure the off-target mutations induced by Cas9 and other nucleases [224]. For example, GUIDE-Seq [211] (Genome-wide, unbiased identification of DSBs enabled by sequencing) relies on the integration of rationally designed double-stranded oligodeoxynucleotides (dsODNs) into a target genome by hijacking the NHEJ machinery. By design, these dsODNs should integrate only in loci which underwent DSB repair. The ODN sequences can then be targeted by next-

generation sequencing, and their abundance and distribution in the genome can be identified. When coupled with Cas9-mediated mutagenesis, this method allows to detect all the loci that have been targeted by Cas9: on-target and off-target sites alike. Using GUIDE-seq and other similar methods, off-target sites in human cells were found with as many as 7 mismatches mostly concentrated at the 3' end of the target site [211,228], with non−NGG PAMs [211], or with "bulge" mismatches [211,229]. Some off-target sites were even found to be targeted more effectively than the actual target site [211].

Similarly to the on-target calculation strategy, the genome-wide data obtained through experiments such as GUIDE-seq have been exploited to train machine-learning algorithms to rank potential off-target sites based on many more parameters such as gRNA melting temperature, methylation pattern, wobble base-pairing, chromatin structure, local DNA geometry, and GC content [230–232]. This last generation of off-target scoring algorithm seems to be performing the best, but are very computationally intensive and are relying on an abundance of available data. Consequently, they are difficult to use and lack flexibility (precomputed results and/or limited target genomes).

When evaluating the off-target efficiency of a given gRNA, scoring algorithms align the target protospacers with the rest of the genomes and identify the sites which are most likely to be targeted by a given gRNA based on the criteria described above. The number and the quality of these potential off-target sites are then weighted to produce a score between 0 and 100, where 100 means that a given protospacer is not expected to produce any off-target mutations if targeted.

Unfortunately, many of the biases inherent to on-target scoring algorithm still apply. Off-target scoring is eukaryotic-centric, relies on NHEJ, and can vary wildly with the experimental context. Even within this narrow frame, the scores computed by the most popular off-target scoring algorithms are only modestly correlated with actual off-target activity [216,219,230].

Some of the most straightforward guidelines to reduce off target effects (at least in human cells) are to avoid protospacers which have less than 4 mismatches with potential off-target sites [224,227,233] – or three mismatches in the seed region [227]; to use truncated gRNAs which only target the 17 or 18 nt closest to the PAM [234]; to flank the gRNA with a guanine dinucleotide (GG) at its 5' end [224], or to induce DSB using an engineered Cas9 nickase and two gRNAs for each target site [235].

## I.6. Objective and aims

The aim of this thesis is to develop a selection of cutting-edge genome editing tools in order to facilitate the engineering and the study of clostridial species – and of the gas-fermenting *C. autoethanogenum* in particular.

The first tool developed, described in Chapter III, consists of a range of genomic "bookmarks" relying on CRISPR-Cas9 HDR mutagenesis, which can be used to streamline complementation studies in any organism. In Chapter IV, a problematic decline in the performance of the conjugation protocol being used was investigated. This resulted in an optimized, more reproducible conjugation protocol, as well as several surprising new insights into some experimental parameters. Having improved the conjugation protocol, the application of the technology of base editing to *C. autoethanogenum* was investigated in Chapter V. First as a proof of concept, targeting a single gene; and then, in Chapter VI, in combination with several multiplex genome editing strategies. The effectiveness of the tool was characterised and different approaches to improve its effectiveness were explored.

It is hoped that these new tools will help understanding and engineering gas fermenting organisms to improve their industrial performances, and thus contribute to mitigating humanity's environmental footprint while reducing the UK's dependence on fossil resource.

# Chapter II: Material and methods

## II.1. Strains

### *II.1.1. Clostridium autoethanogenum*

This thesis focused exclusively on *Clostridium autoethanogenum* DSM10061 (NCBI CP012395.1). In addition to the WT, two mutant strains of *C. autoethanogenum* were also used as chassis or controls (*ΔpyrE* and *ΔpyrE::ACE*) and three other strains were generated: *ΔpyrE::BMa*, *adh1(Q243X) adh3(W90X)*, and *adh1(Q243X) adh3(W90X) CLAU_1794(S251L, W278X)* (cf. Table 3). Refer to Chapter IV.3.1.3) for additional insight into the genotype of *ΔpyrE::BMa* and *ΔpyrE::ACE.*

**Table 3: Summary of all *C. autoethanogenum* strains used in this study.**

| Label | Name | Genotype | Mutagenesis vector | Source |
|---|---|---|---|---|
| **cFS01** | CA_DSMZ_10061 | *WT* | N/A | DSMZ |
| **cFS02** | CA_Δ*pyrE* | *ΔpyrE* | pMTL431511 -CLAU-*pyrE* | [169] |
| **cFS03** | CA_Δ*pyrE*::BMa | *ΔpyrE::BMa* | vFS42 | This study |
| **cFS04** | CA_Δ532Δ534 | *adh1(Q243X), adh3(W90X)* | vFS50 | This study |
| **cFS05** | CA_Δ532Δ534 Δ1794 | *adh1(Q243X), adh3(W90X), CLAU_1794(S251L, W278X)* | vFS74 | This study |
| **cFS06** | CA_ACE | *ΔpyrE::ACE* | pMTL431511 -*pyrE*_ACE | Christopher Humphreys |

### II.1.2. Escherichia coli

Cloning and vector methylation was undertaken in two *Escherichia coli* K12 strains: DH5α (F – Φ80 *lac*Z ΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rK–, mK+) *pho*A *sup*E44 λ– *thi*-1 *gyr*A96 *rel*A1) from New England Biolabs (NEB, Hitchin, UK), and sExpress, respectively. *sExpress* is the Express strain of *E. coli* (*huA2 [lon] ompT gal sulA11 R(mcr-73::miniTn10--*Tet$^S$*)2 [dcm] R(zgb-210::Tn10--*Tet$^S$*) endA1 Δ(mcrC-mrr)114::IS10)* from NEB transformed with the R702 vector from *E. coli* CA434 [236]. R702 is a large conjugative plasmid that mediates the mobilisation of small co-resident vectors from the *E. coli* donor into the clostridial recipient cell.

## II.2. Media and culture conditions

*E. coli* was grown aerobically in LB liquid medium at 37°C and 250 rpm. *C. autoethanogenum* was grown on YTF medium, at 37°C and 0 rpm under an atmosphere of 80% $H_2$, 10% $CO_2$ and 10% $N_2$ in an anaerobic cabinet (Don Whitley Scientific Ltd, Bingley, UK). Plates were made by adding 20 g/L of agar to the LB or YTF medium before autoclaving. Concentrations of supplements such as antibiotics or inducer molecules used with *E. coli* or *C. autoethanogenum* are summarized in Table 4.

**Table 4: Medium supplements with associated working concentrations.** Chloramphenicol and Thiamphenicol both select for the same resistance gene, *catP*, but one is used as selective pressure in *E. coli* and the other in *C.autoethanogenum*.

| Supplement | Stock (mg/ml) | | Working concentration (µg/ml) | |
|---|---|---|---|---|
| | Solvent | Concentration (mg/mL) | *E. coli* | *Clostridia* |
| **Ampicillin** | $ddH_2O$ | 50 | 100 | 100 |
| **Chloramphenicol** | Ethanol | 25 | 12.5 | - |
| **Erythromycin** | $ddH_2O$ | 50 | 500 | - |
| **Clarithromycin** | 70% Ethanol | 6 | - | 6 |
| **Kanamycin** | $H_2O$ | 50 | 50 | - |
| **Spectinomycin** | $H_2O$ | 125 | 250 | - |
| **Thiamphenicol** | 50% Ethanol | 15 | - | 7.5 |
| **IPTG** | $H_2O$ | 238 | 238 | 238 |
| **D-cycloserine** | $H_2O$ | 50 | 250 | - |
| **5-Fluoroorotic acid (light sensitive)** | DMSO | 100 | - | 1000 |
| **Uracil** | $H_2O$ | 0.4 | - | 20 |
| **Theophylline** | DMSO (55°C) | 45 | - | 900 |

### II.2.1. LB-broth, Miller

The powder was purchased directly from Sigma-Aldricht (Merck KGaA, Darmstadt, Germany). 25g of powder dissolved in 1 L of $ddH_2O$ results in a final concentration of 10 g/L NaCl, 10 g/L Tryptone, 10 g/L and 5g/L Yeast Extract.

### II.2.2. Yeast-tryptone-Fructose (YTF) medium

For 1L liquid media:

- 10g yeast extract
- 16g tryptone
- 10g fructose
- 0.2g NaCl
- 1mL 1000x vitamin stock solution [cf. Figure 25(a)]
- 1ml acidic 1000x trace element solution [cf. Figure 25(b)]
- 1mL basic 1000x trace element solution [cf. Figure 25(c)]
- pH adjust to 5.8 using HCl
- Autoclave

To pre-reduce, incubate plates in the anaerobic cabinet for >4h (overnight if more than 2 plates). For liquid YTF or PBS, incubate in anaerobic cabinet for 4 days with lids loosened.

**1000x B-vitamin solution**

| | weighed 1000x mg/500mL (1000x) | Bart Vit1 1000x mg/L (1000x) | Pfennig 100x mg/L (100x) |
|---|---|---|---|
| p-aminobenzoate | 57 | 100 | 5 |
| riboflavin | 52 | 100 | 5 |
| thiamine | 100 | 200 | 5 |
| nicotinate | 103 | 200 | 10 |
| pyridoxin | 255 | 500 | 5 |
| pantothenate (calcium) | 52 | 100 | 5 |
| cyanocobalamin | 39 | 100 | 2 |
| d-biotin | 11 | 20 | 2 |
| folate | 24 | 50 | 2 |
| lipoate/thioctic acid | 25 | 50 | 5 |
| 2-mercapto sulfonic acid | - | - | 2 |

final pH 7.8; store 4 °C

- dissolve p-aminobenzoate separately by titrating with 5 M NaOH until dissolved

- dissolve riboflavin separately at 50 °C (0.012% solubility)

- dissolve lipoate separately by titrating with 5 M NaOH until dissolved

(a)

| **1000x Acidic trace element solution (per L)** |
|---|
| 50 mM HCl |
| 100 mg $H_3BO_3$ |
| 230 mg $MnCl_2.4H_2O$ |
| 780 mg $FeCl_2.4H_2O$ |
| 103 mg $CoCl_2.6H_2O$ |
| 602 mg $NiCl_2.6H_2O$ |
| 78 mg $ZnCl_2$ |
| 50 mg $CuSO_4.5H_2O$ |
| 50 mg $AlK(SO_4)_2.12H_2O$ |

(b)

| **1000x Basic trace element solution (per L)** |
|---|
| 10 mM NaOH |
| 58 mg $Na_2SeO_3$ |
| 53 mg $Na_2WO_4$ |
| 52 mg $Na_2MbO_4.2H_2O$ |

(c)

**Figure 25: Composition of (a) the 1000x vitamin solution, (b) the 1000x acid trace solution and (c) the 1000x base trace solution.**

## II.3. Colony counting

### II.3.1. Conjugation plates

Whenever possible, the conjugation protocol was adapted to count in between 50 and 250 colonies per plate. When the number of colonies on the same plate exceeded 500, the software OpenCFU 3.9 was used [237] and parameters were adjusted manually to capture all colonies and avoid false positives.

### II.3.2. 3x6 drop-count method

In order to estimate the number of live cells within *C. autoethanogenum* or *E. coli* sExpress cultures prior to mating, a 3x6 drop count method was used [238]. A multichannel pipette was used to carry out serial 10-fold dilutions (20 µL into 180 µL PBS) from a 250 µL sample in six replicates in a 96-wells plate. Using the multichannel pipette, one 10 µL droplet from each replicate at three different dilutions are then deposited on a plates with no antibiotics and incubated in appropriate conditions.

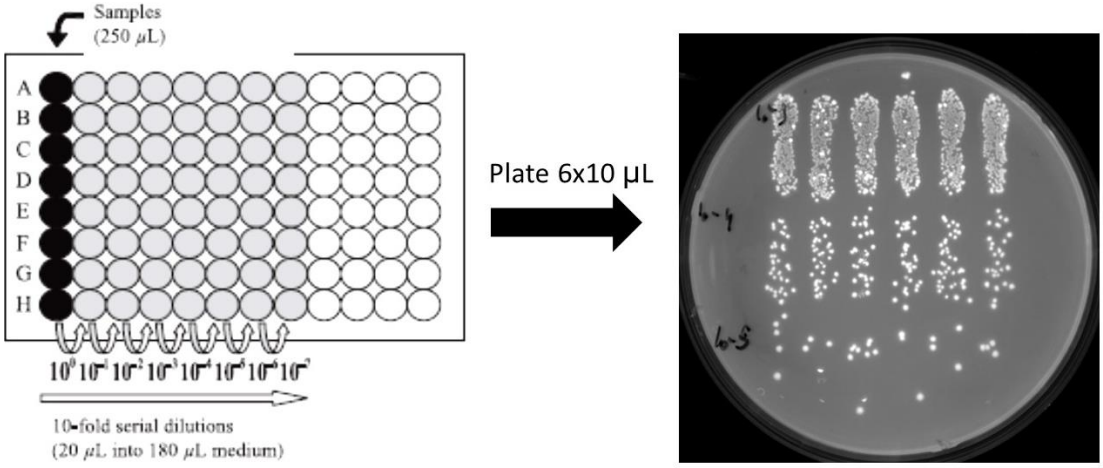



**Figure 26: Schematic representation of the 6x6 drop count experimental protocol, illustrated with the plate obtained from a liquid culture of *C. autoethanogenum* at $OD_{600}$~0.4.** Spreading cell dilutions in this manner saves time, medium and space, all of which are precious when working in a gas cabinet. Modified from Chen et al., 2003 [238].

The average number of individual colonies across all droplets of the dilution which exhibits in between 20 and 70 colonies per droplet can be used to extrapolate the number of live cells in the original solution. The number of cells in the original sample is calculated by multiplying the average number of colonies counted with the dilution factor, and adjusted for the volume of each droplet relative the volume of the sample (Equation 1).

$$CFU_S = D * \overline{CFU}_d * \frac{V_S}{V_d}$$

**Equation 1: Calculation of the number of Colony-forming units (CFU) with the 3x6 drop count method.** $CFU_S$= number of colony forming units in the sample; D= dilution factor (>1); $\overline{\boldsymbol{CFU}}_{\boldsymbol{d}}$= average number of colonies across six droplets; $V_S$= Volume of the sample; $V_d$= Volume of the droplet.

In practice, $OD_{600}$~0.4 cultures of sExpress and *C. autoethanogenum* often showed the best resolution when diluted $10^5$-fold and $10^4$-fold, respectively. In each case, the 10-fold dilution above and below the expected best-resolved dilution were plated.

### II.3.3. Data analysis

All statistical analysis are done using Graphpad Prism 9.

Live cell count of donors, recipients and transconjugants are quantitative, discrete and strictly equal or superior to zero. As such, they should follow a Poisson distribution (Equation 2).

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}, \qquad k = 1,2,3, \dots$$

**Equation 2: Poisson distribution applied to colony counts in CFU.** *k* is the colony count of one plate (in CFU), μ is the average live cell count, and P(*X*= *k*) is the probability that a sample of a population of value μ gives a live cell count of *k*.

Poisson distributions are entirely defined by a single parameter (μ), and, as such, are defined as parametric distributions.

According to the Central Limit Theorem, at high sample size (k>30), parametric distributions can be approximated by a normal distribution. Because colony-forming units (CFU) are typically counted on plates in between 50 and 250 CFU, a normal distribution can be assumed for all CFU measurements and parametric tests can be used.

However, in a Poisson distribution, the variance is equal to the mean. Consequently, when testing the difference between two Poisson means, their variances cannot be assumed equal (since that would also assume equal means and thus defeat the purpose of the test). Consequently, statistically significant difference between two colony counts will be tested using Welch's t-test: a parametric test which does not assume equal variances.

## II.4. Cloning and assembly

Unless otherwise specified, all kits, enzymes and buffers were purchased from NEB. Unless otherwise specified, the instructions of the provider were followed. The primers, vectors, and gblocks are summarized in the supplementary materials (Table S. 6, Table S. 7, and Table S. 8). *In silico* assembly was done with APE - A Plasmid editor v2.0.53, sequence alignment and analysis of sequencing results with Benchling (San Francisco, USA).

### II.4.1. Protocols

#### *II.4.1.1) Restriction*

Check beforehand that the enzyme is compatible with this standard protocol.

- 1 µL NEB Cutsmart buffer.
- 1 µL of each restriction enzyme.
- Top up the reaction mix to 10 µL with DNA template (in between 0.5 µg and 4 µg)
- Incubate at 37°C for 1h (<1 µg) or overnight (2-4 µg).
- If digestion product is a vector backbone, dephosphorylate:
    - Add 1.2 µL of Antarctic phosphatase buffer.
    - Add 1 µL of Antarctic phosphatase.
    - Incubate at 37°C for at least 1h.
- Heat inactivate at 80°C for 20 min.
- Gel purify using NEB gel extraction kit.

#### *II.4.1.2) Ligation*

- Calculate the mass of insert required to meet a 3:1, 7:1 and 10:1 molar ratio. http://nebiocalculator.neb.com/#!/ligation

Do three ligations in parallel, so that at least one of the ratios works out.

- 50 ng of cut and dephosphorylated vector
- x ng of insert
- 0.5 µL of T4 DNA ligase
- 1 µL of T4 DNA buffer 10x
- Top up to 10 µL with ddH2O
- Incubate at 16°C overnight, or at 4°C over the weekend
- Heat kill 10 min at 65°C
- Transform 5 µL of each ligation mix alongside 25 ng of cut vector as a negative control to estimate the number of false positives on the transformation plates.

#### *II.4.1.3) PCR amplification*

##### II.4.1.3)(1) Template preparation

###### (a) Colonies

- Dilute one colony into 40 µL of autoclaved ddH2O. The osmotic shock bursts many cells and releases their DNA in the water. Save the tubes as it will later be used to inoculate positive colonies into overnight cultures. Use PCR tubes in strips or in 96-wells plates so as to facilitate the use of multichannel pipettes.
- For *C. autoethanogenum* colonies: heat shock at 98°C for 10 min to facilitate DNA seeping into the water, then centrifuge 30 sec to pellet the cells debris. Use the supernatant as DNA template.

###### (b) Pure DNA

- Use 1 µL of purified or un-purified PCR product.
- Use only 0.1 µL of purified vectors (or just wash your tip in the purified vector solution) so as to avoid seeing a band for the purified vector in the electrophoresis gel of your PCR product.

##### II.4.1.3)(2) Primer design

- Annealing length >18 nt (up to >16bp if no overhang). Overhang up to a total length of 100 nt or 120 nt depending on the DNA synthesis company. Primers were synthesised either by Integrated DNA Technologies (IDT) Ltd (Sheffield, UK) or Sigma-Aldricht. Especially if used among repetitive sequences or genomic DNA, make sure your primer is specific (unique) to your DNA template.
- Calculate melting temperature (Tm) of the annealing region of the primer on https://tmcalculator.neb.com/#!/main (NEB Tm calculator v1.13.0), using Q5 High-fidelity DNA polymerase as baseline. Try to aim for a Tm of 65°C to have margin when troubleshooting the annealing temperature (Ta). The Tm of two primers used together should be within 5°C of each other.
- Check for secondary structures on http://www.unafold.org/mfold/applications/dna-folding-form.php [239]. If $\Delta G<0$ or close to 0 at the appropriate Ta: no secondary structure within the annealing region – especially not at the 3'-end of the primer. Secondary structures in the overhang are OK.

#### II.4.1.3)(3) PCR mix

- 0.1 µL of each primer (100 µM stock).
- 0.1-1 µL of DNA template.
- Top up to 10 µL with autoclaved dd$H_2O$.
- 10 µL of either:
    - 2X Q5 DNA polymerase master mix (for parts amplification, sequencing or *C. autoethanogenum* colony PCR).
    - 2X Dreamtaq or OneTaq DNA polymerase master mix (for *E. coli* colony PCR).

#### II.4.1.3)(4) PCR program

1. 98°C for 5 min (denaturation; 94°C for OneTaq).
2. 98°C for 30 sec (denaturation; 94°C for OneTaq).
3. Tm-5°C for 30 sec (annealing).
4. 72°C for x min (elongation, x=min/kb; 68°C for OneTaq).
5. Repeat Steps 2 to 4 (25-35 cycles).
6. 72°C for x+2 min (final elongation; 68°C for OneTaq).
7. Hold at 15°.

#### II.4.1.3)(5) Gel electrophoresis

- Cast a 1% (m/v) agarose gel in TAE buffer (Tris base, acetic acid, and EDTA). Boil to dissolve the agarose. If the DNA samples are expected to be under 0.5 kb, you can use a 2% (m/v) agarose gel for better resolution.
- Add 0.5-1 µL of 10,000x SYBR™ Safe DNA gel stain to 25 mL to 150 mL of molten agarose gel, depending on the size of the gel.
- Add a comb to carve wells during solidification of the gel and incubate for 10-30 min at room temperature.
- Add DNA ladder (Generuler 1 Kb Plus DNA ladder, Thermo Fisher Scientific) in at least one of the wells, alongside 3-5 µL of each sample complemented with loading dye in the remaining wells. For gel purification, add all of your sample in wide wells.
- Run at 90V for 40 min in TAE buffer, with samples starting at the cathode (-) and running towards the anode (+).
- Capture gel image under blue light.

### *II.4.1.4) Hifi assembly*

- Design primers so that each part overlaps with 25 bp to 40 bp of the adjacent parts.

- Amplify/digest all parts.
- Gel purify all parts.
- 50 ng of cut and dephosphorylated vector.
- 0.5-1 µL of each other parts (up to 7 parts). If possible, try to keep an equimolar ratio. If parts are smaller than 200 bp, use a 5x molar excess.
- Top up to 5 µL with $ddH_2O$
- 5 µL NEBuilder® HiFi DNA Assembly Master Mix.
- Incubate at 50°C for 1h to 4h.
- Transform 5 µL of each HiFi reaction mix alongside 25 ng of cut vector as a negative control to estimate the number of false positives on the transformation plates.

For large assemblies, it is sometimes easier to amplify the HiFi mix in order to reduce the number of parts before proceeding to a second HiFi with the resulting amplicon and the target vector backbone.

For sgRNA exchange or similar exchange of short sequences, you can use a single primer which overlaps with both extremities of the vector to be changed. The allele must be fully contained in between the overlaps. The 100 µM primer stock should be diluted 100x in NEB buffer 2.1 and 1 µL of the resulting primer solution should be used in a standard HiFi reaction with the cut and dephosphorylated vector. Better results are obtained if the primer is annealed to a reverse-complementary primer beforehand so as to obtain a dsDNA (cf. Chapter 0).

#### *II.4.1.5) Site-directed mutagenesis*

- Primers are designed to contain the intended allele in their overhang, and to linearize the whole vector without overlap after amplification with Q5 DNA polymerase.
- 50 ng PCR Product.
- 5 µL 2X NEB KLD (Kinase, Ligase, DpnI) Reaction Buffer.
- 1 µL 10X NEB KLD Enzyme Mix.
- Up to 10 µL autoclaved ddH2O.
- Incubate at room temperature for 1h.
- Transform 5 µL of each KLD reaction mix alongside 25 ng of cut vector as a negative control to estimate the number of false positives on the transformation plates.

#### *II.4.1.6) Transformation in* E. coli

##### II.4.1.6)(1) Preparation of chemically competent E. coli

###### (a) Day 1

- Inoculate 10 µL of previous chemically competent aliquot into 5 mL LB (+ antibotic if strain with vector, *e.g. sExpress*).
- Book the centrifuge for the next day.

###### (b) Day 2

- Inoculate 1.5 ml of overnight culture into 150 ml LB in a 500 ml sterile conical flask (+ antibotic if strain with vector, *e.g. sExpress*)
    - I usually do 2 x 150 ml in two flasks
- Incubate at 37°C 200 RPM shaking to reach $OD_{600}$=0.3-0.5 (~2h)
    - Pre-chill the centrifuge at 4°C. It takes approximately 20min to cool down.
    - Pre-chill a freezer box into at -80°C.
    - Pre-chill on ice three falcon tubes per conical flask.
- Once at appropriate OD, place conical flask on ice.
- Aliquot the culture into 50 ml falcon tubes (still on ice).
- Centrifuge at 2700g / 4000rpm, 10 min, 4°C.
    - Pre-chill on ice 25 sterile microcentrifuge tubes per conical flask.
- Re-suspend in 10 ml of chilled $CaCl_2$, 20% v/v glycerol (pH7, filter-sterilized).
- Centrifuge at 2700g / 4000rpm, 10 min, 4°C.
- Re-suspend in 500 µL of chilled $CaCl_2$, 20% v/v glycerol (pH7, filter-sterilized).
- Pool all falcon tubes together (still on ice).
- Aliquot 100 µL of cells into each pre-chilled sterile microcentrifuge tubes (still on ice).
- Transfer microcentrifuge tubes to pre-chilled freezer box and store at -80°C.

##### II.4.1.6)(2) Transformation

- Turn on the 42°C water bath.
- Thaw on ice one microcentrifuge tube of competent E. coli for each two constructs to transform.
- Still on ice, place one empty sterile microcentrifuge tube.
- Add 5 µL of vector to transform.
- Add 40 µL of competent cells, flick, and incubate on ice for 30 min.
    - 20 µL is enough if transforming a whole vector (e.g. for methylation).
- Heat-shock 42°C for 30 sec in water bath.

- Rest on ice 2 min.
- Add 300 µL of LB medium
    - 1 mL if transforming a whole vector (e.g. for methylation).
- Incubate 45min-1h in shaker at 37°C.
- Spread 100 µL of transformation mix on plates with adequate antibiotic.
    - For HiFi with more than five parts, plate the whole transformation mix.
- Overnight growth at 37°C.

#### *II.4.1.7) DNA purification and extraction*

- PCR purification: Qiagen (Manchester, UK) QIAquick PCR purification kit; elution in 30 µL after incubation at 55°C for 3 min.
- Gel purification: NEB Monarch® DNA Gel Extraction Kit; elution in 10 µL after incubation at 55°C for 3 min.
- Plasmid extraction: NEB Mornach® plasmid miniprep kit; elution in 50 µL after incubation at 55°C for 3 min.
- Genome extraction: GenElute Bacterial Genomic DNA Kit; elute in 200 µL after incubation at 55°C for 5 min.

#### *II.4.1.8) Conjugation in C. autoethanogenum*

##### II.4.1.8)(1) Original conjugation protocol

This initial protocol was used in Chapter III:, Chapter V.3, and Chapter VI.3. It was quickly optimized throughout Chapter IV:, as described in the next section.

###### (a) Transformation into *E. coli* donor cells

In order to conjugate a plasmid into *C. autoethanogenum*, it is first transformed into *E. coli* sExpress to create a donor strain. To conjugate successfully, the plasmid must carry two origins of replication: one to replicate within sExpress, and one to replicate within *C. autoethanogenum*. Similarly, it must carry an antibiotic resistance marker which is functional both in *E. coli* and *C. autoethanogenum*. It must also carry an origin of conjugal transfer *oriT* to start the process of linearization and translocation to the recipient cell. A family of such vectors has been described in section I.4.1.

###### (b) Concomitant growth of *E. coli* donor cells and *C. autoethanogenum* recipient cells

The next step is to grow the recipient and donor cells so that they reach exponential phase at the same time.

Four days before the mating, *C. autoethanogenum* is inoculated into YTF medium in anaerobic conditions straight from the -80°C freezer.

On the day before the mating, the *E. coli* donor strain is inoculated for the first time into LB medium supplemented with Kanamycin and another antibiotic, in order to maintain the R702 and the shuttle vector to conjugate, respectively. It is incubated at 37°C and 250 rpm overnight; meanwhile, the *C. autoethanogenum* inoculum is subcultured into 1:10, 1:50 and 1:100 cultures.

### (c) Mating

On the day of mating, the donor sExpress strain is subcultured to a 1:100 concentration. When it reaches an $OD_{600}$ in between 0.2 and 0.4, 1 mL of culture is centrifuged at 3000*g, gently resuspended in 500 µL PBS by flicking and inversions of the microcentrifuge tube (no pipetting or vortexing), pelleted down again at 3000*g, and the supernatant is discarded.

Inside the anaerobic cabinet, 200 µL of the *C. autoethanogenum* overnight culture which is the closest to $OD_{600}$ 0.1 is used to gently resuspend the sExpress pellet. The resulting mating mix is poured on a pre-reduced YTF plate with no antibiotic. The mating plate is incubated for 16h to 24h.

### (d) Resuspension in mating slurry and selection of transconjugants

16h to 24h after mating, the mating plate is flooded with 500 µL of pre-reduced PBS and the cells are dislodged by vigorously scraping the plate with a wedge shaped spreader. The plate is tilted to collect the resulting mating slurry, and 100 µL are taken out with a pipette and immediately spread on a pre-reduced YTF plate supplemented with D-cycloserine (to select against *E. coli*)and an adequate antibiotic to select for the vector to conjugate. Serial dilution of the mating slurry in pre-reduced PBS can be done at this stage. The selection plate is then incubated for 72h or until transconjugants colonies start to appear.

#### II.4.1.8)(2) Optimized conjugation protocol

After optimisation throughout Chapter IV:, this protocol was used in all remaining conjugations.

##### (a) Inoculation of *C. autoethanogenum* recipient cells

Four days before the mating (Monday), a whole fresh (<1 year old) cryostock of *C. autoethanogenum* is inoculated into pre-reduced YTF medium in anaerobic conditions straight from the -80°C freezer.

##### (b) Transformation into *E. coli* donor cells and subculture of *C. autoethanogenum*.

Early on the day before the mating (Wednesday), the shuttle vector to be conjugated is transformed into the *E. coli* sExpress donor strain. Alongside your samples, transform a control vector (suggested: pMTL83151). Also transform water onto a plate with only Kanamycin in order to isolate a WT sExpress colony to serve as negative conjugation control. Incubate the plates at 37°C overnight.

Meanwhile, the *C. autoethanogenum* inoculum is subcultured to a starting $OD_{600}$ of 0.05 (and 0.1 as a backup) into 4 mL of pre-reduced YTF.

##### (c) Mating

Early on the day of mating (Thursday), two colonies of each donor strain are inoculated into 5 mL of LB with adequate antibiotics. As soon as one of the inoculums of each donor strain reaches an $OD_{600}$ in between 0.2 and 0.4 (~4h-6h), 1 mL of culture is centrifuged at 3000*G, gently resuspended in 500 µL PBS by flicking and inversions of the microcentrifuge tube (no pipetting or vortexing), pelleted down again at 3000*G, and the supernatant is discarded.

Inside the anaerobic cabinet, 200 µL of *C. autoethanogenum* overnight culture is used to gently resuspend the *sExpress* pellet. The resulting mating mix is poured onto a pre-reduced YTF plate with no antibiotic. The mating plate is incubated for 20h.

##### (d) Resuspension in mating slurry and selection of transconjugants

20h after mating (Friday), the mating plate is flooded with 600 µL of pre-reduced PBS and the cells are dislodged by vigorously scraping the plate with a wedge shaped spreader. The plate is tilted to collect the resulting mating slurry, which is transferred into a micro-centrifuge tube before being normalized to 600 µL with pre-reduced PBS. 50 µL of mating

slurry are taken out with a pipette and immediately spread onto a pre-reduced YTF plate supplemented with D-cycloserine (to select against *E. coli*) and an adequate antibiotic to select for the vector to conjugate. Serial dilution of the mating slurry in pre-reduced PBS can be done at this stage. The selection plate is then incubated for 72h or until transconjugants colonies start to appear.

#### II.4.1.8)(3) Patching of transconjugants

If individual colonies are to be isolated for phenotypic or genotypic characterisation, they must first be patched on a separate plate using an inoculation loop. This is to ensure that (1) the colonies are true transconjugants and are growing adequately (2) there is enough biomass to proceed to colony PCR and (3) the PCR reaction does not get contaminated by a background of WT *C. autoethanogenum* or *E. coli*.

#### II.4.1.8)(4) Isolation of pure colonies

In order to lose the vector used to generate a confirmed mutant, a patch of cells is streaked onto a plate without antibiotics so as to isolate single colonies. The plate is then replicated or individual colonies are patched onto both another plate without antibiotic and a plate with antibiotic selecting for the vector to lose. Patches which grow without antibiotic but cannot grow on antibiotics have lost the vectors. To confirm this, the patch is amplified with primers targeting the vector to be lost, and it is inoculated in liquid culture with the antibiotic. If both the growth in liquid culture and the colony PCR fail to produce any signal (increase in $OD_{600}$ or amplicon, respectively), the patch can be inoculated in liquid culture without antibiotic to prepare cryostocks.

### *II.4.1.9) Cryopreservation*

200 µL of E. coli overnight culture are used to inoculate a Microbank®. After a few minutes, the liquid is removed from the Microbank® and the remaining beads are stored at -80°C. For each subsequent inoculations, a single bead can be used. As demonstrated in Chapter IV.3.2, cryopreservation of sExpress strains already carrying a shuttle vector seems to harm their conjugation potential over time.

180 µL of *C. autoethanogenum* culture at $OD_{600}$ 0.4-1 are complemented with 20 µL of DMSO in a screw-cap tube inside the anaerobic cabinet, then stored at -80°C. The whole aliquot should be used for subsequent inoculation (do not freeze-thaw cryostocks).

### II.4.2. Codon optimization

All the CDS of *C. autoethanogenum* were extracted from its published genome on the NCBI website (CP012395.1) [240,241]. The codon usage table was generated with the CUSP algorithm from the European Molecular Biology Open Software Suite (EMBOSS) (http://www.hpa-bioinfotools.org.uk/, 06/09/2017) [242]. It was then used to train the COOL algorithm (http://cool.syncti.org/, 06/09/2017) [243] and to codon-optimize pmCDA1 with the default parameters. This optimised pmCDA1 sequence was once more optimized by Genescript (NJ, USA) before they synthesized it and inserted in the backbone pUC57 to create the vector vFS07_pmCDA1_COOL_Genescript. The resulting sequence is available in Chapter VIII.2.1.2). All subsequent codon optimizations (gFS05_UGI-LVA and gFS06_Spymac-PAM) were directly done by Genescript.

### II.4.3. Bookmark constructs (Chapter III)

#### *II.4.3.1) Assembly of pMTL431511_BMa*

The KO vector **pMTL431511_BMa (vFS42)** is based on the Cas9 pMTL vector series [169] which, in addition to an antibiotic cassette, a Gram-positive and a Gram-negative replicons and a conjugative transfer function, also carries a functional truncated Cas9 (trCas9), a sgRNA cassette and an editing cassette composed of two 1kb homology arms flanking the mutagenic allele. pMTL431511_BMa was designed to replace the WT *pyrE* gene with a mutant allele (*BMa*) consisting of a 207 nt array of all the bookmark protospacers and their associated PAMs. The vector vFS38, which is a standard Cas9 pMTL vector with a sgRNA targeting the *pyrE* gene of *C. autoethanogenum,* was digested with AsiSI & AscI to remove its gene editing cassette and treated with Antarctic phosphatase to prevent backbone self-ligation. In parallel, a new gene editing cassette was being prepared from *BMa* (gFS04) (synthesised by IDT) and two 1kb homology arms flanking the *pyrE* locus which respectively include the first two and last two codons of *pyrE*. *BMa* was resuspended in 50 µL of ddH2O. Using *C. autoethanogenum* as DNA template, the LHA was amplified using primers oFS109 & oFS119 and the RHA using oFS112 & oFS120. The resulting amplicons were run on a 1% (w/v) electrophoresis gel, extracted and purified using NEB gel extraction kit with 6 µL elution volume. Finally, digested vFS38, left homology arm, right homology arm, and BMa were assembled together in one single step using a NEBuilder® HiFi DNA Assembly Cloning Kit to and cloned into *E. coli* DH5α to form the vector pMTL431511_BMa.

#### *II.4.3.2) Assembly of pMTL431511_BM4 to pMTL431511_BM12*

The nine bookmark complementation vectors, aimed at restoring the WT *pyrE* allele in the *ΔpyrE* strain, were assembled in two steps. First, a new editing template was amplified from *C. autoethanogenum* genome with primers oFS66 & oFS67 in order to restore the WT *pyrE* allele. The resulting amplicon was purified using Qiagen PCR clean-up kit. Finally, the pMTL CRISPR vector vFS11 and the editing template were digested by AsiSI & AscI (vFS11 was dephosphorylated using Antarctic phosphatase), gel purified, eluted in 6 µL of elution buffer and ligated overnight at 16°C with T4 DNA ligase to produce the intermediary vFS17 vector.

In the second assembly step, the sgRNA cassette of vFS17 was replaced by one of nine sgRNA cassettes targeting a different bookmark. vFS17 was digested with SalI, dephosphorylated with Antarctic phosphatase, and gel purified in 6 µL elution buffer. In parallel, 100 µM solutions of ssDNA oligomers (oFS39 to oFS47) coding for the sgRNA targeting each bookmark were diluted 100-fold in NEB buffer 2.1 to form single-stranded sgDNA stocks. Finally, the nine bookmark complementation vectors were assembled in parallel through HiFi assembly by combining 1 µL of their respective single-stranded sgDNA stock with 50 ng of the vFS17 vector previously digested by SalI and 5 µL of HiFI master mix in a total reaction of 10 µL. These vectors were labelled **pMTL431511_BM4** to **pMTL431511_BM12** (**vFS21-29**) for their associated target bookmark protospacer.

### II.4.4. Target-AID constructs (Chapter V)

#### *II.4.4.1) Removal of undesirable restriction sites*

The restriction sites of nCas9 were removed from the vector vFS04_pMTL83151_nCas9_trthl. This plasmid was kindly provided to us by Dr. Daphne Groothuis. This vector places nCas9(D10A) under the control of a truncated, non-functional, $P_{thl}$ promoter from *Clostridium acetobutylicum*. Disrupting five restriction sites required amplifying the whole vector in five parts with Q5® DNA polymerase, each part flanked by ~30 bp overlapping regions which each contain one disrupted restriction sites. This was done with the primers oFS17 & oFS18, oFS19 & oFS20, oFS21 & oFS22, oFS23 & oFS24, and oFS26 & oFS27. The five resulting amplicons were assembled using a NEBuilder® HiFi DNA Assembly Kit and cloned into *E. coli* DH5α to form the vector **vFS08_pMTL83151_nCas9_trthl_RSKO** (for "Restriction Site Knock-Out").

Cau10061II was removed from the pMTL backbone using the Q5® Site-Directed Mutagenesis (SDM) kit (NEB, Hitchin, UK) and the primers oFS03 & oFS04 to amplify the vector pMTL83151. After cloning and plasmid extraction, this resulted in the vector **pMTL83151_RSKO**.

At last, Cau10061II was removed from *lacI* with the help of NEBuilder® HiFi DNA Assembly Cloning Kit. Dr Ryan Hope kindly provided us with the vector vFS06, which encodes Cas9 under the control of Pfacoid [244] and the LacI repressor. vFS06 was digested with MluI, gel extracted, and mixed with oFS64 annealed with oFS76 in a HiFi reaction. This resulted in the vector **vFS20_LacI_RSKO**.

**vFS70_pMTL83151_nCas9_RSKO** and **vFS71_pMTL83151_nCas9** were assembled by digesting pMTL83151, pMTL83151_RSKO, vFS04_pMTL83151_nCas9_trthl, and vFS08_pMTL83151_nCas9_trthl_RSKO with NotI and XbaI, gel extracting the bands with the appropriate length, ligating the RSKO nCas9 with the RSKO backbone to create the vector vFS70_pMTL83151_nCas9_RSKO, and ligating the WT nCas9(D10A) with the WT pMTL83151 to create the vector vFS71_pMTL83151_nCas9.

### *II.4.4.2) Test of different inducible promoters*

The assembly of **vFS35_TA_PlacIq_LacI_*pyrE*** is long and convoluted. It started with the amplification of vFS05_Pfacoid_CatP with the primers oFS53 & oFS54 and vFS20_LacI_RSKO with oFS55 & oFS56. These amplicons were combined into vFS30_Pfacoid_LacI with a NEBuilder® HiFi DNA Assembly Kit. In parallel, the gblock gFS02 – encoding the sgRNA cassette – was synthesised by Genescript (NJ, USA) and amplified with oFS34 & oFS35 to create the amplicon oFS34-35. Afterwards, vFS30_Pfacoid_LacI, oFS34-35 and pMTL83151_RSKO were all digested (respectively with XbaI & NdeI, XbaI & AscI, and NdeI & AscI) and ligated into a single vector, vFS13. Meanwhile, vFS08_pMTL83151_nCas9_trthl_RSKO was amplified with oFS30 & oFS31 to produce the amplicon oFS30-31, and the vector vFS07_pmCDA1_COOL_Genescript was amplified with primers oFS28 & oFS29. Finally, vFS13 was digested by NdeI & NotI and used as a backbone during the HiFi assembly of oFS28-29 and oFS30-31 to produce vFS31_TA_PlacIq_LacI. Later, an error in the design of gFS02, was corrected by SalI digestion of vFS31_TA_PlacIq_LacI and subsequent HiFi assembly with the annealing product of primers oFS115 & oFS116 to produce vFS35_TA_PlacIq_LacI_*pyrE*.

Fortunately, the assembly of **vFS62_TA_Pthl_LacI_pyrE** was more straightforward. The vector pMTL-IC101 [245] and vFS35_TA_PlacIq_LacI_*pyrE* were amplified with the primers oFS245 & oFS246 and oFS247 & oFS248, respectively. In parallel, vFS35_TA_PlacIq_LacI_*pyrE* was digested with SpeI & XbaI. oFS245-246 and oFS247-248 were inserted in the resulting linearized backbone using NEBuilder® HiFi DNA Assembly Kit.

In order to assemble **vFS36_TA_PfdxE_pyrE**, the $P_{fdxE}$ promoter and its associated riboswitch were amplified out of pMTL-IC111-E [245] with oFS93 & oFS94, while the Target-AID backbone of vFS35_TA_PlacIq_LacI_*pyrE* was amplified with oFS95 & oFS96. Both parts were combined using a NEBuilder® HiFi DNA Assembly Kit in order to constitute vFS36_TA_PfdxE_pyrE, the $P_{fdxE}$.

**vFS96_trTA_pyrE** was derived from vFS36_TA_PfdxE_pyrE. The backbone of vFS36_TA_PfdxE_pyrE was isolated after digestion of the vector with NheI and XbaI to remove 5'-end of nCas9 and the $P_{fdxE}$ promoter. In parallel, the trCas9 frameshift mutation was recreated by amplification of the Cas9 vFS36_TA_PfdxE_pyrE with oFS287 & oFS288, while the $P_{thl}$ promoter and the WT ruvC domain were amplified from pMTL431511_BMa (cf. Chapter II.4.3.1)) using oFS285 & oFS286. At last, all three parts were combined in a HiFi assembly to form vFS96_trTA_pyrE.

### *II.4.4.3) Target-AID characterisation controls*

**vFS37_NosgRNA** was assembled by annealing oFS117 with oFS118 and cutting out the sgRNA cassette from vFS36 by digesting it with XbaI & AsiSI. The cut vFS36_TA_PfdxE_pyrE backbone and the oFS117-118 dsDNA oligomer were then combined in a HiFi assembly to assemble the vector vFS37_NosgRNA.

The Cas9 H840A mutation was carried out by Q5 SDM of vFS36_TA_PfdxE_pyrE with oFS91 & oFS92. This resulted in the **vector vFS39_dTA**.

Lastly, the HDR control vector was assembled by digesting vFS36_TA_PfdxE_pyrE with NotI & AscI to remove the whole Target-AID and sgRNA cassettes, and only keep the pMTL83151 backbone. Then, new nCas9 and sgRNA expression cassettes were rebuilt in one block by amplifying vFS36_TA_PfdxE_pyrE with oFS113 & oFS114. In parallel, the *ΔpyrE* gene editing template was amplified from *C. autoethanogenum* genomic DNA using oFS109 & oFS110 for the LHA, and oFS111 & oFS112 for the RHA. All parts were finally assembled in one HiFi assembly step to produce the vector **vFS40_nCas9_PfdxE_HDR**.

### II.4.5. Multiplexing constructs (Chapter VI)

#### *II.4.5.1) First round of mutagenesis*

**vFS50_TA_msgRNA** was assembled in two steps. The first step resulted from a 6-parts Hifi assembly. The CLAU532A, CLAU534A, and CLAU1794A spacers and the tfdx, tecT1 and tectyrS terminators were added to the sgRNA scaffold by amplifying vFS36_TA_PfdxE_pyrE with oFS121 & oFS137, oFS140 & oFS141, and oFS144 & oFS145 respectively. Then, the Pcpth and Pcpfdx promoters were amplified from the vectors vFS46 and vFS47, kindly provided by Dr. Katalin Kovacs [246], using primers oFS142 & oFS143 and oFS138 & oFS139, respectively. These five parts were assembled with vFS36 previously digested with SalI & AsiSI. Unfortunately, this led to an incomplete, truncated construct (vFS49) with the region in between oFS121 & oFS139 missing. In the second round, oFS121-137 and oFS138-139 were re-amplified from vFS36_TA_PfdxE_pyrE and vFS47. oFS121-137 and oFS138-139 where then digested with SpeI and ligated overnight. The resulting ligation product was amplified with oFS154 & oFS155. After purification, oFS154-155 was digested alongside vFS49 with SalI & AatII, and both parts were finally ligated together to constitute vFS50_TA_msgRNA.

In order to assemble the CRISPR array of **vFS51_TA_mCRISPR**, two rounds of amplifications were needed. First, the primers oFS129 & oFS130 were annealed with each other and amplified with oFS131 & oFS158. Then, the resulting amplicon was itself amplified with oFS159 & oFS160. In parallel, the tracrRNA was also reconstituted in two consecutive rounds of amplifications: first, vFS36_TA_PfdxE_pyrE was amplified with oFS163 & oFS164; then, the resulting amplicon was amplified with oFS165 & oFS166. At last, the Pcpth promoter was extracted from vFS47 with oFS161 & oFS162. vFS36_TA_PfdxE_pyrE was digested with SalI and AsisI, then was combined with oFS159-160, oFS161-162 and oFS165-166 into a HiFi assembly to produce vFS51_TA_mCRISPR.

The assembly of **vFS48_TA_mtRNA** was much simpler. The protospacer and tRNA sequences of each sgRNA were simply added by amplification of vFS36_TA_PfdxE_pyrE with oFS121 & oFS122, oFS123 & oFS124, and oFS125 & oFS126. Meanwhile, vFS36_TA_PfdxE_pyrE was digested with SalI & AsiSI. Afterwards, all parts were used in a HiFi assembly to create vFS48_TA_mtRNA.

**vFS52_TA_CA532A**, **vFS53_TA_CA534A** and **vFS54_TA_CA1794A** where assembled by standard sgRNA exchange HiFi assembly. vFS36_TA_PfdxE_pyrE was digested with SalI, then

used in a HiFi assembly with the annealing products of oFS167 & oFS168, oFS169 & oFS170, or oFS171 & oFS172, respectively.

#### *II.4.5.2) Second round of mutagenesis*

**vFS57_TA_msgRNA_CA1794B** was assembled through simple sgRNA cassette exchange. The primers oFS195 & oFS196 were annealed together, then used in a HiFi assembly with a SalI-digested vFS50_TA_msgRNA vector. The same method was used to change the sgRNA cassette of vFS54_TA_CA1794A using the primers oFS193 & oFS194 to create the vector **vFS58_TA_CA1794B**.

In order to replace one of the two tRNAs of vFS48_TA_mtRNA and replace the CLAU1794A protospacer, the whole array had to be replaced. Consequently, vFS48_TA_mtRNA was amplified with oFS121 & oFS197 and oFS198 & oFS199, while vFS36_TA_PfdxE_pyrE was being amplified with oFS126 & oFS200. vFS48_TA_mtRNA was digested with SalI & AsiSI, and all four parts were fused in a HiFi assembly, resulting in the vector **vFS56_TA_mTHRtRNA**.

Ribozymes were added to the tracrRNA of vFS51_TA_mCRISPR by amplifying the vector with oFS201 & oFS202 and oFS203 & oFS204. vFS51_TA_mCRISPR was digested with NheI & AsiSI. In parallel, vFS36_TA_PfdxE_pyrE was amplified with oFS205 & oFS206 and vFS51 was amplified with oFS166 & oFS207. After Hifi assembly of all five parts, this resulted in the vector **vFS60_TA_mCRISPR_HHDV**.

#### *II.4.5.3) Characterisation of TA-iSpymac*

For the assembly of **vFS83_TA-spymac_pyrE,** vFS36_TA_PfdxE_pyrE was first linearized by amplification with oFS253 & oFS254, which also changed the sgRNA cassette to target a NAA PAM in *pyrE*. In parallel, the R221K and N394K mutations were carried out by amplifying vFS36_TA_PfdxE_pyrE with oFS257 & oFS258, oFS259 & oFS260, and oFS261 & oFS262. All the resulting amplicons were finally combined with the gblock gFS06_Spymac-Pam_genescript previously synthesized by Genescript to provide the PAM-interacting domain of *S. macacae*, and the circular vector vFS83_TA-spymac_pyrE resulted from the ensuing HiFi assembly.

#### *II.4.5.4) Third round of mutagenesis*

In order to assemble **vFS72_mTA-NG**, vFS57_TA_msgRNA_CA1794B was used as template for five PCR amplifications – respectively with primers oFS28 & oFS234, oFS235 & oFS236, oFS237 & oFS238, oFS239 & oFS240, and oFS241 & oFS242. VFS57_TA_msgRNA_CA1794B was then digested with XhoI & NotI, and all six parts were assembled by HiFi assembly.

UGI and its LVA-tag was codon-optimized to by Genscript and synthesized by IDT with 35 bp overhangs (gFS05). vFS57_TA_msgRNA_CA1794B was digested with NotI then re-circularized with gFS05 in a HiFi assembly to constitute the vector **vFS75_mTA-UGILVA**.

Shorter sgRNAs were inserted into vFS57_TA_msgRNA_CA1794B by amplifying it with oFS232 & oFS233, digesting it with AatII & XhoI, and combining both parts in a HiFi assembly. Unfortunately, this resulted in the vector vFS66, which had point mutations in its sgRNA scaffold. In order to fix this, vFS36 was amplified with oFS145 & oFS284, and vFS66 was digested with XhoI. Both parts were combined in a HiFi assembly, which resulted in the vector **vFS94_mTA_trsgRNA**.

Finally, three different protospacers targeting CLAU_1794 were inserted into vFS57_TA_msgRNA_CA1794B by digesting in with SalI & XhoI, and by amplifying this vector with oFS249 & oFS250 and oFS251 & oFS252. All three parts were combined into a new vector (**vFS74_mTA-NG_CA1794**) during a HiFi assembly.

## II.5. Whole-genome sequencing

Genome sequencing was achieved at the DeepSeq next generation sequencing facility at the University of Nottingham with an Illumina MiSeq desktop sequencer. Paired reads of 300 bp were generated. Trimming, filtering and mapping of the reads were achieved using CLC Genomics Workbench (v.20.0.4, CLC bio, Denmark). The reads were mapped to the published *C. autoethanogenum* genome (NCBI CP012395.1.) and compared to the original reads used to assemble this genome (SRA SRR2969415). Default parameters were used when not specified otherwise.

### II.5.1. Trimming

Because the read QC report showed important irregularities in the GC-content and ambiguity levels of the first 20 bp of the read populations (likely residual sequences of the adapters used to sequence genomic fragments during Illumina sequencing), the first 20bp of all paired reads were systematically trimmed away.

Quality score = 0.05 (the lower, the more stringent), Maximum number of ambiguities=0, Discard all reads shorter than 270 bp. Depending on the genome, this resulted in 35% to 27% of reads being removed from the dataset, with the average length of reads becoming 279.4 bp.

### II.5.2. Mapping

For the alignment of the reads with the reference sequence, an affine gap cost of 3 to start a deletion or an insertion was used; followed by an extension cost of 1 for both deletions and insertions.

The reads of the original genome assembly (SRR2969415) were used as a control to identify sequencing and mapping artefacts.

### II.5.3. Local realignment

The default parameters were used (Multipass realignment=2; Maximum guidance-variant length=200).

### II.5.4. Detection of regions with no coverage

These regions were manually counted using the user interface connected to the reads map to find regions with coverage =0.

### II.5.5. Fixed Ploidy Variants detection

A fixed ploidy of 1 was used, with the default parameters of 90% variant probability required before calling, ignoring broken pairs, and using a base quality filter with a neighbourhood radius of 5 bp, a minimum central quality of 20 bp and a minimum neighbourhood quality of 15. Variants with a coverage of less than 10 reads and with a frequency of less than 20% were ignored.

# Chapter III: Bookmarks, a "gold standard" *in situ* complementation strategy

## III.1. Introduction

The first tool developed over the course of this thesis was a new methodology to improve and standardize complementation studies in any organism. This introduction first explains the concepts behind a complementation study, then reviews the current best practice; at last, the basis for a new, improved complementation strategy will be discussed, followed by an outline of the experimental design of this chapter.

### III.1.1. Purpose of a complementation study

The standard procedure when studying a gene's function is to knock out the gene, then to compare the phenotype of the KO strain with the phenotype of the WT strain. However, a complemented strain must also be used to verify that the knock-out of the gene of interest is the only factor which can possibly explain the change in phenotype [247–251]. The complemented strain is built by returning the gene of interest to the KO strain with the intention of restoring gene function (Figure 27). If the WT phenotype is restored by complementing the gene of interest, then it can safely be assumed that any difference in phenotype observed between the KO and the WT strains is indeed caused by the knock-out of the gene of interest. On the contrary, if the complemented strain expresses a different phenotype than the WT, then the phenotypical differences between the WT and the KO strains are at least somewhat impacted by other factors, such as unwanted mutations somewhere else in the genome. In the most extreme case, the complemented strain would behave exactly like the KO strain, which would indicate that the knock-out of the gene of interest does not actually contribute to the observed differences between the WT and the KO strains.

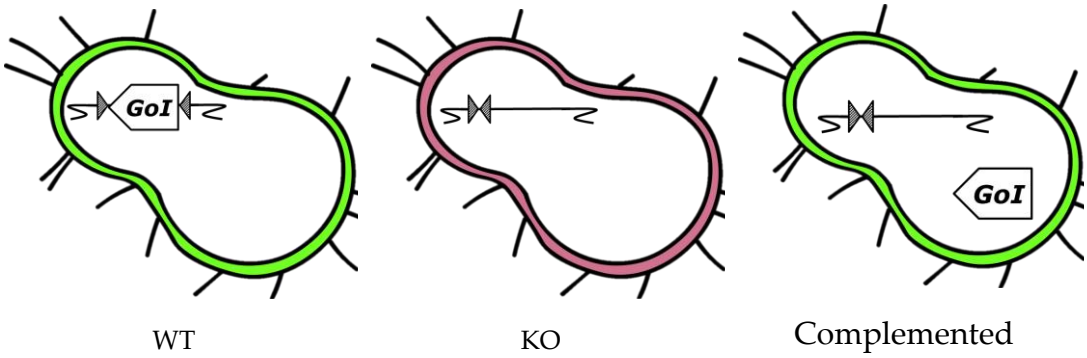
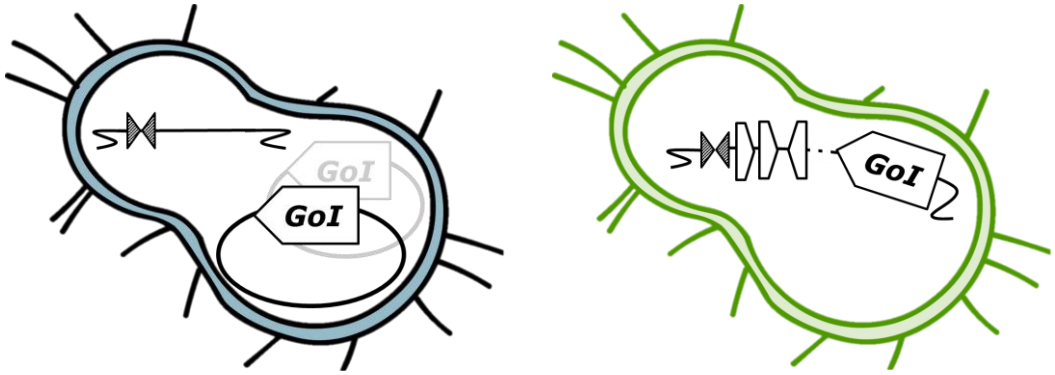



**Figure 27. Illustration of the three strains necessary to study the function of a gene of interest.** The **WT strain** is the native organism used as a control; in the **KO strain**, the gene of interest has been inactivated; in the **complemented strain**, the gene of interest is returned to the KO strain to confirm that the presence or absence of the gene of interest is the sole explanations for the differences in phenotype observed between the WT and the KO strains (here symbolized by a difference of cell wall pigmentation). GoI = Gene of Interest; Green = WT phenotype; Pink = KO phenotype.

It is thus critical to ensure that the complementation method used does not itself produce a change in phenotype in the complemented strain. Current complementation strategies try to minimise that risk, but they all have drawbacks which must be evaluated before choosing the most appropriate complementation strategy.

### III.1.2. Common complementation methods

The two main methods of complementing a gene are to express the gene on a vector [Figure 28(a)] , or to insert it somewhere else in the genome [Figure 28(b)] [250–252].




(**a**) (**b**)

**Figure 28. Schematic of the two established complementation methods. (a)** In vector complementation, a vector expressing the gene of interest is transformed into the KO strain and maintained in several copies through selective pressure (e.g., antibiotics); **(b)** In genomic complementation the gene of interest is inserted back into the genome of the KO strain in a locus thought to not interfere with the expected WT phenotype.

#### *III.1.2.1) Vector complementation*

Complementing a gene on a vector is the most simple but also the most problematic approach. Indeed, vectors are maintained only at great metabolic costs for the cells: they are several kb long and force the cell to retain several copies of them at all time. They also require the presence of an antibiotic (or equivalent selective pressure) to avoid disappearing from the population. Finally, their copy number also impacts the expression level of the complemented gene, potentially raising it to toxic levels or loosening up its regulation [251]. Most of these discrepancies can be alleviated by transforming the WT and the KO strain with an empty complementation vector (that is, a complementation vector without the gene of interest) before comparison with the complemented strain, but it does not fix the issue of the overexpression of the complemented gene.

#### III.1.2.2) *Genomic complementation in different locus*

The second established complementation method is to integrate the gene of interest back into the genome of the KO strain. This method has the advantage of creating a complemented strain with a stable phenotype even without selective pressure, which does not have to replicate several thousands of extra DNA bp, and which conserves the original copy number of the gene of interest. However, it relies on the availability of a hospitable genomic locus. Indeed inserting the gene of interest in an arbitrary locus might have unintended consequences, such as disrupting the expression of neighbouring genes. For this

reason, our group previously identified the *pyrE* locus as a convenient locus to insert complemented genes, with no obvious unintended consequences having been observed so far. The *pyrE* gene codes for orotate phosphoribosyltransferase, an enzyme involved in pyrimidine metabolism. Consequently, *ΔpyrE* mutants are uracil auxotrophes but resistant to FOA, an analogue of orotic acid which becomes toxic when processed into pyrimidine. This dual sensitivity has enabled the ACE genome editing strategy discussed in page 31, which can also be used to carry out complementation studies [174–176].

ACE enables genomic complementation in the *pyrE* locus without scars or heterologous selection markers which might interfere with the phenotype of the complemented strain. However, complementing a strain using ACE requires considerable preparation: it needs to be done in a specifically designed *ΔpyrE* strain and necessitates a specific minimal medium. Another shortcoming of complementing in the *pyrE* locus using ACE is highlighted when functions which are related to genomic structure – such as sporulation – are being studied. In these cases, changing the locus of a gene might impact its function, and thus jeopardize its complementation. In the example of sporulation, the order in which the genome migrates to the forespore affects the order in which genes can be expressed in the forespore [253]. As such, complementing a gene necessary to spore formation in a different locus might not restore WT sporulation behaviour.

### III.1.3. Bookmark, a new *in situ* genomic complementation strategy

Clearly, an optimal complementation strategy would complement the gene of interest back in its original locus, while leaving no ancillary mutations such as a genomic scar or a selection marker which might impact the phenotype. The complemented strain should be indistinguishable from the WT strain, except maybe for a watermark which would allow unequivocal identification of each strain [175].

Peter Rowe hypothesised that the unprecedented precision and reliability enjoyed by Cas9-mediated mutagenesis could be leveraged to design such a gold-standard complementation strategy. This strategy makes use of a standardised, rationally designed short stretch of DNA, named bookmark, to replace a knocked-out gene during a Cas9-mediated genome editing experiment. Once inserted in the genome, this stretch of DNA which includes a PAM can be easily targeted during a subsequent genome editing step – notably with the intention of restoring the deleted gene back into its original locus in order to perform a complementation study. This strategy is only limited by the capabilities of Cas9, and as such it should be applicable in a wide range of organisms. The bookmark complementation strategy is further illustrated in Figure 29.

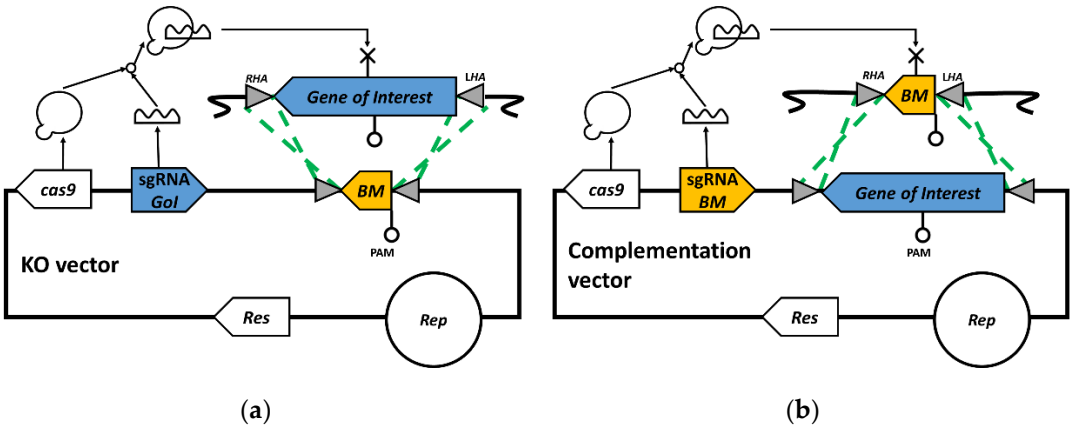



**Figure 29. Overview of the bookmark complementation strategy. (a)** First step — knock-out of the gene of interest and insertion of the bookmark in its genomic locus using a knock-out vector (KO vector) consisting of a Cas9 nuclease and single guide RNA (sgRNA) expression cassettes, as well as an editing template composed of one 24 nt bookmark flanked by homology arms. **(b)** Second step — after isolation of the KO mutant and plasmid loss, another round of Cas9-mediated homology-directed mutagenesis is carried out with the help of a complementation vector, to restore the gene of interest in its original locus. The complementation vector is identical to the KO vector, except for its sgRNA cassette, which targets the genomic bookmark that was previously inserted, and editing template, which consists of the gene of interest flanked by the same homology arms. The gene of interest can be watermarked with a silent mutation for higher reliability of the complementation step. GoI: Gene of Interest, Res: Antibiotic resistance marker, Rep: Replicon, BM: Bookmark, LHA: Left homology arm, RHA: Right homology arm.

### III.1.4. Experimental design

Once individual bookmarks were designed, a proof of principle of their use in a complementation study was executed in *C. autoethanogenum.* First, the sequence of each bookmark was compared to the genome of *C. autoethanogenum* using BLAST and a protospacer design software (Benchling) to identify potential off-target loci. Second, all of the vectors necessary to carry out a knock-out and a subsequent complementation of the *pyrE* gene in *C. autoethanogenum* were assembled. During the assembly process, a side-experiment was performed to optimise the sgRNA cassette exchange protocol. In a third step, the *pyrE* gene was knocked-out and replaced by an array of all bookmark protospacers. In the last step of the bookmark characterisation, each bookmark protospacer was targeted separately by an individual bookmark complementation vector and the complementation efficiency of each bookmark protospacer was calculated by measuring the proportion of transconjugants which reverted to WT genotype after conjugation. The core of the experimental setup was thought out by Peter Rowe, Christopher Humphreys and Nigel Minton, but I refined it and executed all of it.

## III.2. Design consideration

The rational design of the genomic bookmarks can be divided in two main steps: searching for suitable sgRNA targets and reducing the risk of polar effects. The next two subsections detail these two steps. The resulting bookmarks sequences are summarized in Table 5.

**Table 5. Summary of the bookmark sequences.** Nine protospacers were picked from the literature as successful examples of SpCas9 targets in bacterial species [182,183,254]. To constitute a bookmark, each protospacer should be immediately followed by a PAM (suggested: AGG) and complemented with one single nucleotide at either extremity. For each bookmark, the position of the extra nucleotide as well as the orientation of the bookmark relative to the coding sequence (CDS) it replaces are given to avoid internal STOP codons. "+": same direction as the target CDS, " – ": reverse-complementary direction relative to the target CDS, "+/-": either same or reverse-complementary direction relative to the target CDS, *S. pneumonia: Streptococcus pneumonia, B. subtilis: Bacillus subtilis, L. reuteri = Lactobacillus reuteri.*

| Bookmark | Bookmark sequence (24 nt) | | | | Orientation | Origin |
|---|---|---|---|---|---|---|
| | Extra nt | Protospacer (20 nt) | PAM | Extra nt | | |
| **BM4** | G | AGGGTTGTGGGTTGTACGGA | AGG | / | +/- | *S. pneumonia* |
| **BM5** | / | ATTTCTGATATTACTGTCAC | AGG | A | +/- | *S. pneumoniae* |
| **BM6** | / | ACCGATACCGTTTACGAAAT | AGG | A | +/- | *S. pneumoniae* |
| **BM7** | G | TGAAGATCAGGCTATCACTG | AGG | / | + | *B. subtilis* |
| **BM8** | G | TCCGGAGCTCCGATAAAAAA | TGG | / | +/- | *B. subtilis* |
| **BM9** | G | TATTGATTCTCTTCAAGTAG | AGG | / | - | *B. subtilis* |
| **BM10** | / | CCATTGTACTATCATGCTAG | AGG | A | +/- | *L. reuteri* |
| **BM11** | G | ATGCAGTCGGCTGTAGAAAG | AGG | / | +/- | *L. reuteri* |
| **BM12** | G | CGACTGCATTTTATTATGTA | AGG | / | +/- | *L. reuteri* |

Incidentally, the protospacer of BM8 has a restriction site (BspEI = T*CCGG*A, where * represents a cut site) which might be exploited for *in vitro* analyses in some experimental setups [175].

### III.2.1. Protospacer design

For a bookmark to be targetable by Cas9, its most basic requirement is to have a 20bp protospacer directly upstream of a 3bp PAM. Although a wide array of protospacer prediction software are available, the safest way to design protospacer sequences is to look in the literature for protospacers which have already been found to be targetable by Cas9. Another important criterion for a functional bookmark is that its sequence must be absent from the original genome of the organism in which it is being used, for fear of causing off-target mutations during Cas9-mediated mutagenesis. As such, functional protospacers must be sought in different organisms, so that an alternative can be picked if a bookmark happens to be identical to a sequence already found in the target genome. To that end, Peter Rowe identified in the literature nine protospacers which had been successfully targeted by Cas9 in three different organisms to serve as foundation for our Bookmark design [182,183,254].

### III.2.2. Prevention of polar mutations

In the next design step, it was sought to minimise the possibility of inducing polar effects upon insertion of a bookmark in a genome. As discussed in Chapter I, this requirement can be summarised by the imperative of avoiding nonsense mutations inside an operon upon insertions of a bookmark [165]. Removing the coding sequence altogether could also produce polar mutations in some contexts [164]. The final bookmark was thus designed to be inserted in the same reading frame as the coding sequence that it would replace, flanked by the first two and the last two nucleotides of the original gene so as to keep a token coding sequence. To prevent disrupting the original reading frame, the length of the bookmarks was increased to 24bp (equivalent to 8 codons) by the addition of one extra nucleotide to either end of the sequence composed by a protospacer and its associated PAM. The extremity at which this nucleotide was added, as well as the recommended direction of the insertion of a given bookmark inside the mutant reading frame, were picked in order to avoid nonsense codons within the sequence of the bookmark.

## III.3. Off-target analysis

To confirm that the protospacers detailed in Table 5 could be used in *C. autoethanogenum* with a minimal chance of off-target mutagenesis, each bookmark protospacer and their associated PAM was compared to *C. autoethanogenum* using BLASTn (https://blast.ncbi.nlm.nih.gov) [255]. The search parameters summarised in Table 6 were optimised to find short sequences with little similarity to the query but with no gaps so as to find potential off-target sites.

**Table 6. Parameters of the basic local alignment search tool (BLAST) of all the bookmark protospacers and their associated protospacer adjacent motif (PAM) against *C. autoethanogenum* genomes.**

| Search parameter | Value |
|---|---|
| Program | BLASTn |
| Word size | 7 |
| Expect value | 1000 |
| Hitlist size | 100 |
| Match/Mismatch scores | 1, –1 |
| Gapcosts | 5, 2 |
| Filter string | F |
| Genetic Code | 1 |
| **Database** | |
| Posted date | Mar 24, 2020 9:05 AM |
| Number of letters | 8,771,469 |
| Number of sequences | 18 |
| Entrez query | Includes: *C. autoethanogenum* DSM 10061 (taxid:1341692) |

The two hits with the highest level of identity were BM5 and BM9 (Figure 30), respectively with 18 and 19 nt which were found to be identical to a 23 nt sequence in *C. autoethanogenum* genome. Although the PAM of the off-target site was a perfect match in both cases, at least two mismatches were located in the seed region of each protospacer and at least four mismatches were present in the overall protospacer. These two characteristics have been found to lead to low mutagenesis efficiency when using Cas9 in mammalian cells [227], and, as such, all Bookmarks have been preliminarily judged suitable for use in *C. autoethanogenum.*

```
Range 2: 223444 to 223466
Identities:19/23(83%),  Gaps:0/23(0%), Strand: Plus/Minus

BM9    1       TATTGATTCTCTTCAAGTAGAGG  23
               ||| |||| ||||||| || |||
Sbjct  223466  TATGGATTATCTTCAATTACAGG  223444

Range 16: 1235296 to 1235318
Identities:18/23(78%),  Gaps:0/23(0%), Strand: Plus/Plus

BM5    1        ATTTCTGATATTACTGTCACAGG  23
               ||| ||||||| ||||  | |||
Sbjct  1235296  ATTGCTGATATGACTGGAATAGG  1235318
```

**Figure 30. Two top hits of the BLASTn analysis performed on all bookmark sequences against *C. autoethanogenum* genome (accession CP012395.1).** The bit score and Expect value are not shown. "Range" is the position of the hit in the genome; "Identities" shows the number of subject nucleotides which are identical to the query; "Gaps" shows the number of gaps in the subject sequence; "Strand" shows the strand orientation of the query then the strand orientation of the subject separated by a "/". Matches between the query and the subject are represented by a vertical line which connects identical nucleotides ("|").

However, because BLASTn is not well suited to look for such short homology regions [216] (it will miss any potential off-target site which have less than seven contiguous bases identical to the query), a specialized off-target scoring algorithm [222] was used to confirm this hypothesis. The algorithm from Hsu et al. (2013) was applied using Benchling, with the parameters presented in

Table 7. No bookmark received an off-target score lower than 98%, which gave further indication that all nine bookmarks could be used in *C. autoethanogenum* with minimal probability of producing off-target effects.

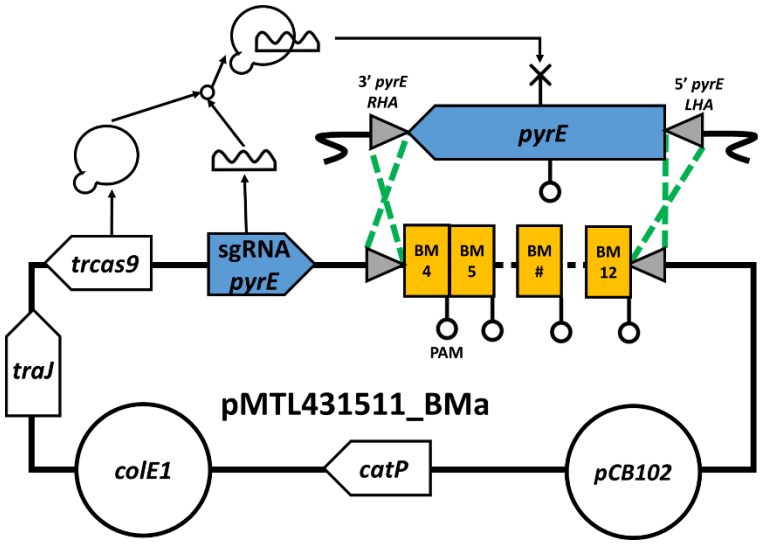
**Table 7. Off-target score (specificity score) of all nine bookmark protospacers calculated with Benchling with the algorithm of Hsu et al, 2013** [222]**.** A score above 50 is generally considered a suitable protospacer.

| Reference genome: | GCA_000484505.1 (*C. autoethanogenum* DSM 10061) | | |
|---|---|---|---|
| Design type | Single guide | | |
| Guide length | 20bp | | |
| PAM | NGG (SpCas9, 3'side) | | |
| **Bookmark protospacer** | **Sequence** | **PAM** | **Specificity Score** |
| **BM4** | AGGGTTGTGGGTTGTACGGA | AGG | 100 |
| **BM5** | ATTTCTGATATTACTGTCAC | AGG | 99.95 |
| **BM6** | ACCGATACCGTTTACGAAAT | AGG | 99.8 |
| **BM7** | TGAAGATCAGGCTATCACTG | AGG | 98.2 |
| **BM8** | TCCGGAGCTCCGATAAAAAA | TGG | 100 |
| **BM9** | TATTGATTCTCTTCAAGTAG | AGG | 99.5 |
| **BM10** | CCATTGTACTATCATGCTAG | AGG | 100 |
| **BM11** | ATGCAGTCGGCTGTAGAAAG | AGG | 100 |
| **BM12** | CGACTGCATTTTATTATGTA | AGG | 100 |

## III.4. Integration of a bookmark protospacer array in the *pyrE* locus

The *BMa* array of all the bookmark protospacers and their associated PAMs was integrated into the *pyrE* locus of *C. autoethanogenum* through conjugation of the plasmid pMTL431511_BMa (vFS42) to produce a *ΔpyrE::BMa* strain [Figure 31(a)]. After conjugation, eight transconjugants were patched on YTF plates complemented with FOA to isolate pure *ΔpyrE* mutants. After four days, six patches grew and were screened by colony PCR using the primers oFS105 and oFS106. The WT (amplicon of 2 kb) and a clean *ΔpyrE* strain (amplicon of 1.5 kb) were used as controls. The expected band of 1.7 kb corresponding to a successful knock-in of the Bookmark array was observed in four of the six strains, although the signal was weak in lanes 3 and 4 [Figure 31(b)]. One of the amplicons (lane 1) was sent to Sanger sequencing to confirm the replacement of *pyrE* with the bookmark array, and the corresponding *ΔpyrE::BMa* colony (cFS03) was re-streaked on an YTF plate without antibiotic to lose the vector before storage in cryostocks.

(a)




(b)

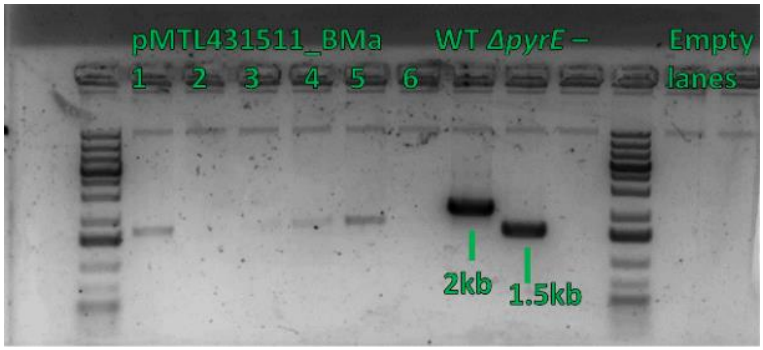



**Figure 31. Knock-out of *pyrE* and knock-in of the bookmark protospacer array. (a)** Integration of an array of bookmark protospacers into the genomic locus of *pyrE* using pMTL431511_BMa. The bookmark protospacer array (BM4 to BM12) is flanked by two homology arms of 1 kb, each homologous to the genomic region directly upstream (LHA) and downstream (RHA) of the *pyrE* locus and which include the first two and last two codons of *pyrE*. Homology-directed mutagenesis of the *pyrE* locus replaces the *pyrE* gene with the protospacer array. trCas9

nuclease and an sgRNA expression cassette counter-select wild type (WT) transconjugants by cutting the genomic DNA of cells which have a WT *pyrE* locus. **(b)** Electrophoresis gel of six *C. autoethanogenum* colonies obtained after conjugation of pMTL431511_BMa. The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% (w/v) agarose gel alongside a Generuler 1 Kb Plus DNA ladder, Thermo Fisher Scientific. The expected size of the amplicon of a successfully knocked-in bookmark array (BMa) is 1.7 kb, versus 2 kb for the WT *pyrE* locus and 1.5 kb for a clean Δ*pyrE* genotype without the bookmark array. No DNA template was added to the PCR mix of the negative control (–). *traJ*: conjugative transfer function, *colE1*: Gram-negative replicon, *catP*: chloramphenicol/thiamphenicol resistance cassette, *pCB102*: Gram-positive replicon, BM#: Protospacer sequence of any bookmark, where # is an integer between 4 and 12.

An unexpected band was detected at 20 kb in all lanes, including lanes with no PCR reactions. To explain this phenomenon, it was hypothesised that a genomic DNA contamination occurred in the TAE buffer used to run the electrophoresis gel. To confirm that this band could safely be ignored, the PCR and the gel electrophoresis were repeated on all the controls and the *ΔpyrE::BMa* strain using a fresh stock of TAE buffer (Figure 32).

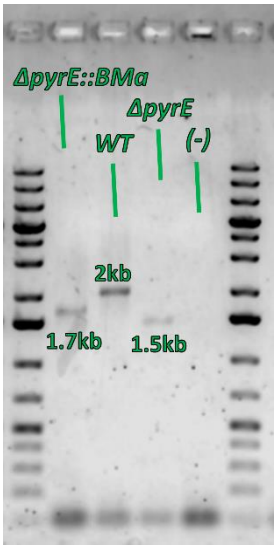



**Figure 32. Subsequent PCRs of the *ΔpyrE::BMa* strain first screened in Figure 31.** In accordance with the hypothesis that the 20 kb band observed in Figure 31 could be safely dismissed as the product of a contamination of the TAE buffer, subsequent PCRs of the strains used as template in Figure 31 did not exhibit any 20 kb band. Generuler 1 Kb Plus DNA ladder, Thermo Fisher Scientific.

As expected, the 20 kb band was no longer visible, and could thus safely be dismissed as a contamination of the TAE buffer.

Optimisation of the sgRNA exchange protocol

The opportunity presented by the assembly of the nine bookark complementation vectors was exploited to improve the method used to exchange the sgRNA cassette of CRISPR vectors. The original protocol relied on ssDNA inserts to exchange the sgRNA cassette of a CRISPR vector; however, this was regularly leading to less than 10 transformants per construct, which is barely above the number of false positives that could be expected from undigested vectors. Occasionally, no transformants were observed at all. Considering the mechanism of HiFi assembly, it was hypothesized that using dsDNA inserts would improve the efficiency of the assembly, and as such, the transformation efficiency of the HiFi reaction mix. Because of the cost of NEBuilder® HiFi DNA Assembly Cloning Kit and because of the necessity to have a high number of samples to gain statistical significance in a process as inconsistent as bacterial transformation, it was decided to combine this protocol optimisation experiment with the necessary assembly of the nine bookmark complementation vectors.

### III.4.1. Experimental setup

Double-stranded DNA (dsDNA) oligomers coding for each bookmark sgRNA were built by annealing together complementary ssDNA oligomers. Annealing was achieved by mixing 5 µL of 100 µM solution of each bookmark sgRNA cassette oligomer (oFS39 to oFS47) with its respective reverse-complementary oligomer (oFS79 to oFS87), then incubating them 30 sec at 98 °C, 5 min at 72 °C, and 5 min at 50 °C before holding the temperature at 15 °C. The resulting double-stranded sgDNA solutions was diluted 100x in NEB buffer 2.1 to form double-stranded sgDNA stocks.

To confirm annealing, 1 µL of each 100µM ssDNA oligomer solution and 1 µL of each double-stranded 100 µM sgDNA solution were diluted 10 times in ddH2O, complemented with 2µL of loading dye, and 5 µL of each resulting mix was then loaded on a 2% (w/v) agarose gel (cf. Figure 33) with 5 µL of ddH2O used as negative control. The electrophoresis was run for 30 min at 120 V and 100 mA.

The band of the double-stranded sgDNA solution was visibly more intense than the band of each separate ssDNA oligomer at the same concentration, confirming that the DNA in the double-stranded sgDNA solution was indeed double-stranded.

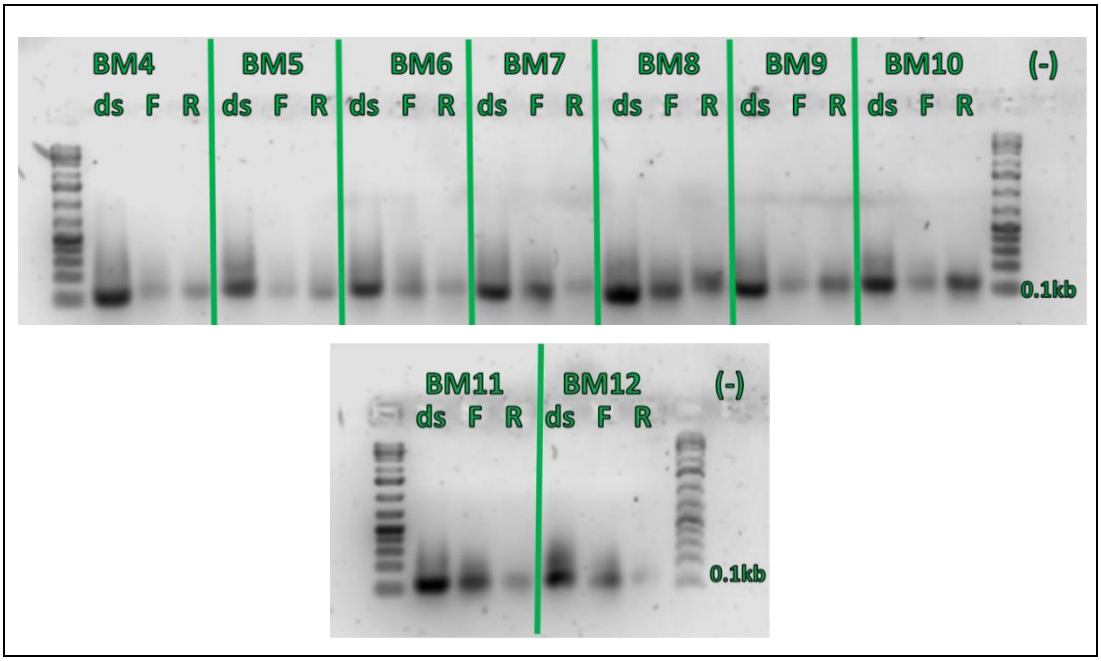



**Figure 33. Electrophoresis gel of nine sgRNA exchange cassettes after annealing (ds) of two complementary ssDNA strands (F and R).** 1 µL of each 100µM ssDNA oligomer solution and 1 µL of each double-stranded 100 µM sgDNA solution were diluted 10 times in ddH2O, complemented with 2µL of loading dye, and 5 µL of each resulting mix was then loaded on a 2% (w/v) agarose gel alongside a Generuler 1 Kb Plus DNA ladder, from ThermoFisher Scientific.

As in the HiFi assembly of the bookmark complementation vectors using ssDNA, 1 µL of double-stranded sgDNA stock was used with 50 ng of digested and dephosphorylated vFS17 in a 10 µL HiFi reaction mix. After 1h of incubation at 50°C, 5 µL of HiFi reaction mix (including either ssDNA or dsDNA sgRNA exchange cassettes) was used to transform 40 µL of chemically competent *E. coli* DH5α. The HiFi mix of pMTL431511_BM11 assembled with dsDNA was only transformed into 20 µL of competent cells, and as such it was removed from subsequent data analysis. 25 ng of cut, dephosphorylated vFS17 was diluted into 5 µL of sterile ddH2O and transformed alongside the HiFi reactions to control for the background colonies resulting from self-ligated or undigested vectors. Half of the resulting transformation mix was plated onto LB plates with chloramphenicol. The colony-forming units of each transformation plate were counted and the number of colonies obtained through sgRNA exchange with ssDNA was compared to the number of colonies obtained through sgRNA exchange with dsDNA.

### III.4.2. Results and discussion

Overall, pMTL4131511_BM8 (vFS25) showed the lowest transformation efficiency with only nine colonies obtained with ssDNA and two colonies obtained with dsDNA (Figure 34.a). It is also the only construct which showed a lower transformation efficiency with dsDNA than with ssDNA. pMTL431511_BM10 (vFS27) had the biggest relative difference in CFU with no colonies obtained with ssDNA but 205 obtained with dsDNA. As many as 645 colonies were counted in pMTL431511_BM6 (vFS23) when assembled with dsDNA, while 69 was the highest number of colonies obtained with ssDNA (in pMTL431511_BM4, or vFS21). The cut vector control still gave 4 colonies, which is higher than the colony count obtained by pMTL4131511_BM8 dsDNA and pMTL431511_BM10 ssDNA. Since the same amount of cut backbone vector was transformed in each case, but the backbones in the samples were processed by the HiFi reaction mix, the lower colony count in some samples might indicate that some of the enzymes in the HiFi mix reduce the chance of vector self-ligation.

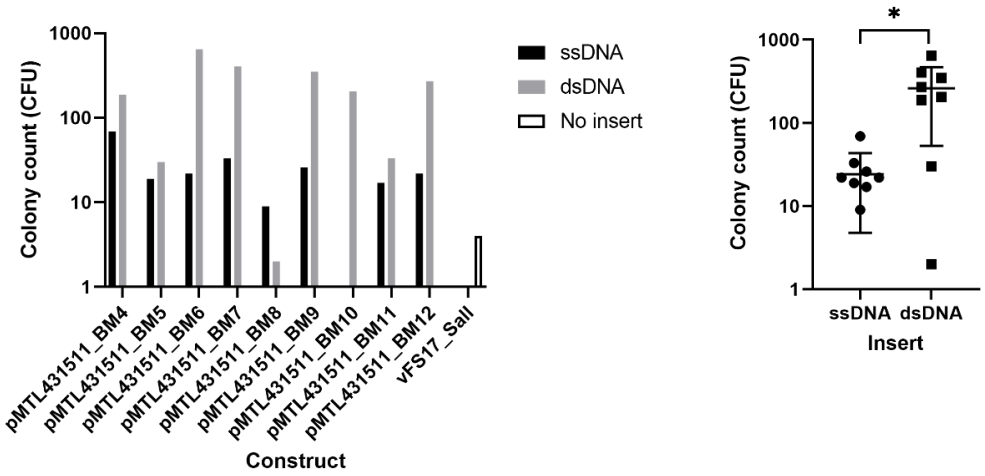



**Figure 34. Colony count of cloning of nine CRISPR vectors assembled by HiFi reaction with either ssDNA or dsDNA to exchange the sgRNA cassette. (a)** Individual colony count obtained for each construct. pMTL431511_BM11 assembled with dsDNA was transformed in only 20 µL of competent cells. **(b)** Average colony count and standard deviation of all constructs assembled with either ssDNA (n=9) or dsDNA (n=8) on a logarithmic scale. pMTL431511_BM11 assembled with dsDNA was excluded from data analysis. No colonies were obtained in pMTL431511_BM10 assembled with ssDNA (not visible on log-scale). * = $p<0.05$ (two-tailed Welsh's t-test).

When grouped together [Figure 34(b)], all the constructs assembled with ssDNA or dsDNA, respectively, averaged 24 ± 19 CFU (n=9) or 261±209 CFU (n=8) (mean ± SD), and differed from 238 ± 74 CFU (mean ± SEM) with a 95% confidence interval ranging from 63 CFU to 412 CFU. This difference was found to be significant ($p<0.05$) under a two-tailed (unpaired) Welsh's t-test. Consequently, it can be concluded that using dsDNA would drastically improve the established sgRNA exchange protocol which was originally based on ssDNA.

## III.5. Complementation of *pyrE* through nine individual Bookmarks

During the complementation step, nine complementation vectors (pMTL431511_BM4 to pMTL431511_BM12, assembly described in Chapter II.4.1.8)(4)) were used to target each of the bookmark protospacers previously knocked-in in the *pyrE* locus, and restore the WT *pyrE* allele in its original locus (Figure 35.a). Each Bookmark complementation vector simply consists of standard CRISPR vector whose sgRNA cassette targets a particular bookmark and whose editing template is the WT *pyrE* allele flanked by the same 1 kb homology arms as the KO vector. Each complementation vector was conjugated into the *ΔpyrE::BMa* strain over the course of three independent experiments, and the resulting transconjugants were patched and screened by colony PCR to ascertain whether they carried the WT or the mutant allele. Complementation efficiency for each bookmark complementation vector was calculated by dividing the number of transconjugants which reverted to the WT allele by the total number of colonies screened [cf. Figure 35(b)].

Unfortunately, the number of colonies screened varied between conjugations due to a combination of many colonies failing to grow after patching and many colony PCR failing. The electrophoresis gels of all colony PCRs are presented in the supplementary materials (cf. Figure S. 2 through Figure S. 10). All constructs were screened with an average and a median of 7 colonies overall, but the range starts at only two and ends at twelve screened colonies for one of the replicates of BM10 and BM6, respectively (cf. Table S. 9). As such, a statistically significant analysis of the results is impossible.

A large variability in the raw number of transconjugant colonies was observed between conjugations, which can easily be explained through experimental bias: the concentration of recipient cells, donor cells, and standard liquid handling errors in each conjugation all contribute to inconsistencies in the raw number of transconjugants across several conjugations. As such, only the complementation efficiency – and not the raw colony count – was used to compare the bookmarks.

The average complementation efficiency was 91%±15%, with BM9 having consistently the maximal efficiency (out of seven, seven, and five colonies screened for each replicate). The lowest efficiency was achieved by BM12, with one of its three replicates achieving only 33% efficiency (only two out of six colonies were successfully complemented); however, the other two replicates of BM12 achieved 100% efficiency (four out of four colonies were complemented in both cases).

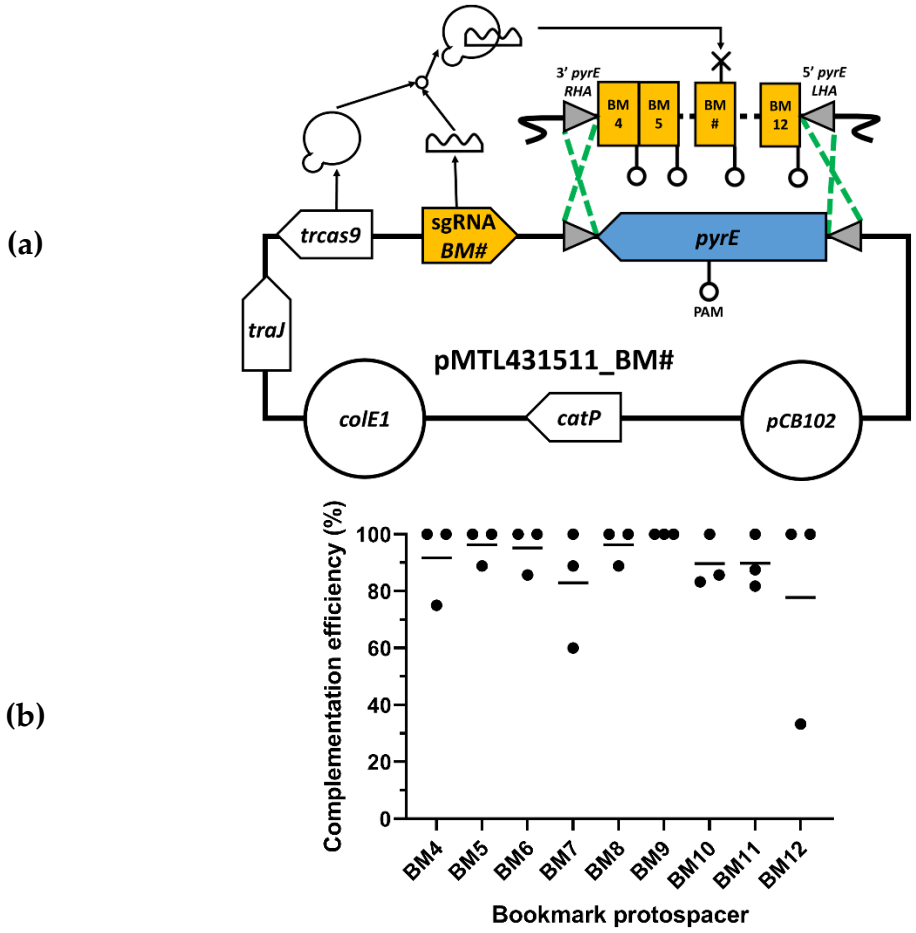



**Figure 35. Integration of the *pyrE* gene back in its original locus using pMTL431511_BM# to target different bookmark protospacers. (a)** Each homology arm consists of a 1 kb region directly upstream (LHA) or downstream (RHA) of the *pyrE* locus, including the first two and the last two codons of the *pyrE* gene, respectively. Together with the *pyrE* gene they constitute the *pyrE* editing template. trCas9 nuclease and sgRNA expression counter-selects *pyrE*Δ::*BMa* transconjugants by cutting the genomic DNA of cells with an intact bookmark protospacer array. **(b)** Complementation efficiency of nine protospacers in *C. autoethanogenum.* Complementation of a *pyrE*Δ::*BMa* strain of *C. autoethanogenum* using different bookmark protospacers was successful in 91 ± 15% of all screened colonies, with little to no variations observed in between the particular heterologous protospacers targeted by each bookmark complementation vectors. (•): complementation efficiency of one separate conjugation, (–): arithmetic mean of the complementation efficiency of three independent conjugations for each bookmark, *TraJ*: conjugative transfer function, *colE1*: Gram-negative replicon, *CatP*: chloramphenicol/thiamphenicol resistance cassette, *pCB102*: Gram-positive replicon, BM#: Protospacer sequence of any bookmark, where # is an integer between 4 and 12.

Overall, these results show that all Bookmark complementation vectors are probably equivalent in their ability to target the protospacers used in each genomic bookmark and to replace them by a gene such as *pyrE*. Much larger genes or genes with particular functions (for example, the genes essential to HR) might suffer from worse complementation efficiencies.

## III.6. Conclusion and perspectives

A new complementation strategy was successfully designed and illustrated. It solves many of the issues inherent to traditional complementation methods such as vector complementation or genomic complementation in a different locus. Because bookmark complementation complements genes in the genome, it has none of the issues of vector complementation related to vector maintenance or overexpression of the gene of interest. Because this genomic complementation of the KO strain is *in situ* rather than in a different locus, it bears no risk of disrupting the WT phenotype by disrupting a second locus. Instead, when the KO strain has been built as intended, bookmark complementation should always reproduce the WT phenotype – which is precisely what is expected of a complemented strain. As such, whenever the phenotype of the bookmark-complemented strain is different than the phenotype of the WT, one can deduce with near certainty that an unknown feature is affecting the experimental design. This feature could be an unknown mutation in the genome of the KO strain, or a discrepancy in the way the phenotype is measured in each strain.

To enable the unambiguous identification of the bookmark-complemented strain, it is recommended to introduce a watermark in the complemented allele in the shape of a silent mutation – a mutation which is associated with the same aa as the WT codon. Such a mutation spread out on several pre-determined codons is unlikely to happen by chance, but easy to introduce in the complementation process, and it should not affect the phenotype of the strain.

One possible drawback of characterising genes using bookmarks is their potential lack of modularity. In the scenario briefly described previously of a gene whose function is intrinsically linked to its locus (for example, a sporulation gene), Bookmark complementation would enable perfect complementation, but it would not discriminate between the function of the gene and its locus. A different complementation strategy would fail to reproduce the WT phenotype, but this failure might help researchers identify the reliance of the gene on the locus to be functional. It also means that if that gene was to be used in an arbitrary locus in another organism (*e.g.,* in a metabolic engineering context), the gene of interest might not deliver its expected function. However, these are niche concerns which would still arise using a conventional complementation method.

The usefulness of bookmarks is not limited to complementation strategies. Such convenient genomic targets could be exploited for *in vivo* assembly pipelines using Cas9-mediated mutagenesis, or they could enable multiplex mutagenesis mediated by a single sgRNA, in a similar strategy as the one employed by Finnigan and Thorner (2016) [256]. Bookmarks could also be exploited as standard targets for *in vivo* characterisation of different CRISPR-associated nucleases. Indeed, these different Cas are often characterized *in vitro* on rationally designed standard targets, but rarely *in vivo* – where researchers must rely on naturally occurring genomic targets with different loci, PAM and protospacer sequences [198,257–261]. This can make unbiased comparison of different nucleases difficult. Using bookmarks as standard genomic targets would standardise the comparison of different CRISPR-associated nucleases. For example, the efficiency of Cas9 proteins engineered to recognize alternative PAMs could be characterized *in vivo* on the same bookmark protospacer simply by changing the PAM, keeping all other variables (locus and protospacer sequence) constant. BM8 is a good candidate for this strategy, as it incorporates a restriction site which could be targeted in *in vitro* screening using commercially available restriction enzymes [175].

Some problems inherent to complementation have not been addressed by bookmark complementation: indeed, with each cloning step arises the risk of random mutagenesis events; and the occurrence of Cas9-mediated off-targets mutations, albeit rare, cannot be ignored – especially in organisms with a functional NHEJ DNA repair pathway. This latter concern could be alleviated by using engineered Cas9 or sgRNA structures with increased specificity [199,201,202,224,233,262,263]. The dependence on a cloning step is unfortunately shared with all complementation methods. The current complementation standard in *Clostridium* is ACE complementation, which requires two successive cloning steps (one to engineer a *ΔpyrE* strain and the other to generate the actual complementation); as such, bookmark complementation constitutes a significant improvement, with only one cloning step involved. Vector complementation requires less re-streak steps since it does not entail losing the complementation vector – but it also demands that the WT and KO strain be transformed with an empty vector, which offers more opportunities for undesired mutations. Ultimately, whatever the complementation method, the only way to ensure the absence of undesired mutations in the genome of the complemented strain is to proceed to whole genome sequencing.

To be entirely satisfactory, the characterisation of BM4 through BM12 should have included a phenotypic test of each complemented strain (growth on uracil-deprived medium) as well as a sequencing of the genome of each strain to estimate the probability of ancillary mutations when using each bookmark. However, these goals were seen as accessory to the demonstration of the fundamental principle of bookmark complementation, and they were deemed of little practical interest. Indeed, the *pyrE* mutant had already been extensively characterised by our group, and any ancillary mutation detected through whole genome sequencing would have given little information on the probability that this mutation would occur in another experiment involving the same bookmark. It was thus decided to publish the results obtained at that stage of the project, and to dedicate ourselves to the next chapter of this thesis. It resulted in the following publication in *Genes*:

Seys, M.F.; Rowe, P.; Bolt, E.L.; Humphreys, C.M.; Minton, N.P. A Gold Standard, CRISPR / Cas9-Based Complementation Strategy Reliant on 24 Nucleotide Bookmark Sequences. *Genes (Basel).* **2020**, *11*, 1–8.

# Chapter IV: Protocol development

## IV.1. Introduction

A basic, yet important, step in synthetic biology is the transfer of a genetic construct into a target population of cells. The effectiveness of this step is commonly evaluated by counting the colonies which managed to grow after being plated onto a selective medium. Only the cells which have successfully received the genetic construct can grow to form a colony. The relative toxicity of a genetic construct can often be observed at this stage, in the form of a low number of colony forming units (CFU). It is a crucial step, because there is no point in designing a state-of-the-art genome editing tool if it cannot be reliably transferred into the cells. Ideally, to facilitate experimental design, the yield of each genetic transfer should also be consistent across experiments.

This chapter is stemming from a meta-analysis of all the experiments of this thesis which involved a step of genetic transfer into *C. autoethanogenum.* It became apparent that the bacterial conjugation protocol that I was using was incomplete, as it had yielded widely different numbers of colonies, and it had even led to many total failures. Worse, the protocol was producing fewer colonies after each execution, to the point where whole experiments sometimes had to be cancelled. Eventually, in an effort to recover a yield of transconjugants which would be both high and consistent across experiments, the conjugation protocol was itself investigated in a series of experiments. These efforts finally resulted in a new protocol which consistently delivers a high yield of transconjugants, with a 100- to 1000-fold increase in yield and conjugation efficiency relative to some of my worst conjugations.

### IV.1.1. High variability of transformation yield

At the beginning of the project, the bacterial conjugation protocol was producing many thousands of colonies per experiment. In fact, it was so effective that it could be inconvenient: plates would be overcrowded with hundreds if not thousands of colonies, which was making counting individual colonies difficult and unreliable. Additionally, large differences in conjugation yields between different constructs went unexplained: was a lower count of transconjugants an indicator of toxicity of that particular genetic construct, or was it just a consequence of the many events which could impact the conjugation yield of any construct? Finally, many conjugations failed for no apparent reasons. Sometimes the experiment failed before the protocol could be completed (for example, the inoculum would not grow); some other times, the experiment would yield very few transconjugants, or even none at all.

At first, these seemed like minor problems which could be worked around; for example, by simply repeating the occasional failed experiment and only analysing ratios of cells within the same conjugation. However, as time went on and unexpected results kept accumulating, the inconsistency of conjugation yield became increasingly disruptive. Even worse: as can be seen in Figure 36, the yield seemed to slowly decrease over time.

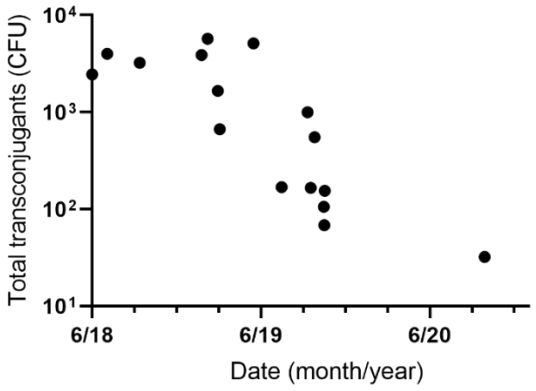



**Figure 36. Yield of WT and *ΔpyrE::BMa C. autoethanogenum* transconjugants over time.** The total number of *C. autoethanogenum* colony forming units (CFU) for a given conjugation is plotted on a log-scale and calculated by extrapolating the fraction of the mating slurry which was plated during the conjugation to the total mating slurry ($CFU_{TOTAL}=CFU_{MEASURED}/V_{PLATED}*V_{TOTAL}$). Whenever available, the transconjugants yield of the same construct (pMTL83151) was used; the yield of the first three conjugations was calculated from the construct with the highest yield (respectively vFS39, vFS23 and vFS23). Conjugations which failed to yield any colonies were not plotted.

It was thus decided to systematically characterize some of the variables which could be affecting the conjugation protocol. To approach these variables critically, a deeper insight into the mechanisms of DNA transfer in prokaryotes might first be desirable.

### IV.1.2. Horizontal gene transfer in prokaryotes

Transfer of genetic material between two cells is a basic feature of life. In prokaryotes, it can serve many purposes: from providing nutrients to spreading phages. As such, many mechanisms of DNA transfer exist. In this section, the best known DNA transfer mechanisms will be briefly described through the lens of their application in biotechnology: transformation, transduction and conjugation. Particular attention will be dedicated to the mechanism of conjugation, which will be described at the protocol-level to provide a reference for the rest of the chapter.

#### *IV.1.2.1) Barriers to DNA transfer*

Many obstacles stand between exogenous DNA and their stable replication in a recipient cell. There are physical barriers such as the cell wall, and enzymatic inhibitors such as the recipient's cell CRISPR system (cf. Chapter I.5.1.2) or restriction-modification (RM) systems.

Briefly, RM systems cut foreign DNA by identifying its methylation pattern. They have been divided into four types based on their enzymatic conformation and requirements, the relation between their cut site and their recognition sequence, and their sensitivity to methylation status [264–266]. Types I to III endonucleases cut non-methylated DNA, while Types IIM and IV cut DNA with a specific methylation pattern. If the DNA double-helix is hemimethylated (for example right after DNA replication), the methyltransferase of RM systems types I to III will methylate the unmethylated strand, protecting this DNA molecule from the activity of their cognate endonuclease. Type IV RM systems do not have an associated methyltransferase.

Finally, once the DNA has successfully penetrated the cell, it has to share some level of genetic compatibility with its host if it is to be replicated successfully [267]. Some parasitic DNA molecules hijack the host's molecular machinery to express the proteins necessary for their insertion into the genome or for their own independent replication. Alternatively, the exogenous DNA could rely on simple sequence homology to integrate the genome of the recipient cells via HR (cf. Chapter I.3.5).

#### *IV.1.2.2) Transformation*

Transformation is the process of acquiring DNA straight from the external environment [268]. It relies entirely on the recipient cell. The ability to transform provides competent bacteria with several advantages: they can use DNA as a source of nutrients, as a mean to increase their genetic diversity, or as a repair template for their own genome. Many bacteria

are naturally competent, but competence can also be induced in the laboratory. Common techniques involve applying a strong electromagnetic field to the cells to disrupt their cell wall (electroporation), or exposing the cells to chemicals and temperature changes which facilitate DNA precipitation and cell wall disruptions (chemotransformation) [269,270]. This allows dsDNA molecules to seep inside the cells, which exposes them to the endogenous RM systems [266]. In contrast, natural transformation occurs through specialized membrane receptors and transmembrane channels which only import ssDNA. Consequently, natural transformation should be less vulnerable to the host's RM systems. It is a versatile method which can transfer linear or circular DNA, ssDNA or dsDNA, and lengths ranging from one base to dozens of kilobases [268].

#### *IV.1.2.3) Transduction*

Transduction is the transfer of genes between bacteria through the intermediary of a phage particle [268]. After lysis of their host's genome, phage particles can sometimes encapsulate some of their host's DNA alongside their own (specialized transduction) or even instead of it (generalized transduction). In both cases the phage particle binds specific membrane receptors and injects the DNA into a new host. As such, each phage can only infect a narrow range of species. Specialized transduction can lead to integration of the foreign DNA into the new host's genome alongside the lysogenic phage.

#### *IV.1.2.4) Conjugation in Clostridium autoethanogenum*

During conjugation, a donor bacterium infects a recipient bacterium with a linearized single-stranded plasmid, through direct contact (a.k.a. mating). The plasmid is then circularized and replicated in the recipient cell to produce a transconjugant cell. It is a complex process but it is the gene transfer technique with the broadest range of recipients and it can transfer the largest DNA molecules (up to several hundreds of kb). Some vectors, called self-mobilizable, carry all the genes necessary to mediate their own conjugation, but so-called mobilizable plasmids require the participation of other vectors to initiate their conjugation into a new host.

Because DNA transfer occurs as ssDNA, conjugation should be protected against RM systems; however, in practice, knocking-out or evasion of the RM system of the recipient cell dramatically improve the yield of conjugation [236,266,271–273].

In *Clostridium* biotechnology, conjugation is mediated by the self-mobilizable plasmid R702 [236,274]. It is a plasmid found in Gram-negative bacteria; it carries resistance genes against

streptomycin, tetracycline, kanamycin, sulphonamides and mercury salts, as well as all the genes necessary to infect *Clostridium* species by conjugation. R702 was conjugated into a commercial *E. coli* strain (NEB Express) which lacks the Dcm methyltransferase. The resulting *E. coli* sExpress strain keeps cloning convenient and protects the DNA it replicates against the Type IV RM system of *C. autoethanogenum* [236]. R702 is able to mobilize any vector which harbours the origin of transfer *oriT*.

#### IV.1.2.4)(1) Kinetic models of conjugation

To help us quantify the relative effectiveness of a given conjugation experiment, a mathematical model of the conjugation process is necessary. In a foundational paper, Levin et al. established that, in exponential phase, conjugation kinetics could be approximated by a mass-action model – i.e., the conjugation rate is directly proportional to the concentration of donor and recipient cells [275]. If the concentrations of cells are held constant in a chemostat, the evolution of the number of transconjugants over time can be described by Equation 3:

$$\dot{T} = \gamma * R * D$$

**Equation 3: Rate of change in the concentration of transconjugants in a chemostat (**[275]**).** Ṫ is the conjugation rate (CFU/min), γ is the transfer rate constant ((CFU*min)$^{-1}$), R is the number of recipient cells (CFU) and D is the number of donor cells (CFU).

In the chemostat equation, the transconjugants population is assumed small compared to the total number of cells; as such its contribution to the conjugation rate is negligible. This is fitting theconjugation protocol, as R702 is not able to replicate inside Gram-positive bacteria; *C. autoethanogenum* is thus unable to spread the shuttle vector to other cells. This model suggests that the transfer rate is a constant which could be used to compare the effectiveness of different conjugations. Additionally, the relation between conjugation rate and the relative abundance of donor cells (R+D=N) shows that the maximal rate of conjugation is achieved when the concentration of donor and recipient cell is equal (equivalent to D/R=1 or D/N=0.5) (Equation 4).

$$\dot{T} = \gamma * R * \frac{N}{N} * D * \frac{N}{N} = \gamma * \frac{R}{N} * \frac{D}{N} * N^2 = \gamma * \left(1 - \frac{D}{N}\right) * \left(\frac{D}{N}\right) * N^2, \qquad N = R + D$$

**Equation 4: Relation between the rate of change in the concentration of transconjugants in a chemostat, relative to the total cell concentration.** Ṫ is conjugation rate (CFU/min), N is the total number of cells (CFU), γ is the transfer rate constant ((CFU*min)$^{-1}$), R is the number of recipient cells (CFU) and D is the number of donor cells (CFU).

However, this simple relation assumes that donor, recipient and transconjugants share the same growth rate, which is not the case of *E. coli* and *C. autoethanogenum.* Additionally, the chemostat is a poor representation of the conjugation conditions. In the protocol, the mating mix of donor and recipient cells is spread on a plate with rich medium and no antibiotics for 20h. Consequently, it resembles more the conditions of a batch than a chemostat. This difference is not trivial: on top of the impact of bacterial growth on the final number of transconjugants already modelled by Levins et al. [275], MacDonald et al. have shown that substrate availability significantly affects transfer rate [276]. To account for the finite amount of nutrients available on the mating plate, Philipsen et al. suggested a model of the transfer rate which is dependent on substrate availability, based on Monod's logistic equation [277] (Equation 5). If the $K_c$ value of the transfer rate equation is much lower than the $K_c$ value of the bacterial growth equation, then nutrient limitation will limit growth well before it affects the conjugation rate, and Levin's model will be an adequate model of conjugation for as long as donor and recipient cells are in exponential growth.

$$\gamma = \frac{\gamma_{max} * S}{K_c + S}, \qquad S\epsilon[0,1]$$

**Equation 5: Expresison of the conjugation transfer rate as a function of nutrient availability.** γ is the transfer rate constant ((CFU*min)$^{-1}$), $\gamma_{max}$ is the maximum achievable transfer rate, S is the relative concentration of the limiting substrate, and $\kappa_c$ is the relative substrate concentration at which half the maximum transfer rate is reached.

In a routine conjugation, the transconjugant colonies are only counted after 20h. Consequently, the conjugation rate cannot be calculated. However, the kinetic models of conjugation presented above can help us anticipate the relation between the final number of transconjugants, and the number of donor or recipient cells (a.k.a. conjugation efficiency).

At the beginning of the mating, the transfer rate is close to its theoretical maximum. At this stage, the conjugation rate is dependent on the density of cells and the relative abundance of donor and recipient strains. The higher the density, and the most similar the concentration of donor and recipient cells, the faster the conjugation. However, the transfer rate γ will fall as the medium is depleted (or as cells leave the exponential phase). Because a higher density of cells in the mating mix would deplete the medium faster, the final number of transconjugants might actually be lower with a higher initial cell density. More specifically, because *E. coli* has a much faster growth rate than *C. autoethanogenum*, *E. coli* can be expected to deplete the medium faster than *C. autoethanogenum* and compete against *C. autoethanogenum* transconjugants. As such, the optimal ratio of *E. coli* to *C.*

*autoethanogenum* might be expected to be lower than the 1:1 modelled by Levin's mass-action model in a chemostat.

#### IV.1.2.4)(2) Calculation of conjugation efficiency

The total number of colonies is a relatable and practical parameter. However, in a process as complex as conjugation, it can also be misleading. Indeed, the concentration of recipient and donor cells involved in the same volume of culture can vary in between each experiments. For this reason, the number of transconjugants are often reported in the literature relative to the number of donor or recipient cells [236,272,278].

Biologically speaking, this ratio of transconjugants to donor cells is most relevant when all colonies are counted while the mating is occurring in order to calculate a conjugation rate [275–277]. However, calculating conjugation rate for each construct during routine conjugations is impractical; consequently, the number of donors or recipients at the start [236,245,272] or at the end [278] of the mating are often used as proxy. Counting the number of donor and recipients at the end of mating makes the most biological sense as it also accounts for the growth of each population (transconjugants, donors and recipients) through cell division during the 20h-long mating. Nonetheless, from a protocol design perspective, the only variable over which the experimenter has control is the volume and concentration of the cell cultures involved at the start of the mating, not the final number of cells at the end of the mating.

#### IV.1.2.4)(3) Initial protocol of conjugation in *C. autoethanogenum*

At this stage, please consult the original conjugation protocol for *C. autoethanogenum* described in Chapter II.4.1.8)(1), page 53. This initial protocol constitutes the starting point for the whole chapter.

### IV.1.3. Experimental design

The goal of this chapter is to find a solution to (1) highly variable conjugation yields and (2) decreasing conjugation yields. It was suspected from the start that the roots of these two problems might be multiple, complex and overlapping. A long process of trial and error was started to isolate and define different explanatory variables. For clarity, the experiments made to investigate each problem have been organized into two stages.

In the first stage, ambiguous or vague parameters of the original conjugation protocol were investigated. Maybe the protocol was just not precise enough, and having stricter guidelines would suffice to increase reproducibility. A series of minor parameters were fixed arbitrarily based on my empirical experience of the protocol. Then, different definitions of conjugation efficiency were explored (per donor or per recipient cells) to facilitate the comparison between different conjugations.

In the second stage, my past conjugation record was critically examined to find hidden parameters which would not be accounted for by rigidly following the protocol. The genome of my *C. autoethanogenum* strains were sequenced in order to find explanations for their different conjugation potential. The impact of different ways of inoculating the sExpress donor strain was measured, and the impact of cryopreservation on *C. autoethanogenum* was questioned.

## IV.2. Optimisation of the conjugation protocol

This section will mostly consider the decrease of my conjugation yield as a consequence of the variability built into the protocol. The underlying hypothesis is that if the conjugation protocol is just tweaked a little, if the execution is slightly more thoughtful or if just a little more data is collected during its execution, then the variations (and decrease) in conjugation yield will disappear – or, at least, it will become more manageable.

### IV.2.1. Immobilisation of minor parameters

Several stages of the original protocol left considerable flexibility to the experimenter. It makes the execution of the protocol more convenient, and it is good enough for most applications where qualitative results are mostly sought. However, this flexibility also makes comparisons between experiments more difficult. As big differences in conjugation yield in between two experiments were expected, it took some time to notice that the conjugation yield had been steadily decreasing. Another consequence of these loosely defined parameters was that recipient and donor strains sometimes failed to grow at the expected speed after inoculation, leading to delays and often cancellations of whole experiments.

The following parameters were thus fixed in a process of trial and errors:

- *Inoculation of C. autoethanogenum from cryostocks:* do not scrape a cryostock and put it back in the freezer. Instead, inoculate a whole cryostock of 200 µL of fresh culture (less than 1 year old) frozen at an $OD_{600}$ in between 0.2 and 1.
- *Subculture of C. autoethanogenum:* instead of diluting your inoculum at 1:100, 1:50 and 1:10, measure $OD_{600}$ and start your aliquot at $OD_{600}$ 0.05. This is usually equivalent to a 1:50 dilution. A backup starting at $OD_{600}$ 0.1 is recommended if the inoculum only reached a final $OD_{600}$<1.0.
- *Inoculation of E. coli sExpress:* inoculate at least 19h before the intended time of subculture to guarantee that the inoculum will reach OD>0.4 by then (it is usually in between $OD_{600}$ 1.0 and 2.0).
- *Subculture of E. coli sExpress:* instead of simply diluting your inoculum 1:100, measure $OD_{600}$ and start your aliquot to an $OD_{600}$ of 0.05. The subculture should reach $OD_{600}$ 0.4 within 2.5 h of inoculation.
- *Mating:* Instead of waiting for 16h-24h, wait for 20h before harvesting the mating slurry.

- *Selection of transconjugants:* Do not flood the plate with 500 µL of pre-reduced PBS and do not immediately plate 100 µL of the resulting slurry accumulating at the bottom of the tilted plate. Instead flood it with 600 µL of PBS, collect the resulting slurry in a microcentrifuge tube. Measure the volume of the slurry using your pipette and normalise it to 600 µL with pre-reduced PBS. Plate 50 µL of the normalized slurry. Calculate the total number of transconjugants by reporting the volume of slurry plated to the total volume of slurry.
- *Controls:* always conjugate an aliquot of recipient cells with WT donor cells (without shuttle vector) to check for biological contamination. Always conjugate an aliquot of recipient cells with pMTL83151 (or an equivalent empty vector) as a positive control.

Most of these suggestions are minor and actually make the protocol more convenient in the long run. For example, subculturing all sExpress strains at the same starting $OD_{600}$ maximises the chance that they will all reach the appropriate final $OD_{600}$ at the same time, allowing the experimenter to process them all at once.

The imperative of inoculating a whole cryostock of 200 µL of fresh *C. autoethanogenum* ensures that the inoculum consistently reaches a measurable $OD_{600}$ before subculturing. Indeed, if a smaller inoculum is used instead, the culture might sometimes not grow at all. This also has implications for the long-term management of strains. As briefly discussed in Chapter IV.3.3.1), it seems that *C. autoethanogenum* does not tolerate cryostocks as well as could be expected. Consequently, cryostocks should be refreshed regularly and special attention should be given to not accumulate single-nucleotide polymorphisms (SNP) and other mutations by serial subcultures of the same lineage.

The penultimate recommendation to normalise the mating slurry to a given volume has two purposes. First, it keeps the viscosity of the mating slurry low; this ensures that the volumes pipetted for plating are accurate and consistent. Second, it is meant to make sure that the same proportion of the slurry is always plated. Indeed, prior to normalisation, the mating slurry can have a volume anywhere in between 150 µL and 450 µL. Plating 100 µL of this original mating slurry without taking the starting volumes into account can thus be equivalent to plating 75% or 22% of the whole population of transconjugants, depending on how quickly the plate has absorbed the PBS. That being said, even after normalising the mating slurry, this step is still expected to be the biggest source of variability in the conjugation protocol. There is no way to know how many of the cells never make it into the slurry and stay immobilized on the mating plate. To improve on this would mean to overhaul

the protocol completely, for example by plating the mating mix on a filter and dipping it into 10 mL of pre-reduced YTF after mating to recover all the cells. Such a modification might increase reproducibility of the protocol but it would also complicate its execution, and it would require much more optimisation than could be justified within the scope of this chapter.

### IV.2.2. Gas transfer precautions

In this section, anecdotal evidence of the impact of gas flow inside the anaerobic cabinet on the conjugation yield is presented. Early in the project, irregularities in the morphology of colonies and the shape of some plates have led us to suspect that differences in the gas exchanges between and within each plate could be impacting conjugation yield significantly. Oxygen contamination is a constant threat when working with obligate anaerobes. As such, strict procedures have been put in place to minimize exposure of the cells to oxygen. Additionally, because *C. autoethanogenum* grows so slowly, plates can sometimes dry out before all the potential transconjugants have had the time to form visible colonies.

#### *IV.2.2.1) Oxygen contamination*

In order to avoid oxygen contaminations inside the anaerobic cabinet, plates were incubated in the anaerobic cabinet overnight instead of the recommended four hours. This was to ensure that no matter how many plates were incubated, the oxygen dissolved in the medium would have the time to transfer out of the plates and into the air of the cabinet, where it could be reduced by a catalyst. However, after several experiments with a large amount of plates failed or produced strange colony morphologies (Figure 37), it was suspected that just incubating plates for longer before plating might have been insufficient.

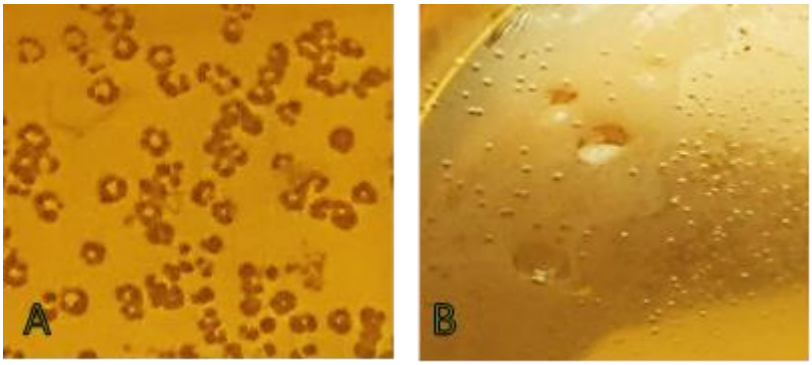



**Figure 37: Atypical morphologies of *C. autoethanogenum* colonies from two transconjugants selection plates (YTF+D-cycloserine+Thiamphenicol). A**= ring-like structure; **B**= transparent colonies. In this particular case, each plate was conjugated with a different plasmid (A=vFS56, B=pMTL83151), but these morphologies have occurred across multiple constructs interchangeably. After more than 5 days, transparent colonies sometimes all acquired pigmentation at the same time.

As an additional precaution, instead of using the sleeves of the anaerobic cabinet to move plates inside, the use of the airlock was generalized every time any number of plates had to be moved inside. This noticeably reduced the occurrence of total failures and strange morphologies, although it did not eradicate them completely.

### *IV.2.2.2) Desiccation*

The inside of an anaerobic biosafety cabinet is subject to a high temperature (37°C) as well as constant airflow. As such, evaporation rates are important enough to visibly desiccate a plate in less than a week. Loss of water can affect the availability and concentration of nutrients and antibiotics in the plate. Because colonies can take 3 to 8 days to grow, the health and number of microbial colonies can be affected by their exposure to the airflow of the anaerobic cabinet. Plates at the bottom of a stack can be expected to lose their water slower than the ones at the top, and plates closer to the edge of the cabinet can be expected to dry faster than the ones at the back of it. This can sometimes be observed at the scale of a single plate, as is presented in Figure 38.

**Figure 38: Spatial gradient of *C.autoethanogenum* colony sizes across a transconjugants selection plate.** The colonies on one side of the plate (right) were noticeably larger and more brightly pigmented than on the other side of the plate (left). A continuous gradient can be observed in between these two extremes. It was suspected that continuous exposure to direct airflow on the left side of the plate might have desiccated that side faster, resulting in higher antibiotic concentrations, higher osmotic pressure and slower colony growth on that side.

Several tactics were attempted to slow down the desiccation rate. Leaving an open petri dish filled with water in the gas cabinet was only mildly effective in slowing down desiccation and very unwieldy. The most effective and convenient technique to slow down desiccation was to seal the plates with parafilm. This does not create a perfect seal, as the parafilm cracks after only a few hours, but it is sufficient to visibly slow desiccation. In an experiment where 50 µL of the same pMTL83151 mating slurry was spread on three plates which were sealed with parafilm and three plates which were kept unsealed, the parafilm seal seemed to decrease the number of colonies by about 20%, but the difference was not statistically significant (Welch's t-test, p=0.07) (Figure 39). Nonetheless, as precaution, parafilm seals should only be used to conserve plates which have already been counted.

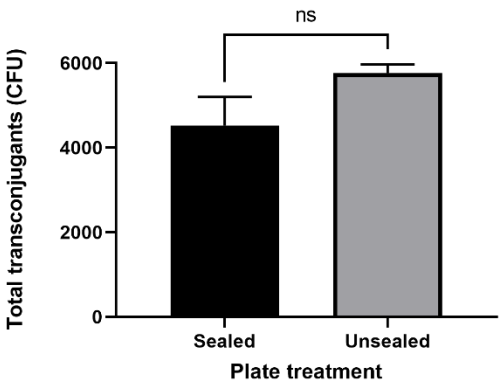



**Figure 39: Effect of sealing pMTL83151 transconjugants selection plates with parafilm.** Sealing plates with parafilm is expected to hinder gas transfer in general, and desiccation in particular. However, there was no significant difference in the number of colonies between plates treated with parafilm and the unsealed controls (p=0.07, n=3, two-tailed Welch's t-test).

In the end, the best technique to keep plates hydrated is to maximize their starting humidity (for example, by keeping the lids on when they solidify in order to keep water vapour inside the plate), and to store them the furthest from the edge of the anaerobic cabinet as possible. Only when all colonies have been counted should the plates be sealed with parafilm (two layers of standard parafilm are recommended for more than 10 days of storage). Special care should be taken to keep all the plates from the same experiment at the same distance from the edge of the gas cabinet to prevent differences in gas transfer from affecting the number of colonies on each plate (Figure 40).

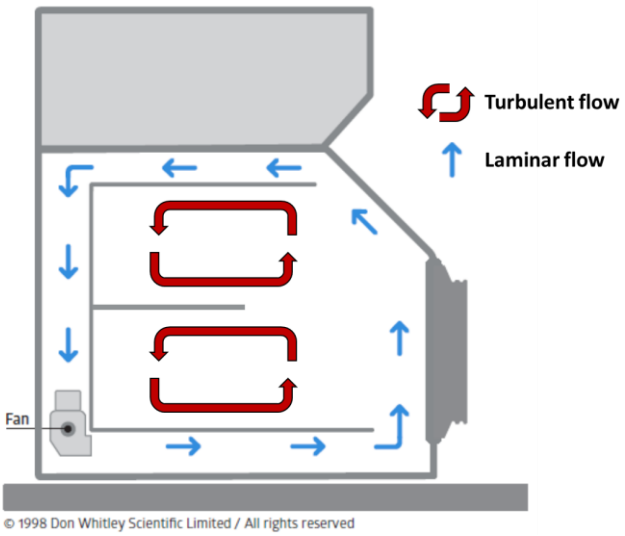



**Figure 40: Schematic of the airflow in the MG1000 anaerobic workstation from Don Whitley Scientific ltd.**

### IV.2.3. Normalisation of conjugation yield

Conjugation is a complex process which depends on many factors. For practical reasons, the original protocol operates with volumes of cell cultures. However, it does not give any $OD_{600}$ requirement for the recipient cells, and it is relatively permissive regarding the $OD_{600}$ of the recipient cells ("in between 0.2 and 0.4"). As such, it was suspected that a large part of the observed variation in the number of transconjugants could come from variations in the starting number of donor and recipient cells.

To test this hypothesis, I measured the number of transconjugants obtained by varying the amount of initial donor and recipient cells. The aim was to find which of donor or recipient populations are limiting in the conjugation protocol, and use this variable to measure conjugation efficiencies across all conjugations.

#### *IV.2.3.1) Experimental setup*

One arm of the experiment examined the impact of varying the amount of recipient cells by mating the same volume of sExpress culture (1 mL) with varying volumes of *C. autoethanogenum* (0.1 mL, 0.2 mL and 1 mL) [Figure 41(a)]. This is the aspect of the experiment which is the most directly relevant to the conjugation protocol, as the concentration of recipient cells is the most likely to change in between different conjugations. In parallel, the impact of varying the amount of donor cells was investigated by mating the same volume of *C. autoethanogenum* culture (0.2 mL) with varying volumes of sExpress (0.5 mL, 1 mL and 2 mL) [Figure 41(b)]. Because the updated protocol instructs that cells are harvested at a specific OD (OD~0.4), this aspect of the experiment should be less directly relevant to the optimisation of the conjugation protocol. After all, the concentration of donor cells should be roughly equal across all experiments if they are all harvested at the same OD. It was still deemed interesting to measure, if only to ensure that the protocol was not unexpectedly sensitive to the concentration of donor cells. Potential interactions emerging from the proportion of recipient to donor cells were also investigated by plotting the number of transconjugants against the ratio of recipient to donor cells, as per Levin's mass action model [275] [Figure 41(c)].

The concentration of live cells present in donor and recipient cultures was measured at the beginning of mating with the 3x6 drop count method detailed in Chapter II.3.1. The same *C. autoethanogenum* culture was conjugated with three aliquots of sExpress stemming from three different single colonies after transformation with pMTL83151, inoculated directly from the sExpress transformation plate. All three sExpress subcultures were harvested in

between $OD_{600}$ 0.3 and 0.4. After 20h of mating, each mating plate was resuspended in PBS and normalised to 0.6 mL. 0.2 mL of the resulting mating slurry was plated on transconjugants selection plate (YTF+D-cyc+Tm) and the number of transconjugants on each plate were counted after four days.

### *IV.2.3.2) Results*

The concentration of cells in each sExpress culture was calculated to be $(4.23\pm1.30)*10^8$ CFU/mL, $(3.92\pm1.22)*10^8$ CFU/mL, and $(4.50\pm0.54)*10^8$ CFU/mL. The concentration of cells in the *C. autoethanogenum* culture was $(2.32\pm0.59)*10^8$ CFU/mL. Following the standard conjugation protocol, where 0.2 mL of recipients are used to resuspend a cell pellet obtained from 1 mL of sExpress, would thus result in a roughly 10-fold excess of donors relative to the population of recipients.

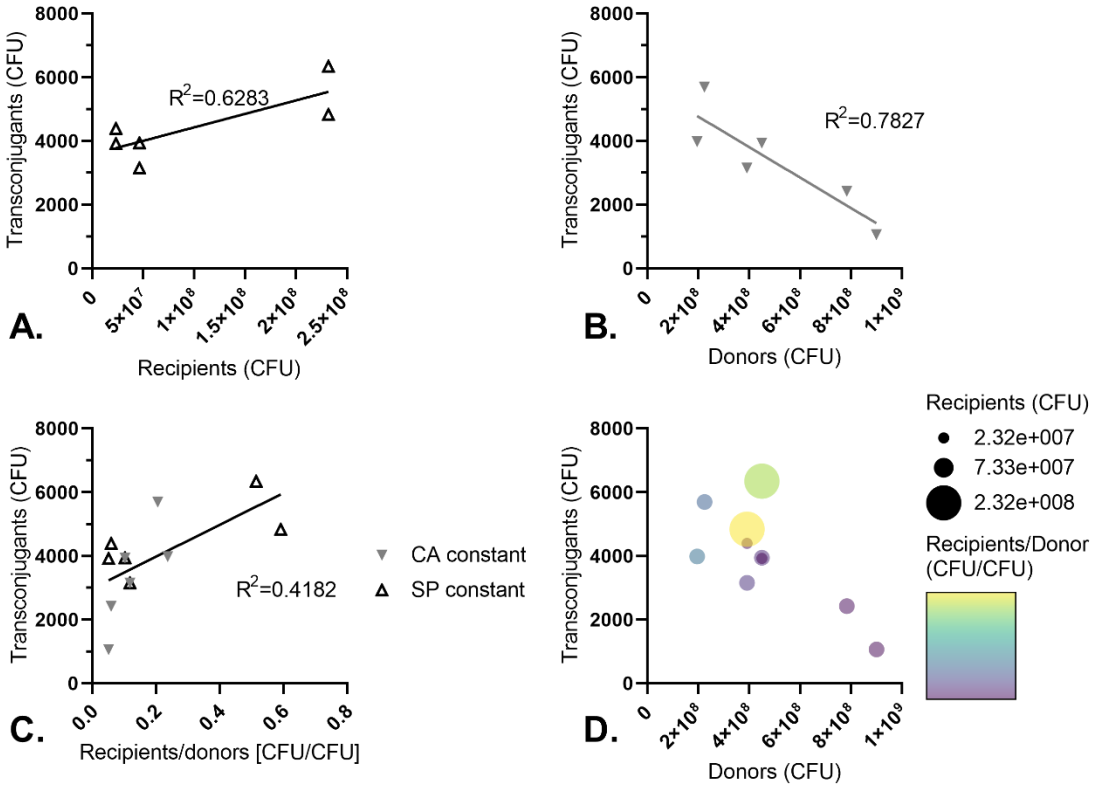



**Figure 41: Relationship in between the number of donor and recipient cells at the start of the mating and the final number of transconjugants**. Each point represents the colony count of the transconjugants from one plate multiplied by the ratio of the total volume of the mating slurry to the plated volume. **(a)** The number of transconjugants has a weak, positive, linear correlation with the initial number recipients ($R^2$=0.63); **(b)** The number of transconjugants has a stronger, negative, linear correlation with the initial number of donors ($R^2$=0.78). **C.** There is no linear relation in between the ratio of initial numbers of recipients and donors ($R^2$=0.41); **D.** Summary of the multivariate analysis.

The first replicate only yielded about 15% as many transconjugants as the other two and was thus removed from the analysis as an outlier.

The remaining transconjugants counts showed a weak positive correlation with the starting concentration of recipient cells ($R^2$=0.63) and a slightly stronger negative correlation with the starting concentration of donor cells ($R^2$=0.78). Indeed, a 10-fold increase in recipient cells (from 0.1 mL to 1 mL) increased the average number of transconjugants by approximately 35%; similarly, a four-fold increase in donor cells (from 0.5 mL to 2 mL) has led to an approximate decrease of 65% in the average number of transconjugants.

#### *IV.2.3.3) Discussion*

##### IV.2.3.3)(1) Kinetic interpretation of the relationship between donor and recipient cells

Interestingly, these trends indicate that the protocol uses an excess of donor cells so large that it is actually detrimental to conjugation efficiency. In Levin's mass action model, this could be interpreted as the consequence of departing from the 1:1 ratio of donors to recipients [275]. This is illustrated by the linear regression in Figure 41.C – although the relationship seems to break down at very low ratios of donors to recipients (R/D~0.05) ($R^2$=0.42). However, Levin's mass action model describes conjugation rates, not the final number of transconjugants; because 20h elapses before the transconjugants are counted, a slower rate of conjugation caused by a high donor to recipient ratio might still have been expected to yield the same number of transconjugants – simply because time is not limiting. Consequently, in this scenario where conjugation rate is mainly influenced by the probability that a random encounter between two cells involves one recipient and one donor, the final number of transconjugants could have been expected to remain constant as long as there is enough excess donors to guarantee that virtually all recipient cells are paired with at least one donor cell.

The decrease in the number of transconjugants as the number of donors increases might thus be better understood through the lens of a model of conjugation limited by nutrients availability [276,277]. In this model, the growth of the donor population would deplete available nutrients before all recipient cells had a chance to undergo conjugation. A larger excess of donor cells would deplete nutrients faster, which would further reduce the time available for mating, and as a consequence the final number of transconjugants. The importance of this process would depend on how sensitive the growth rate and the conjugation transfer rate are to nutrient depletion.

If the rate of conjugation is less vulnerable to nutrient depletion than the growth rate of transconjugants, the negative relationship in between the number of transconjugants and

the number of donors could also be explained solely on the basis of competition during bacterial growth, after the mating is over. In this model, there is enough time and resources that all conjugable recipient cells undergo conjugation. Consequently, the same number of transconjugants is obtained at the end of mating independently from the extent of the excess number of donors. However, once the mating is over, the transconjugant population must compete with the donor population for the remaining nutrients in order to grow by standard cell division. A larger excess of donor cells would reduce further the share of the nutrients available to the transconjugants population, resulting in a slower growth of the transconjugants population.

Of course, stochastic encounters, mating and growth are simultaneous process. As such, all three phenomena probably contribute to the observed relation in between the number of transconjugants, donors, and recipients. Measuring the final number of recipient cells should inform us about which process is dominant. Indeed, if transconjugants are competing with the donors for nutrients, recipient cells should be equally affected. Consequently, if the final number of recipient decreases in the same way as the number of transconjugants when the initial number of donors increases, then nutrient availability is most likely impacting the growth rate of transconjugants more than it is impacting the conjugation rate. If the final number of recipients stays the same or decreases less than the number of transconjugants when the initial number of donors increases, it would mean that nutrient availability is impacting the conjugation rate more severely than it is impacting bacterial growth. Tentatively, the medium of the mating plates could be optimized to prevent competition in between *E. coli* and *C. autoethanogenum*.

#### IV.2.3.3)(2) The raw number of transconjugants is a more reliable metric than conjugation efficiency per final number of donor or recipients

For all its insights into the processes at play within the conjugation protocol, such a large variation (4-fold) in the number of donor cells would not occur during a normal execution of the conjugation protocols. Indeed, all sExpress cultures are usually harvested at the same $OD_{600}$ (in the case of this particular experiment, in between $OD_{600}$=0.3 and $OD_{600}$=0.4). Interestingly, because the relation between donors and transconjugants was negative, decreasing the starting number of donor cells should actually increase the conjugation yield. It would not be practical to implement a large change to the protocol as ~4000 CFU per conjugation is a perfectly sufficient yield and reducing the number of transconjugants would

require additional dilution steps. Rather, donor cells should simply be harvested at $OD_{600}$=0.2 rather than $OD_{600}$=0.4 whenever possible to minimize the excess of donor cells.

The weaker, positive relationship between the final number of transconjugants and the starting number of recipient cells was not directly proportional (Y = 8.468e-006*X + 3581); in fact, it was not significantly different from a horizontal line (F-test, n=6, p=0.06). Consequently, normalising the final number of transconjugants by the number of recipient cells at the start of the mating would have little meaning.

As a matter of fact, it had been decided to start measuring the number of recipient cells shortly after the very first failed conjugations, so the data is available to put this hypothesis to the test. Normalising the number of transconjugants by the starting number of recipient cells actually increases the scatter of the transconjugants rather than decreasing it (cf. Figure 42). Expressed mathematically: the relative standard error of the distribution of the raw number of transconjugants doubles from 112% of the mean to 211% of the mean when it is normalized by the number of recipients (although neither distribution is normal, so this has little meaning). Consequently, correcting for the variability of the number of recipients at the beginning of the mating does not mitigate the variability of the conjugation yield [cf. Figure 42(a)].

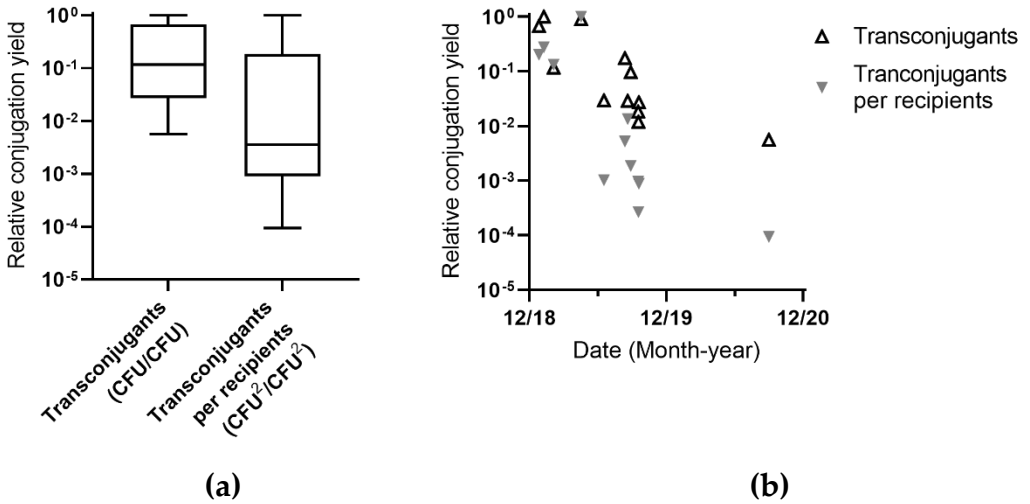



**Figure 42: Comparison between the number of WT and *ΔpyrE::BMa C. autoethanogenum* transconjugants and their associated conjugation efficiencies per recipients for all successful conjugations of pMTL83151 on a log-scale (a) in a boxplot and (b) over two years.** Each series was normalized by its highest value. The whiskers of the boxplot represent the range in between the minimum and the maximum; the box represents the range in between the 25th and 75th percentiles; the middle horizontal line is the median.

On the other hand, the conjugation efficiency per recipient decreases faster than the raw conjugation yield [cf. Figure 42(b)]. This might be a hint of some causal relationship: maybe a slow process – such as the potential degeneration of *C. autoethanogenum* cryostocks – is

gradually impairing the ability of the *C. autoethanogenum* strain to be conjugated. Alternatively, the faster decrease in conjugation efficiency per recipient might only share a common cause with whatever is actually decreasing the raw conjugation yield. It is impossible to say at this stage.

#### IV.2.3.3)(3) Alternative metrics to measure conjugation efficiency

In respect to finding a parameter with which to normalize and compare the conjugation efficiency of different experiments, the number of recipient cells at the time of transconjugants selection might be a more relevant metric than at the start of the mating. It is more difficult to measure when several conjugations are made in parallel, but, as previously discussed, it should account for the bacterial growth of recipients and transconjugants alike. However, if nutrient availability affects the conjugation rate more than it affects the growth rate of the recipient and transconjugants, then the relation between transconjugants and the final number of recipients might not be linear either. For this reason, some authors use instead the transconjugants count of a standard empty vector as an internal control. This is even more convenient than measuring the number of recipient cells at the start of mating and it should be effective when comparing different constructs, but it does not allow to investigate variations in the absolute conjugation efficiency (i.e., what is affecting the conjugation efficiency of the internal control?). The gold-standard to compare different conjugations is probably to measure and calculate their respective kinetic constants ($\gamma$, $K_c$ or $V_{max}$) for the conjugation of each vector, but this is an unrealistic expectation for routine conjugations.

#### IV.2.3.3)(4) Limitations of the study

*A posteriori*, the main limitation of this experiment was that the range of the number of donors and recipients was too narrow – or, at least, too high. A follow-up experiment exploring the impact of 10-fold dilutions on donor and recipients might be even more informative. Dilutions of donor cultures might allow us to approach the concentration at which the maximal yield of transconjugants could be achieved, even if there would be little practical use for this information since the yield is already good enough. Dilutions of the recipient culture, on the other hand, might have allowed us to cover the low end of the range of possible starting concentrations of recipients. Indeed, cultures of recipient cells rarely exceed $OD_{600}$~1 (which is a scenario reproduced in this experiment) but sometimes stay stagnant at $OD_{600}$~0.05 (which was not reproduced here).

Another weakness of the experimental setup is that the number of transconjugants obtained was much higher than anticipated. The actual number of colonies counted per plate ranged from 354 to 2115, which is much higher than the 50 to 250 CFU usually recommended for a reliable cell count. Reproducing this experiment by plating less mating slurry or plating several dilutions of the mating slurry might improve the resolution of the data and give stronger correlations.

#### *IV.2.3.4) Conclusion*

In summary, the conjugation protocol was found to be relatively robust to large variations in the starting number of donor (4-fold) and recipient cells (10-fold). By extension, small variations in the starting number of donor or recipient cells cannot explain the large inconsistencies that were observed in the conjugation yield over time. The relation between the number of transconjugants and the starting number of recipients or donors was found to not be directly proportional; consequently, these are not good variables to use in order to normalize and compare the number of transconjugants of routine conjugations. However, these observations might differ at lower ranges of cell concentrations. The final concentration of recipient cells, although less convenient to measure, should be a better metric to normalize the final number of transconjugants – but this too should be tested. Using an internal control to compare the conjugations of efficiencies of various constructs across experiments might be a better compromise in between accuracy and convenience.

## IV.3. Exploration of new parameters affecting conjugation efficiencies

More precise conjugation parameters, stricter gas transfer precautions, and a better awareness of the impact of the relative number of donors and recipients did not increase the yield or even the reproducibility of conjugation efficiency. More fundamental parameters had to be investigated, parameters which would have been ignored in the original protocol and thus could not be fixed just by improving its execution. Hidden parameters were sought by critically comparing successful and unsuccessful conjugations and reviewing any common pattern outside of the procedures described in the protocol.

A hypothetical genetic drift of the *C. autoethanogenum* stocks strain was first investigated by whole genome sequencing, along with all the *C. autoethanogenum* variants used in the project. In parallel, the consequences of inoculating the donor strain from a cryostock or from a transformation plate was scrutinized. These two experiments then led us to question if cryopreservation could also have an impact on the recipient strain.

### IV.3.1. Comparison of the conjugation potential of different *C. autoethanogenum* strains

One of the *C. autoethanogenum* strains was seemingly unaffected by the decrease in conjugation yield. This pattern was investigated further by sequencing the genomes of all my strains. This led us to uncover unexpected mutations in some strains, but ultimately did not explain the decrease of the conjugation yield.

#### *IV.3.1.1) CO-adapted C. autoethanogenum strain is more easily conjugated*

As illustrated in Figure 43, although the conjugation yield of WT and *ΔpyrE::BMa C. autoethanogenum* was decreasing over time, a third strain of *C. autoethanogenum* seemed largely unaffected by this phenomenon: *ΔpyrE::ACE,* a strain of *C. autoethanogenum* engineered by Dr. Christopher Humphreys to be used in the ACE complementation strategy [174,279] (cf. Chapter I.4.3. Allele Coupled Exchange). It originates from the so-called *col.3*, a *C. autoethanogenum* strain which was first evolved to grow faster on carbon monoxide by Dr. Anne Henstra. Dr Christopher Humphreys truncated the *pyrE* gene of *col.3* so that it was no longer functional, but could be fully restored through a successful round of ACE complementation. *C. autoethanogenum ΔpyrE::ACE* was used in the context of the multiplex metabolic engineering attempt of Chapter VI in order to facilitate an eventual complementation step.

Importantly, the *ΔpyrE::ACE* strain was not known to behave any differently than WT under heterotrophic conditions, or to have any different potential for conjugation. Consequently, it took us some time to notice that this strain might actually be intrinsically easier to conjugate than the WT.

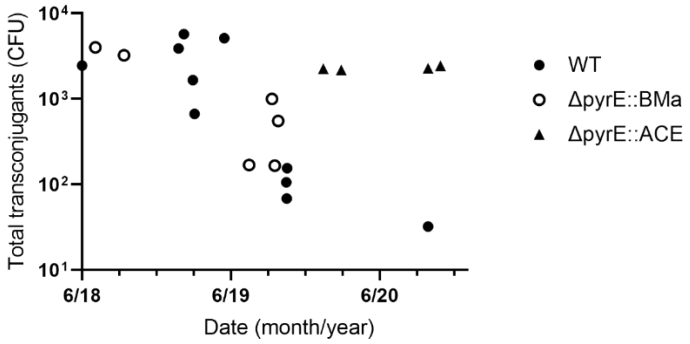



**Figure 43: Yield of WT, *ΔpyrE::BMa* and *ΔpyrE::ACE C. autoethanogenum* transconjugants over time.** The total number of *C. autoethanogenum* colony forming units (CFU) for a given conjugation is plotted on a log-scale and calculated by extrapolating the fraction of the mating slurry which was plated during the conjugation to the total mating slurry ($CFU_{TOTAL}=CFU_{MEASURED}/V_{PLATED}*V_{TOTAL}$). Whenever available, the transconjugants yield of the same construct (pMTL83151) was used; the yield of the first three conjugations was calculated from the construct with the highest yield (respectively vFS39, vFS23 and vFS23). Conjugations which failed to yield any colonies were not plotted.

In order to test this hypothesis, the vector pMTL83151 was conjugated into WT and *ΔpyrE::ACE* in parallel (Figure 44). No biological replicates were made, but each mating slurry was plated in three technical replicates (3x50 µL were plated from the same mating slurry).

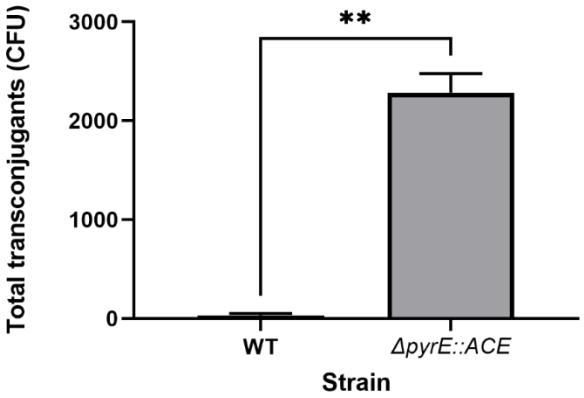



**Figure 44: Comparison of the total conjugation yield of pMTL83151 in two *C. autoethanogenum* strains.** The total number of *C. autoethanogenum* colony forming units (CFU) for a given conjugation is calculated by extrapolating the fraction of the mating slurry which was plated during the conjugation to the total mating slurry ($CFU_{TOTAL}=CFU_{MEASURED}/V_{PLATED}*V_{TOTAL}$). The difference in between WT and *ΔpyrE::ACE* is highly significant ($p<0.005$,n=3, two-tailed Welch's t-est).

The difference in between *ΔpyrE::ACE* and WT *C. autoethanogenum* was highly significant, with *ΔpyrE::ACE* conjugating approximately 70 times better than WT. The concentration of

recipient cells of each strain was roughly equivalent ([1.4±0.2]*$10^8$ and [0.8±0.1]*$10^8$ for WT and *ΔpyrE::ACE*, respectively*)*.

Tentatively, the cause of the decrease of conjugation efficiency in the WT strain could be unravelled by sequencing the two strains.

#### *IV.3.1.2) C. autoethanogenum WT cryostocks have not genetically degenerated*

Because the conjugation yield of the WT strain used to be comparable to the *ΔpyrE::ACE*, it was suspected that the cryostock might have degenerated over time. Potentially, storage at -80°C was somehow selecting a subpopulation of mutant cells which were less receptive to conjugation. In other words, maybe the WT stock had evolved out of the WT genotype despite my best efforts to keep the strain pure. Conversely, *ΔpyrE::ACE* might benefit from mutations which boost survival in the cryostock, or might just have originated from a strain less degenerated than my WT. In order to verify these hypotheses, the whole genome of the WT stock was sequenced along with *ΔpyrE::ACE* and *ΔpyrE::BMa.*

The next-generation sequencing (NGS) was undertaken at the Deepseq facility at the University of Nottingham, and the sequence analysis was achieved using CLC Genomics Workbench (v.20.0.4, CLC bio, Denmark). The reads were mapped to the published genome of *C. autoethanogenum* (NCBI CP012395.1.) [240]. The original reads (SRA SRR2969415) used for the re-sequencing of the published genome were also mapped as a reference for the coverage and the variants identified among the reads of the WT, *ΔpyrE::ACE* and *ΔpyrE::BMa* strains. Variants and regions without coverage were identified for each strain. All the variants shared with the reference reads were discarded as sequencing artefacts. More details on the Whole Genome Sequencing (WGS) data processing are available in Chapter II.5.

In contradiction to my main hypothesis, the WT strain matched the reference genome perfectly. In conclusion, genetic drift of the cryostocks cannot be the reason why the conjugation yield of WT *C. autoethanogenum* is plummeting. A detailed analysis of the WGS data is presented in the following section, but focuses on the variants identified within *ΔpyrE::ACE* and *ΔpyrE::BMa.*

### *IV.3.1.3) Analysis of the WGS of ΔpyrE::ACE and ΔpyrE::BMa strains*

#### IV.3.1.3)(1) Genomic regions without reads coverage

The same three regions without coverage were found among the reads of both *ΔpyrE::ACE* and *ΔpyrE::BMa* (Table 8). Two of these are also in common with the reads of the WT and the reference strain (SRA SRR2969415). These were highly repetitive regions; as such, the low coverage is likely an artefact resulting from the short reads (<300 bp) used in Illumina NGS technology. Unsurprisingly, the third genomic regions without coverage which was common to *ΔpyrE::ACE* and *ΔpyrE::BMa* was located in the *pyrE* locus (CLAU_RS07060).

**Table 8: Regions without coverage within four strains of *C. autoethanogenum.*** Each region is identified by its locus (CLAU_RS), and the start of the region without coverage and its size in bp are specificed for each strain.

| | Coverage=0 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Locus** | **Reference** | | **WT** | | ***ΔpyrE::BMa*** | | ***ΔpyrE::ACE*** | |
| **CLAU_RS** | **Start** | **Size (bp)** | **Start** | **Size (bp)** | **Start** | **Size (bp)** | **Start** | **Size (bp)** |
| **17590** | 3872231 | 1508 | 3873109 | 127 | 3873063 | 182 | 3872040 | 1170 |
| **19680** | 4317568 | 998 | 4317539 | 921 | 4317534 | 1070 | 4317448 | 1287 |
| **07060** | | | | | 1579704 | 318 | 1579744 | 24 |
| **12230** | | | | | | | 2672510 | 30 |
| **19175** | | | | | | | 4203316 | 269 |

However, two additional regions without coverage were found in *ΔpyrE::ACE*. The first one (CLAU_RS12230) can be confidently discarded as another sequencing artefact, because its coverage in the other genomes is also very small: consistently lower than 3. Yet, the second additional region without coverage (CLAU_RS19175) is likely significant, as it is extensively covered in other genomes (>100). It corresponds to the first 70 bp directly upstream of *nupC* and its first 67 codons. If confirmed, such a deletion is likely to knock-out the gene.

In *E. coli*, NupC was shown to participate to the transport of extracellular pyrimidine nucleosides inside the cell [280]. Consequently, the deletion of *nupC* should limit the availability of pyrimidines nucleotides inside the cell when nucleosides are supplied in the medium. It is thus surprising that such a mutation would have been selected in a *ΔpyrE::ACE* chassis, which is unable to produce its own pyrimidine nucleotides and must import them from the medium [281,282]. Tentatively, in *C. autoethanogenum,* NupC might function as a pyrimidine exporter under certain conditions. In this case, *ΔnupC* could help a *ΔpyrE* strain conserve high intracellular pyrimidine concentrations. Interestingly, Wu and colleagues [283] have shown that an *E. coli ΔnupC ΔnupG* strain optimized for uridine synthesis secreted less

uracil derivatives, which might support this hypothesis. Alternatively, *C. autoethanogenum* NupC might facilitate the transport of orotate and FOA inside the cell. In this case, the *ΔnupC* mutation might have been selected by exposure to FOA during the engineering of the strain.

#### IV.3.1.3)(2) Variants detection

Analysis of the five variants found in *ΔpyrE::BMa* (Table 9) showed that only one of them (*hprK*, R170L) was not shared with *ΔpyrE::ACE.* The unique *hprK* variant has only a frequency of 57% (all other variants occurred in 100% of the reads covering their respective locus); this means that the mutation probably occurred after the genetic engineering of the strain, perhaps during the growth in liquid medium for the preparation of the cryostock. HprK participate to the regulation of the central carbon metabolism [284] and is thus unlikely to have any relation with conjugation efficiency.

**Table 9: Variants detection within three strains of *C. autoethanogenum.*** The variants shared with the reference reads (SRA SRR2969415) and the variants in the *pyrE* locus were filtered out for clarity. The nature of each variant is specified in the column "Mutation" by (N/M) where N is the reference base and M is the variant allele. The presence of a variant in a given strain is represented by a 1 (in a red cell), while the absence of a variant (i.e. the WT genotype) is represented with a 0 (in a green cell). If the coding region associated with a variant was not associated with a known gene, the closest related gene was identified using BLASTx [255].

| Position | Mutation | Amino- acid change | WT | *ΔpyrE::BMa* | *ΔpyrE::ACE* | Annotation |
|---|---|---|---|---|---|---|
| **305800** | C/A | R170L | 0 | 1 | 0 | hprK |
| **483403** | C/A | D175Y | 0 | 0 | 1 | hydA |
| **611671** | G/A | D86N | 0 | 0 | 1 | acoR |
| **1649394** | G/T | | 0 | 1 | 1 | tRNA-Gly-TCC-1-3 |
| **2570271** | G/A | A71V | 0 | 1 | 1 | pssA |
| **2605689** | C/A | L46I | 0 | 0 | 1 | CLAU2375 |
| **3138683** | G/T | V167L | 0 | 0 | 1 | RNA-methyltransferase |
| **3330660** | C/A | P37H | 0 | 0 | 1 | csrA2 |
| **3469169** | A/C | Q52P | 0 | 1 | 1 | spoA |
| **3469310** | C/A | A99D | 0 | 1 | 1 | spoA |
| **3480533** | A/- | | 0 | 0 | 1 | Non coding |
| **3778483** | G/T | M110I | 0 | 0 | 1 | Hypothetical |
| **3837709** | G/T | | 0 | 0 | 1 | yycG |

That all the other four variants identified in *ΔpyrE::BMa* were shared with *ΔpyrE::ACE* hints at a common ancestry between the two strains. To understand better this relationship, the variants profile of *ΔpyrE::BMa* and *ΔpyrE::ACE* were compared to the variant profiles of *C. autoethanogenum col.3* – the ancestral strain of *ΔpyrE::ACE* – in addition to the hypothetical "progenitor" strain of *col.3* as presented in the thesis of Dr. Peter Rowe [285].

The comparison between the strains, described in Table 10, reveals that *ΔpyrE::BMa* was probably generated from a "progenitor" background (and not from a WT background as claimed in Chapter III:. As expected, *ΔpyrE::ACE* shares all the variants of its ancestral strain (col.3); however, it has also acquired four additional SNPs of its own. The mutation in position 3837709 is silent, and as such probably has no effect on the phenotype. The mutations in positions 3138683 and 3330660 are not silent and both affect mRNA-binding proteins. Three variants from *col.3* and two variants from the progenitor could not be confirmed in *ΔpyrE::BMa* or *ΔpyrE::ACE* with absolute certainty because they occurred in regions with low coverage (<10 to 0, cf. Table 8), but it can be assumed that these variants also exist in *ΔpyrE::BMa* and *ΔpyrE::ACE* – or that they do not actually exist in *col.3* and the progenitor strains either.

From the variants identified in Table 9 and the outline of their probable function presented in Table 10, it is not obvious why *ΔpyrE::ACE* would conjugate better than the WT strain. The variant in position 3837709 (*yycG*, cell wall remodelling during cell division) is interesting, but the mutation is silent; as such, it should not produce any phenotypical difference. The variant in position 3330660 (*csrA2*), however, might be worth investigating for its involvement in mRNA stability and stress response, which might be relevant during conjugation.

If the causes between the different conjugation phenotypes of *ΔpyrE::ACE* and WT are not genetic, they might just be phenotypical [286]. Indeed, identical cells can still exhibit different phenotypes if they are in different regulatory states. Under this hypothesis of population heterogeneity, a subpopulation of WT *C. autoethanogenum* which is not receptive to conjugation might have been selected even though no mutation had taken place. If this hypothesis is correct, then the loss of conjugation yield might be reversible through consecutive rounds subculturing. This might give the opportunity to the poorly conjugable population to diversify the regulatory states of its individual cells and leave the state which is poorly receptive to conjugation. Interestingly, such a sensitivity to regulatory states could also affect the donor population [287,288]. After all, in a conjugation, the donor strain does most of the work.

**Table 10: Comparison of variants observed in *ΔpyrE::BMa* and *ΔpyrE::ACE* with the ancestral strain of *ΔpyrE::ACE* (col.3) and the hypothetical ancestral strain of col.3 (Progenitor) as described in Dr. Peter Rowe's thesis** [285]**.** Three variants had too low coverage to be part of the variants detection summarized in Table 9 but were checked manually and added here to enable a full comparison with the ancestral strains: the coverage was either <10 (*****) or absent (**x**). The presence of a variant in a given strain is represented by a 1 (in a red cell), while the absence of a variant (i.e., the WT genotype) is represented with a 0 (in a green cell). The putative function of the gene in which the variant was determined using BLASTx [255].

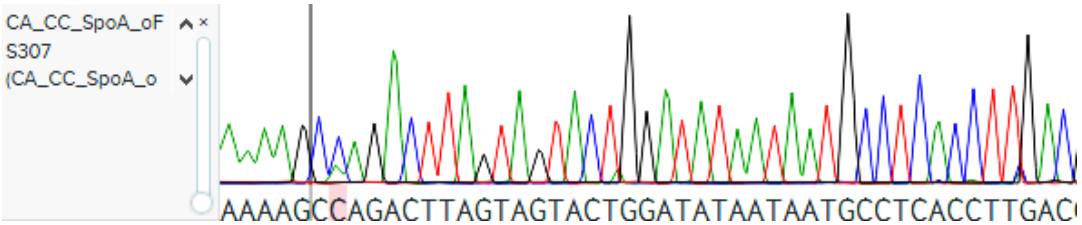
| Position | Progenitor | *ΔpyrE::BMa* | Col.3 | *ΔpyrE::ACE* | Putative gene function |
|---|---|---|---|---|---|
| **305800** | 0 | 1 | 0 | 0 | Regulation of carbon metabolism. |
| **483403** | 0 | 0 | 1 | 1 | Crp/FnR transcriptional regulator. |
| **611671** | 0 | 0 | 1 | 1 | Transcriptional activator of acetoin dehydrogenase operon. |
| **1649394** | 1 | 1 | 1 | 1 | tRNA. |
| **2570271** | 1 | 1 | 1 | 1 | Phosphatidyltransferase, lipid metabolism. |
| **2605689** | 0 | 0 | 1 | 1 | Inosine-5'-monophosphate dehydrogenase; purine metabolism. |
| **3138683** | 0 | 0 | 0 | 1 | 16S rRNA methyltransferase (uracil(1498)-N(3))-methyltransferase). |
| **3330660** | 0 | 0 | 0 | 1 | Key translational regulator; binds mRNA (usually 5'UTR) to regulate translation initiation and/or mRNA stability; mediates global changes in gene expression, shifting from rapid growth to stress survival. |
| **3469169** | 1 | 1 | 1 | 1 | Sporulation regulator. |
| **3469310** | 1 | 1 | 1 | 1 | Sporulation regulator. |
| **3480533** | 0 | 0 | 1 | 1 | Unknown. |
| **3778483** | 0 | 0 | 0 | 1 | Unknown. |
| **3837709** | 0 | 0 | 0 | 1 | Essential sensor histidine kinase; coordinates cell wall remodeling with cell division in Gram-positive bacteria. |
| **4053355** | 0 | 0* | 1 | 1* | Non-coding. |
| **4317806** | 1 | x | 1 | x | Collagen-like protein. |
| **4318262** | 1 | x | 1 | x | Collagen-like protein. |

### IV.3.1.3)(3) Identification of the progenitor cryostock

Interestingly, the WT strain matched the reference genome but the *ΔpyrE::BMa* strain – which should have been generated in a WT chassis – actually matched the profile of the hypothetical progenitor strain. This means that at least two different WT *C. autoethanogenum* stocks are available in the lab: one is the progenitor strain, and the other is true WT. In order to discriminate between the true WT stocks and the stocks of the so-called progenitor strain, the primers oFS307 and oFS308 were designed to screen the *spoA* locus and cover the SNPs in positions 3469169 and 3469310 (cf. Table S. 6).

The primers oFS307 and oFS308 were thus used to screen the *C. autoethanogenum* WT strain kept in the culture collection. They revealed a mixed base (with the expected A/C mutation) in position 3469169 9 (Figure 45), but no mutation in position 3469310. The original stock in the culture collection thus seems to be partially mutated already, but it is not yet showing the full progenitor genotype.

**Figure 45: Trace of the Sanger sequencing read of the *spoA* locus of the *C. autoethanogenum* WT stock from green cryostock of the culture collection (CC3142).** A mismatch in position 3469169 (red) exhibits a partial peak with from correct base, betraying a mixed culture of pure WT and mutant.




I recommend that all existing stocks of *C. autoethanogenum* WT be screened with oFS307 and oFS308 before use in order to confirm the strain background. More straightforwardly, all stocks could be replaced by my personal *C. autoethanogenum* WT stock which was just shown to be identical to the reference genome.

### IV.3.2. SExpress inoculation method impacts conjugation yield.

Having eliminated genetic drift as a potential cause of loss of conjugation yield, my attention shifted to factors which could affect the epigenetics of the donor or recipient cells. After careful examination of my experimental records, I realized that a large drop in conjugation efficiency occurred when I started inoculating the sExpress donor strains from cryostocks (Figure 46).

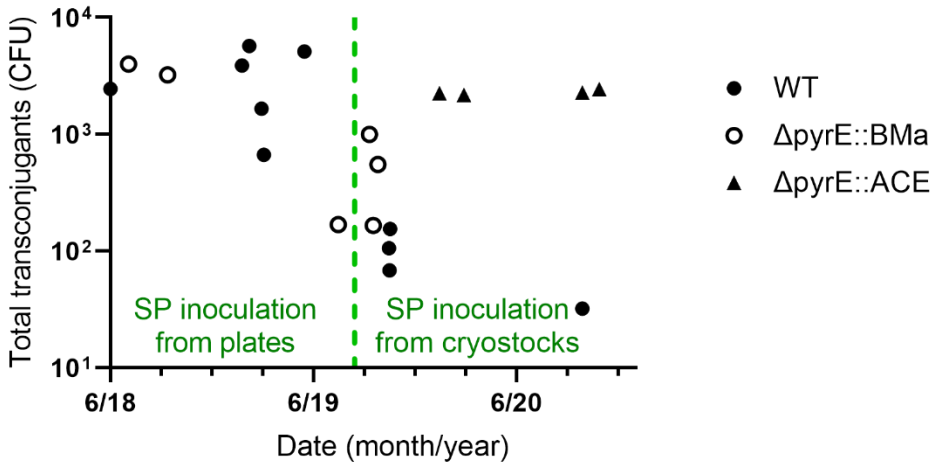



**Figure 46: Yield of WT, *ΔpyrE::BMa, and ΔpyrE::ACE C. autoethanogenum* transconjugants over time, with change in the inoculation source of the donor strain (SP) highlighted.** The total number of *C. autoethanogenum* colony forming units (CFU) for a given conjugation is plotted on a log-scale and calculated by extrapolating the fraction of the mating slurry which was plated during the conjugation to the total mating slurry ($CFU_{TOTAL}=CFU_{MEASURED}/V_{PLATED}*V_{TOTAL}$). Whenever available, the transconjugants yield of the same construct (pMTL83151) was used; the yield of the first three conjugations was calculated from the construct with the highest yield (respectively vFS39, vFS23 and vFS23). Conjugations which failed to yield any colonies were not plotted.

Prior to September 2019, a new batch of sExpress was transformed with the shuttle vector before each new conjugation. After that date, because the original conjugation protocol does not specify whether the inoculum should come directly from a plate or from a cryostock, it was decided to prepare a cryostock of each donor strain. Inoculating the donor strains from their respective cryostocks instead of from new transformation plates would ensure that all the conjugations involving the same construct would be done with the exact same donor strain background, and it would give us the opportunity to extract and sequence the shuttle vector prior to the conjugation. This ensured that no mutation had occurred in the shuttle vector during the transformation of *E. coli* sExpress. It would also allow us to trace back the exact donor strain which would have been used in any given experiment, should anything go wrong.

The overlap with the drop in conjugation yield is not perfect, and it seemed to not be affecting the *ΔpyrE::ACE* strain anyway, but with so many parameters potentially affecting the conjugation yield, it was decided to test the possibility that inoculating the donor strains from cryostocks might have somehow harmed their conjugation potential. Additionally, because the conjugation yield seemed to decrease over time, I wondered if the duration of the cryopreservation (or the number of freeze-thaw cycles endured by the cryostocks) might also affect conjugation yields.

For this reason, in addition to simply comparing the conjugation yield of pMTL83151 in WT *C. autoethanogenum* when the donor strain was inoculated from cryostocks or from plates, two controls were added to account for some hidden variables.

- A fresh cryostock was prepared to test whether the duration (or, potentially, the number of freeze-thaw cycles) of the cryopreservation was impacting the conjugation yield. At the time of this experiment, the original sExpress pMTL83151 was fourteen-months old and had endured more than 10 freeze-thaw cycles.
- To check whether the act of plating was improving the conjugation potential of the strain by itself – instead of the cryopreservation process decreasing it – the original cryostock was spread on a plate 24h before mating in order to inoculate the strain from the old cryostock straight from a plate on the day of the mating.

#### *IV.3.2.1) Experimental setup*

All hypotheses and their associated controls are summarized in Table 11.

Because all hypotheses could be phrased as increase or decrease compared to their control, one-tailed Welch's t-tests were used during data analysis.

A fresh cryostock of *E. coli* sExpress harbouring the reference vector pMTL83151 was prepared 7 days before the mating. The old and fresh pMTL83151 cryostocks were each inoculated into 5 mL of liquid LB supplemented with chloramphenicol and kanamycin on the day before the mating. Also, on the day before mating, the old cryostock was spread on a petri-dish with solid LB-agar medium (also supplemented with chloramphenicol and kanamycin); a cryostock of *E. coli* sExpress was once again freshly transformed with pMTL83151 and also incubated overnight on a solid LB-agar medium supplemented with chloramphenicol and kanamycin.

**Table 11: Summary of the four hypotheses and of different treatments of the donor strain used to test them. SP= sExpress**

| Label | Hypothesis | Test | Donor strain |
|---|---|---|---|
| $H_0$ | Cryopreservation of pMTL83151 SP does not affect its conjugation yield. | | Old stock |
| $H_1$ | Cryopreservation of pMTL83151 SP deteriorates its conjugation yield over time or freeze-thaw cycles. | Compare the conjugation yield of the 14 months-old pMTL83151 SP cryostock with a fresh one. | Fresh stock |
| $H_2$ | Cryopreservation itself reduces the conjugation yield of pMTL83151 SP, no matter the duration or the number of freeze-thaw cycles. | Compare the fresh pMTL83151 SP cryostock with the fresh pMTL83151 SP strain inoculated straight from transformation plates. | Plate |
| $H_3$ | Inoculating from a plate improves the conjugation yield. | Compare the original pMTL83151 SP cryostock with itself, but inoculated straight from a plate on the mating day. | Plate from old stock |

On the day of the mating, the liquid overnight cultures of the old and fresh pMTL83151 cryostocks were subcultured to $OD_{600}$~0.05 in 5 mL of LB with chloramphenicol and kanamycin; a single colony of the old pMTL83151 cryostock and fresh pMTL83151 transformants were also each inoculated into 5 mL of LB with chloramphenicol and kanamycin. This was done in duplicate (two colonies or overnight liquid cultures from each treatment of the donor strain).

The rest of the protocol was executed as usual: harvest of 1 mL of each donor strain at $OD_{600}$~0.4, wash with 500 µL PBS; resuspend inside the gas cabinet with 200 µL WT *C. autoethanogenum* recipient strain at $OD_{600}$~0.7; mating on YTF plate without antibiotic for 20h; resuspension and normalisation of mating slurry in 600 µL; and spread of 200 µL on transconjugants selection plates.

### *IV.3.2.2) Results and discussion*

As illustrated in Figure 47, the number of transconjugants actually decreased when the original glycerol stock was streaked on a plate prior to inoculation, resulting in a $p>0.5$. As such, the hypothesis that plating is by itself improving the conjugation yield can be rejected. On the other hand, the increase in transconjugants colonies when using a fresh cryostock was highly significant ($p<0.005$), going from 83±15 CFU to 1077±34 CFU – more than a 10-fold increase. Lastly, although the average number of transconjugants almost doubled when

the donor strain was inoculated straight from the transformation plate instead of from a fresh glycerol stock (going from 1077±34 CFU to 1833±322 CFU, respectively), the increase was not statistically significant (p=0.09).

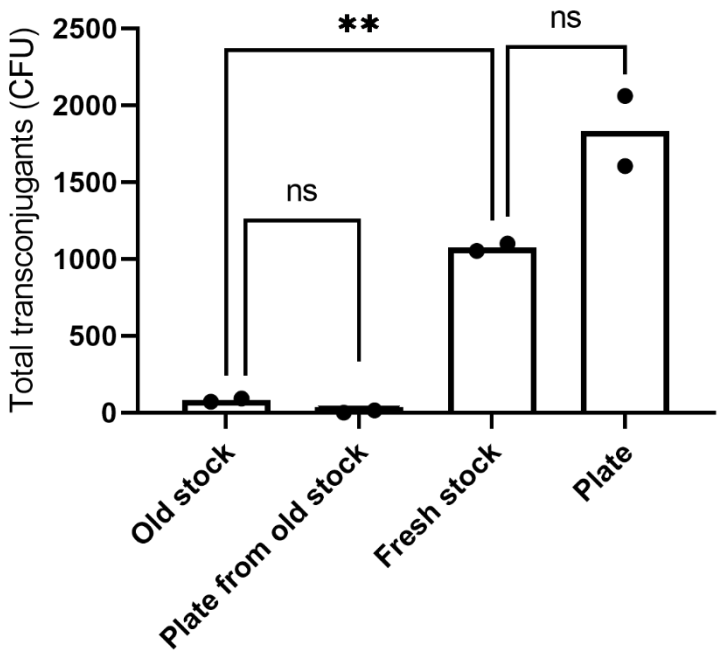



**Figure 47: Comparison of the conjugation yield between different origins of the inoculum of the sExpress donor strain.** The average number of transconjugants is represented by a bar, while the spread of each average is represented directly by each data point. The total number of *C. autoethanogenum* colony forming units (CFU) for a given conjugation is calculated by extrapolating the fraction of the mating slurry which was plated during the conjugation to the total mating slurry ($CFU_{TOTAL}=CFU_{MEASURED}/V_{PLATED}*V_{TOTAL}$). Old stock= original pMTL83151 cryostock; Plate from old stock = plate obtained from streaking the original pMTL83151 cryostock; Fresh stock = seven days-old pMTL83151 cryostock; Plate= fresh pMTL83151 transformation plate. ns= p>0.05; **= p<0.005; One-tailed Welsh's t-test, n=2.

Beyond statistical tests – which should not be taken at face value given the low number of replicates – it is striking to observe how such a minor difference in the treatment of the donor cells can impact the final number of transconjugants. For reasons which are not entirely clear, the duration of the cryopreservation (or the number of freeze-thaw cycles) definitely has an impact on the ability of the strain to conjugate effectively, even after being given the opportunity to recover over the course of an overnight culture and a subsequent subculture. Interestingly, the inoculum from the original cryostock actually reached the threshold of $OD_{600}$~0.4 faster (in 2 h 30 min) than the fresh cryostock (in 4 h 40 min) and all the other strains (between 4 h and 5 h), so the loss of conjugation potential is not due any loss of general fitness.

Three hypotheses could explain the decrease in conjugation yield during long period of cryopreservation.

#### IV.3.2.2)(1) Cryopreservation might generate or select subpopulations of donor cells with truncated R702 vectors

Parts of the R702 conjugation plasmid might have been lost in some of the sExpress cells, resulting in an overall loss of conjugation efficiency. Importantly, because the strains are still resistant to kanamycin, the R702 vector can only be damaged, not lost. These cells with truncated R702 might cope better during cryopreservation, or recover faster from it after inoculation, quickly outnumbering the sExpress cells which are still harbouring the whole R702 plasmid during overnight growth and the subsequent subculture. Such genetic instability might be favoured by the *dcm*- genotype of sExpress, although similar sensitivity to DNA damage has only been described for *dam*- and not *dcm*- strains, to the best of my knowledge [289–292].

Because cryostocks are generated from single colonies, and because the decrease in conjugation yield was observed among all constructs and not just pMTL83151, this hypothesis of sExpress subpopulations with truncated R702 vectors would need to fulfil at least one of three conditions:

- The original truncation event of R702 consistently occurs over the course of two overnight cultures: either when the cryostock is first generated or when it is revived prior to conjugation.
- The original truncation event of R702 is consistently induced during cryopreservation or during the freeze-thaw cycles.
- Each *E. coli* sExpress cell already harbours a mixed population of truncated and intact R702 vectors.

If this hypothesis is correct, the additional (but, technically, not significant) increase in conjugation yield when a fresh donor strain is inoculated straight from the transformation plate could be due to the loss of opportunity for R702 to be truncated – or loss of opportunity for the sExpress subpopulation with truncated R702 to take over – over the course of the generation of the cryostock, cryopreservation, and/or overnight recovery.

On the face of it, this hypothesis does not seem to be particularly likely, because it would imply that the R702 vector is inherently unstable – at least under the conditions of cryopreservation. Because cryopreservation is a routine microbiological practice, such an instability would have been spotted decades ago. Moreover, a putative R702 instability during cryopreservation should also affect the conjugation potential of the WT sExpress strain which is similarly kept in cryostocks and is regularly revived to generate new stocks of competent sExpress cells. No such decrease in conjugation yield is observed with the wider WT sExpress stock.

#### IV.3.2.2)(2) Cryopreservation might lock donor strains in an epigenetic state unfavourable to conjugation

A more phenotypical explanation to the decrease in conjugation yield could lie on the impact of cryopreservation and freeze-thaw cycles on the cell's membrane and regulatory state rather than the integrity of its R702 vector.

When cells divide, they inherit their membrane and cytoplasm from their progenitor, meaning that traumas like cryopreservation could potentially impact the state of cells for a few generations [293,294]. Conjugation is a costly, complex and tightly regulated process [287,288], so it might be affected by cryopreservation purely from an epigenetic perspective.

#### IV.3.2.2)(3) Chemically competent cells might be more proficient at conjugation

Contrary to the sExpress strains already carrying a shuttle vector, the conjugation efficiency of the WT sExpress stocks seems to be unaffected by cryopreservation. There are only two differences between these two cryostocks:

1) WT sExpress is not subjected to repeated freeze-thaw cycles; instead, a different aliquot of WT sExpress is used for each transformation.
2) WT sExpress is stored in a chemical transformation buffer of pre-cooled 100 mM $CaCl_2$ solution with 20% glycerol. On the other hand, the *E. coli* strains already carrying a shuttle vector are stored in a proprietary Microbank® buffer.

Consequently, the Microbank® buffer might just be worse than the chemotransformation buffer at protecting the cells from the deleterious impact of cryopreservation. Conversely, the chemocompetency of cells or their associated buffer could actually be improving conjugation efficiency for a few generation, but these improvements would be progressively lost during cryopreservation in Microbank® buffer and the subsequent subculture steps.

### *IV.3.2.3) Overnight culture is not necessary when inoculating straight from the sExpress transformation plate*

A small modification was brought to the conjugation protocol when inoculating the donor strain straight from the transformation plate: the overnight liquid culture step was removed. The donor strain liquid culture is instead inoculated straight from the transformation plate on the day of the mating. It then reaches $OD_{600}$~0.4 within 4h-8h of inoculation. To make sure that this modification does not affect the effectiveness of the conjugation protocol, impact of removing this step was assessed by measuring the final number of transconjugants. This was done alongside the experiment illustrated in Figure 47, but tested with a two-tailed Welch's t-test.

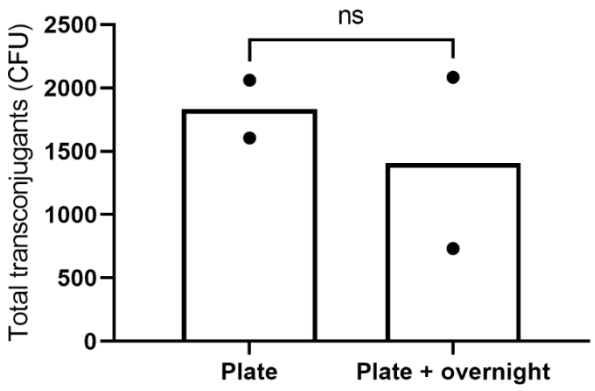



**Figure 48: Comparison of the conjugation yield with or without removing the donor strain overnight culture step when inoculating straight from the sExpress transformation plate.** The total number of *C. autoethanogenum* colony forming units (CFU) for a given conjugation is calculated by extrapolating the fraction of the mating slurry which was plated during the conjugation to the total mating slurry ($CFU_{TOTAL}$ = $CFU_{MEASURED}/V_{PLATED}*V_{TOTAL}$). Plate = pMTL83151 transformation plate without subsequent overnight; Plate + overnight = pMTL83151 transformation plate with subsequent overnight. ns= p>0.05. Two-tailed Welsh's t-test, n=2.

The removal of the overnight culture step made no significant difference (Figure 48). If anything, it increased the yield and reduced the variability. As such, all future conjugations will be done by inoculating the donor strain straight from the sExpress transformation plate, without additional overnight incubation.

### IV.3.3. Paths to further improve conjugation yield

#### *IV.3.3.1) Cryopreservation of C. autoethanogenum might affect the conjugation yield*

Given the dramatic impact of the long-term cryopreservation of the donor strain on the conjugation yield (Chapter IV.3.2), the absence of mutations within the WT genome (Chapter IV.3.1.2), and the comparatively excellent yield of conjugation in a different *C. autoethanogenum* strain without obvious genetic justification (Chapter IV.3.1.3), I started suspecting that long-term cryopreservation of the recipient strain might also have a deleterious impact on the conjugation yield without having to impact its genetics. It was hypothesized that serial sub-culturing the *C. autoethanogenum* WT strain might allow it to leave an epigenetic state induced by cryopreservation which would be unfavourable to conjugation. The superior performance of *ΔpyrE::ACE C. autoethanogenum* strain when conjugated with a donor strain from a cryostock might thus be explained by a shorter stay in cryopreservation than WT *C. autoethanogenum.*

The last aliquot of my personal *Clostridium* WT autoethanogenum stock was inoculated into 4 mL of YTF medium without antibiotics and subcultured three times over a period of eight days before making fresh cryostocks when it reached an $OD_{600}$~0.4 after the final subculture step.

Unfortunately, no formal experiment was done to compare this refreshed stock with the original because the original, non-refreshed cryostock of *C. autoethanogenum* was depleted. However, conjugations done with the refreshed stock did seem to yield about 50% more transconjugants than the best yields that were obtained before the *C. autoethanogenum* cryostock had been refreshed in this way, providing anecdotal evidence that serial subctulture of a cryostock might help increase the conjugation yield of a strain preserved in cryostocks (cf. Figure 49). The difference is still not statistically significant (one-tailed Welch's t-test, p=0.053, n=3), however, and the effect might also be at least partially due to the fact that less excess donor cells were used than before by harvesting the donor strain at $OD_{600}$~0.2 instead of $OD_{600}$~0.4.

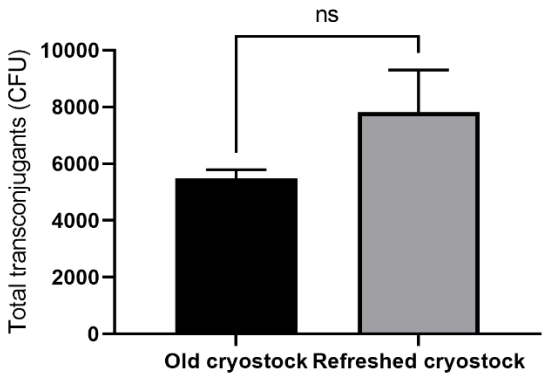



**Figure 49: Putative effect of refreshing the recipient strain by three serial subcultures.** The average of the top three conjugations yield obtained with pMTL83151 before the cryostock was refreshed by three serial subculturing (old cryostock) is compared to the next three conjugation yields obtained after refreshing the cryostock. The total number of *C. autoethanogenum* colony forming units (CFU) for a given conjugation is calculated by extrapolating the fraction of the mating slurry which was plated during the conjugation to the total mating slurry ($CFU_{TOTAL}=CFU_{MEASURED}/V_{PLATED}*V_{TOTAL}$). ns= non-significant difference (p=0.053, one-tailed Welch's t-test, n=3).

This hypothesis should be tested more rigorously by starting over from the original *C. autoethanogenum* stock of the culture collection: it could be refreshed in the same manner with three consecutive subcultures, then the conjugation yield of its original and refreshed cryostocks could be formally compared. The conjugation yield of the original and refreshed cryostocks could also be compared when conjugated with an old and refreshed cryostock of pMTL83151 donor strains, to see if cryopreservation and epigenetics could explain on their own why *ΔpyrE::ACE C. autoethanogenum* performed so much better than WT when conjugated with an old cryostock of pMTL83151 sExpress donor cells. Whole genome sequencing of the methylation patterns of each strain [295–297] might then reveal some of the epigenetic markers responsible for their different performances during conjugation.

#### *IV.3.3.2) Serial conjugations might select subpopulations more receptive to conjugations*

Interestingly, conjugation of pMTL83151 in a *C. autoethanogenum* strain recently engineered with a vector derived from pMTL83151 yielded an even greater number of transconjugants (upwards of 20,000 CFU, or more than double the previous record) (Figure 50). Dr Christopher Humphreys hypothesised that this might also find an explanation in epigenetics, as this strain necessarily descends from the small subpopulation of recipient cells actually receptive to conjugation which enabled the initial conjugation event in a first place. However, this hypothesis requires further testing as it was the result of only one single experiment and might not be reproducible.

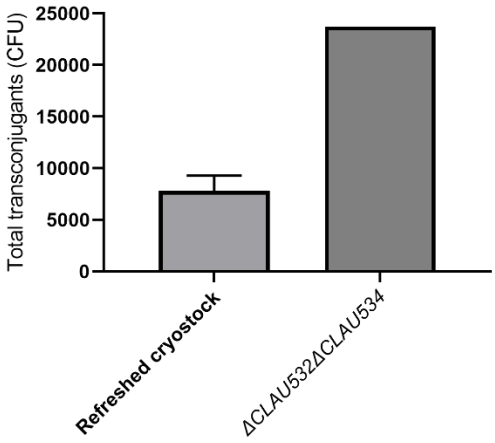



**Figure 50: Putative effect of conjugating a *C. autoethanogenum* strain which has previously received (then lost) a similar shuttle vector through conjugation.** The yield of the three conjugation yields obtained after refreshing the cryostock are compared to the yield of the one conjugation made in *ΔCLAU532ΔCLAU534 C.autoethanogenum*, a strain previously engineered with a pMTL83151-derived vector. The total number of *C. autoethanogenum* colony forming units (CFU) for a given conjugation is calculated by extrapolating the fraction of the mating slurry which was plated during the conjugation to the total mating slurry ($CFU_{TOTAL}=CFU_{MEASURED}/V_{PLATED}*V_{TOTAL}$). ns=non-significant difference (p=0.053, one-tailed Welch's t-test, n=3).

The great and consistent conjugation yield measured in *ΔpyrE::ACE* even when the donor strain was inoculated from an old cryostock might also be partially explained by this hypothesis, as it was also generated by conjugation with a pMTL83151-derived vector. On the other hand, *ΔpyrE::BMa C. autoethanogenum* has also already acquired and subsequently lost a pMTL83151-derived vector, but it showed a poor conjugation efficiency in the same conditions.

## IV.4. Summary and perspectives

This chapter was born out of frustration and curiosity. The conjugation protocol regularly failed to yield any transconjugants. When it worked, the yield of the conjugation protocol was unpredictable, and worse, seemed to be steadily decreasing. With a combination of trial-and-error and rigorous experimentation summarized in Figure 51, various parameters involved which could improve and stabilize the performance of the conjugation protocol were explored and defined. It resulted in an improved conjugation protocol which reliably yields 7821±1486 CFU of *C. autoethanogenum* carrying the pMTL83151 vector when using *E. coli* sExpress as the conjugal donor. In order of importance, this progress was a consequence of: (1) inoculating the donor strain straight from its transformation plate on the day of the mating instead of inoculating it from an old cryostock; (2) by reducing the number of donor cells at the start of the mating (from 1 mL at $OD_{600}$~0.4 to 1 mL at $OD_{600}$~0.2), and; (3) by taking extraordinary measures to prevent desiccation and oxygen contamination.

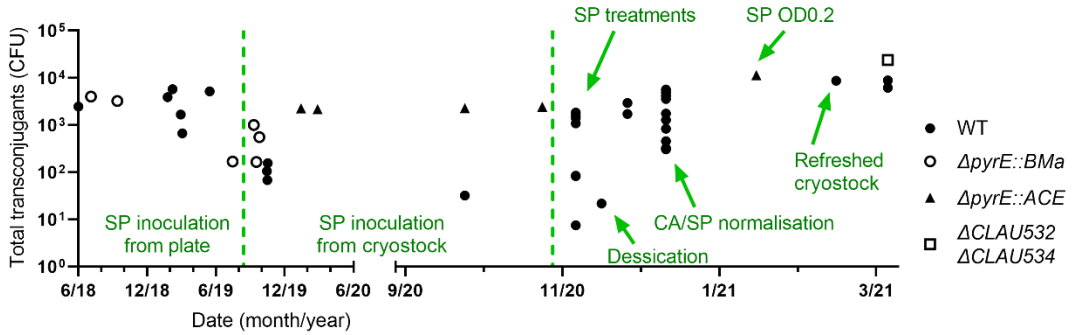



**Figure 51: Yield of WT and *ΔpyrE::BMa C. autoethanogenum* transconjugants over time.** The total number of *C. autoethanogenum* colony forming units (CFU) for a given conjugation is plotted on a log-scale and calculated by extrapolating the fraction of the mating slurry which was plated during the conjugation to the total mating slurry ($CFU_{TOTAL}=CFU_{MEASURED}/V_{PLATED}*V_{TOTAL}$). Whenever available, the transconjugants yield of the same construct (pMTL83151) was used; the yield of the first three conjugations was calculated from the construct with the highest yield (respectively vFS39, vFS23 and vFS23). Conjugations which failed to yield any colonies were not plotted. The change in treatment of the sExpress donor strain, as well as various experiments mentioned in this chapter, are annotated in green. SP treatment= Chapter IV.3.2; CA/SP normalisation = Chapter IV.2.3; Refreshed cryostock= Chapter IV.3.3.1).

In the process, it was found that small variations in the starting number of donors or recipients were unlikely to make a difference to the final number of transconjugants, and consequently that conjugation yield was most likely limited by nutrient availability. Importantly, it was established that normalizing conjugation yield by the starting number of donors or recipients could not be justified mathematically. Instead, it is more appropriate to normalise the conjugation yield of a given construct by the conjugation yield of an internal control, such as the empty vector pMTL83151.

Although genome sequence analysis of the *C. autoethanogenum* recipient strains in some cases revealed some unexpected SNPs, the differences in conjugation efficiency observed between these strains is more likely due to epigenetic changes rather than mutations. In support to this hypothesis, it was shown that a long period cryopreservation, or many freeze-thaw cycles, of the donor strain negatively impacts its ability to conjugate effectively. Anecdotal evidence showed that this might also be true for the recipient strain.

# Chapter V: Characterisation of a base editor in *C. autoethanogenum*

## V.1. Introduction

In an effort to improve the Cas9-genome editing toolbox and facilitate multiplex mutagenesis, I sought to apply a new, state-of-the-art genome editing technique to the *Clostridium* genus, and to *C. autoethanogenum* in particular: base editing. Base editors allow to mutate a single DNA base among the whole genome, bypassing HR entirely (cf. Chapter I.3.5). As discussed previously, *C. autoethanogenum* has a rather poor HDR efficiency relative to other organisms; consequently, base editors might also facilitate multiplex mutagenesis.

On the other hand, base editing is inherently less versatile than an HDR-mediated genome editing tool. Indeed, instead of changing, adding or removing hundreds of thousands of bases, base editing is limited to changing at most a single codon. However, this is enough to conduct a complementation study by changing this codon to a STOP codon. Such a nonsense mutation interrupts the translation of the target CDS, which can effectively knock-out the gene. The underlying hypothesis of this chapter is that the convenience and the efficiency of gene knock-out through base editing should outweigh its lack of versatility.

### V.1.1. Base editors exploit the Mismatch Repair pathway

In 2016, Nishida et al. and Komor et al. each designed a fusion protein described as a base editor to carry out targeted point mutations in eukaryotic genomes [298,299]. It is based on CRISPR-Cas9, and, as such, it has the same specificity, versatility, and potential for multiplexing. However, contrarily to the standard Cas9, base editors insert point mutations without relying on the HDR or NHEJ pathways. Indeed, base editing mutates one single strand of the DNA helix without introducing a DSB, and exploits the MMR pathway (cf. Chapter I.3.2) to mutate the second strand.

The first part of the protein fusion, Cas9-nickase or nCas9, targets a specific locus and nicks the DNA strand complementary to the strand that will be mutated by the second part of the protein fusion. This fulfils three functions:

- The nCas9-gRNA duplex targets a particular locus in the genome with high specificity.
- The nCas9-gRNA duplex nicks the WT allele, providing selective pressure against it.

- The nCas9-gRNA duplex nicks the non-edited strand, which primes the MMR pathway to repair it by using the sequence of the edited strand as template if a mismatch occurs in between both strands (cf. Chapter I.3.2).

The second part of the protein fusion is a cytidine deaminase (CDA), which effectively turns a cytosine on the edited strand into a uracil (C→U). Uracil has the same bonds configuration as a thymine (T), which can lead to a mutation of the non-edited strand upon DNA replication or DNA repair through the MMR pathway (G→A) (cf. Chapter I.3.2). In that case, the uracil would then be irreversibly replaced by a thymine (U→T) through the BER pathway (cf. Chapter I.3.3.1) using the now mutated non-edited strand as a template. In short, the base editor is simply substituting a specific C:G base-pair with a T:A base-pair.

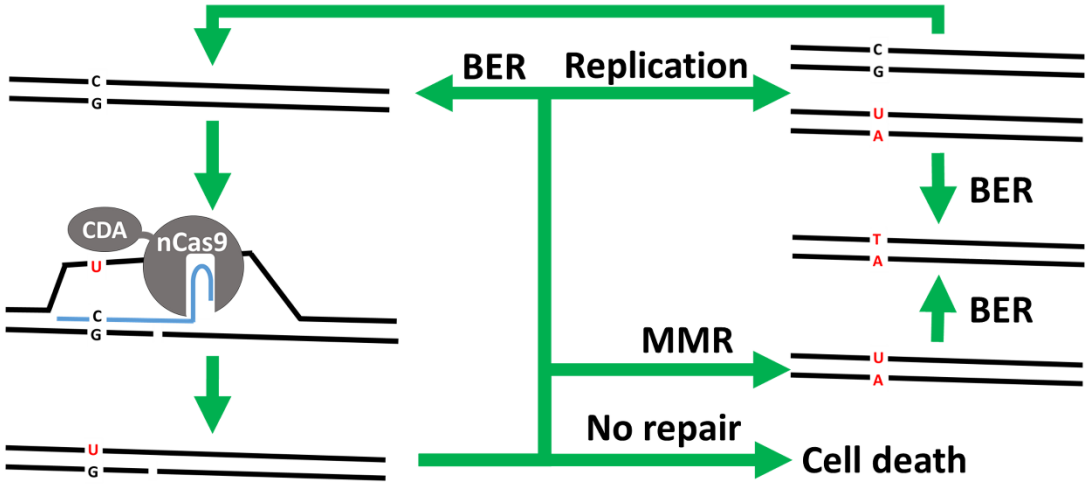



**Figure 52: Summary of the base editing mutagenesis strategy.** In a clockwise order starting from the top left with the **WT chromosomic DNA**: the **base editor-sgRNA duplex** unwinds the DNA double helix around the protospacer, creating a R-loop, nicking the non-edited strand and exposing a cytosine (C) to the deaminase activity of CDA on the edited strand – which changes it into a uracil (U); **without repair** of the nick on the non-edited strand, the cell is unable to replicate its DNA and dies; if the DNA helix undergoes **MMR**, the nicked non-edited strand will be repaired and mutated to match the edited strand, replacing the guanine (G) of the non-edited strand with an adenine (A); in the last step, the uracil of the edited strand will be removed and replaced by a thymine (T) through the **BER pathway**; alternatively, **DNA replication** of the edited strand can also produce a new chromosome which would include the desired G→A mutation, along with a WT chromosome resulting from the replication of the non-edited strand; at last, immediate ligation strand (not shown) of the nicked non-edited and subsequent repair of the edited strand through the **BER pathway** would result in a stable WT chromosome which would be exposed to another cycle of mutagenesis for as long as the base editor and its sgRNA cassette are being expressed.

This mutagenesis strategy is thus pitting DNA replication and the MMR pathway against the BER pathway (Figure 52). As long as the BER pathway removes the deaminated cytosine before either the MMR pathway or DNA replication has had a chance to mutate the non-edited strand, the base editor should keep binding the WT allele, deaminating the target

cytosine, and nicking the non-edited strand to provide selective pressure and prime the MMR pathway. As part of a base editor, CDA only targets cytosines within a roughly 4bp window inside the protospacer. Consequently, if the mutation is successful, the resulting mismatch in between the mutant allele and the gRNA spacer should prevent binding and protect the mutant cell from the nickase activity of Cas9.

### V.1.2. Obstacles to base-editing mutagenesis

Some DNA repair pathways can interfere with the base editing mutagenesis strategy (cf. Figure 53). If the edited strand is cut by a spontaneous break or by AP endonuclease (cf. page 18) before the nick on the non-edited strand could be mended by DNA ligase, this would generate a DSB which could trigger the NHEJ pathway, leading to indel mutations [300]. In *clostridia*, however, this course of event would be lethal since they do not possess a NHEJ pathway. Error-prone DNA polymerases could also mutate the protospacer or the base editor and its gRNA, respectively producing random mutations of the protospacer or inactivating the base editor altogether [112]. At last, translesion DNA synthesis (cf. Chapter I.3.1) could replace the bases complementary to the AP site left by UDG (cf. Chapter I.3.2) with a random nucleotide, potentially leading to a C→A or C→G mutation instead [112,116–118,301]. In mammals, alternative DNA repair pathways such as these have been blamed for the occasional occurrence of indels and other unintended mutations instead of the specific C→T transition.

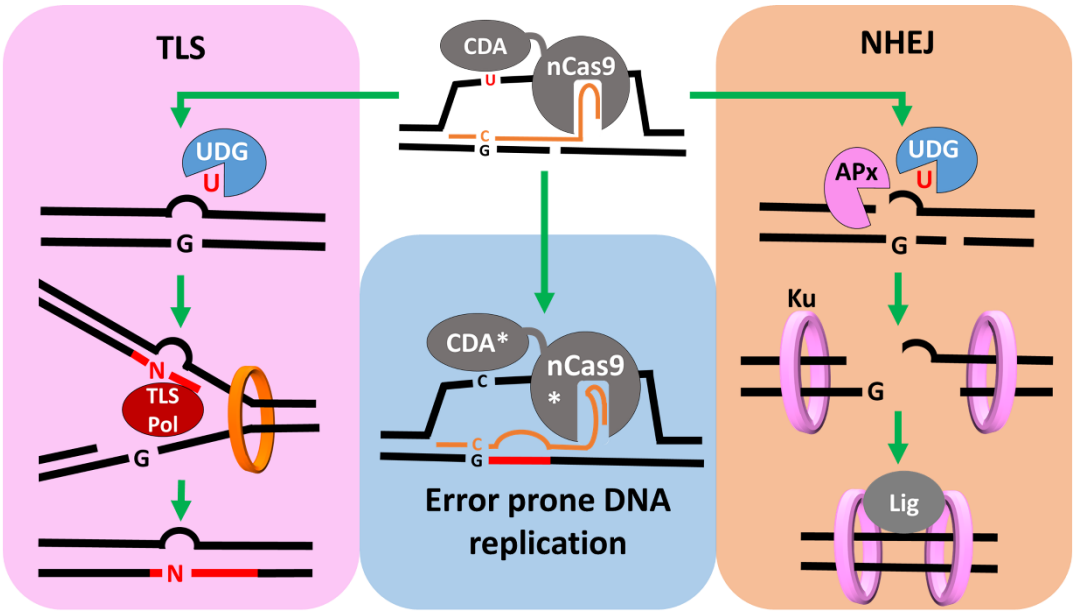



**Figure 53: Mutagenic processes competing with base editing.** During **translesion DNA synthesis (TLS)**, the nick in the unedited strand is repaired immediately but DNA polymerase stalls on the AP site left by UDG; subsequently DNA replication is rescued by a TLS polymerase (TLS pol) which inserts random nucleotides over the AP site and the adjacent bases. In the occurrence of **error prone DNA replication**, random point mutations in the genomic locus targeted by the base editor (red), in the regulatory regions of the base editor (not schematized), in the CDS of the base editor (Cas9* or CDA*) or in its associated sgRNA (misshapen sgRNA) can all completely or partially inhibit the activity of the base editor. At last, the **NHEJ pathway** is activated if AP endonuclease (APx) nicks the edited strand before the non-edited strand could be mended by DNA ligase, resulting in a DSB which recruits Ku proteins and is eventually ligated – often at the cost of indels at the locus of the DSB.

To inhibit some of these competing pathways, a third enzyme is often added to the protein fusion: a uracil glycosylase inhibitor (UGI) from bacteriophage PBS [298–300] (Figure 54). The PBS phage uses exclusively uracil instead of thymine in its DNA, and as such it needs to deactivate UDG and the BER pathway of its host cell in order to replicate itself [302,303]. Accordingly, fusing UGI to a base editor locally inhibits the BER pathway for as long as the base editor binds to the target locus. In mammals, it has been shown to effectively reduce the occurrence of indels and increase the C→T mutagenesis efficiency [298,299,303]. Similar gains in efficiency have been observed in *E. coli*; however, it also proved toxic for *E. coli* and it significantly increased the number of off-target mutations [304]. Toxicity was mitigated by using the last three aa of an LVA protein degradation tag to reduce the stability of the base editor and by using a dCas9 instead of an nCas9.

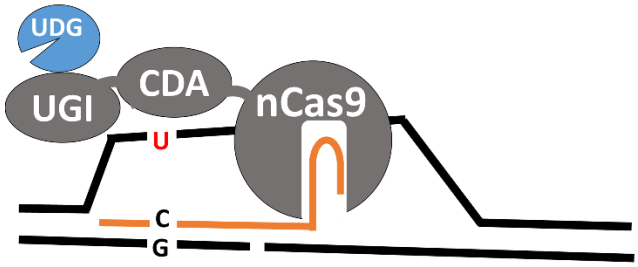



**Figure 54: Uracil Glycosylase inhibitor (UGI) is fused to a base editor (composed of nCas9 CDA) to inhibit uracil DNA glycosylase (UDG) and maximize the chances of producing the intended mutation.**

### V.1.3. Further development of base editors

Komor et al. named their fusion protein BE, while Nishida et al. used the nomenclature Target-AID. Both systems are similar, but differ in the specific CDA that is fused to Cas9, the position of the UGI (C- or N-terminal) and the linker used between each protein.

Several base editors have since been developed on the same principle. A number of increasingly effective iterations of base editors [259,305–307] were eventually complemented with an adenine base exchanger (ABE), in which CDA was replaced by an engineered tRNA adenosine deaminase (TadA) to enable A:T to G:C transitions [308]. Over the course of this project, different groups have also designed base editors with expanded genetic spaces to target additional loci by exploiting Cas9 proteins which have been engineered to recognize PAMs other than NGG [129,200,260,309–317].

Both Target-AID and BE were explored through a collaboration with Li and colleagues [318]. Li tested BE in *C. beijerinckii*, while I characterized Target-AID in *C.autoethanogenum*. The choice turned out to be relevant, because Nishida et al. later showed their system to be working in *E. coli* too, although they had to use a nuclease-deficient Cas9 (dCas9) instead of a nCas9. As such, this chapter aims at replicating the Target-AID fusion protein (spCas9-pmcDA1) to insert premature stop codons in *C. autoethanogenum*.

### V.1.4. Inducible expression systems for Cas9-mediated and base editing mutagenesis

As introduced in Chapter I.5.1, Cas9 can be toxic in eukaryotes, and expressing it under the control of an inducible promoter is recommended [183,184]. This is also true of Target-AID, which was reported to be toxic in *E. coli* even under the control of a temperature-inducible promoter [304].

Interestingly, in Chapter III:, Cas9 was already used very successfully without any repression system. To be more specific, a truncated Cas9 (trCas9) had been used. It resulted from a frameshift mutation at the beginning of the *cas9* gene [169]. Since trCas9 is demonstrably active in spite of the frameshift mutation, it is suspected that an internal methionine (M90) serves as an alternative START codon. The resulting trCas9 protein would thus be shorter than Cas9 and would have lost part of the RuvC nuclease domain (cf. Figure 55). This would effectively mean that the strong $P_{thl}$ constitutive promoter controlling the expression of trCas9 actually results in the very low expression of a Cas9 nickase. Because the nCas9(D10A) used in Target-AID has also a disrupted RuvC nuclease domain, and because constitutive expression of trCas9 seems to be well tolerated by *E. coli* and *C. autoethanogenum*, I wondered if using trCas9 instead of nCas9(D10A) would also enable constitutive expression of Target-AID.

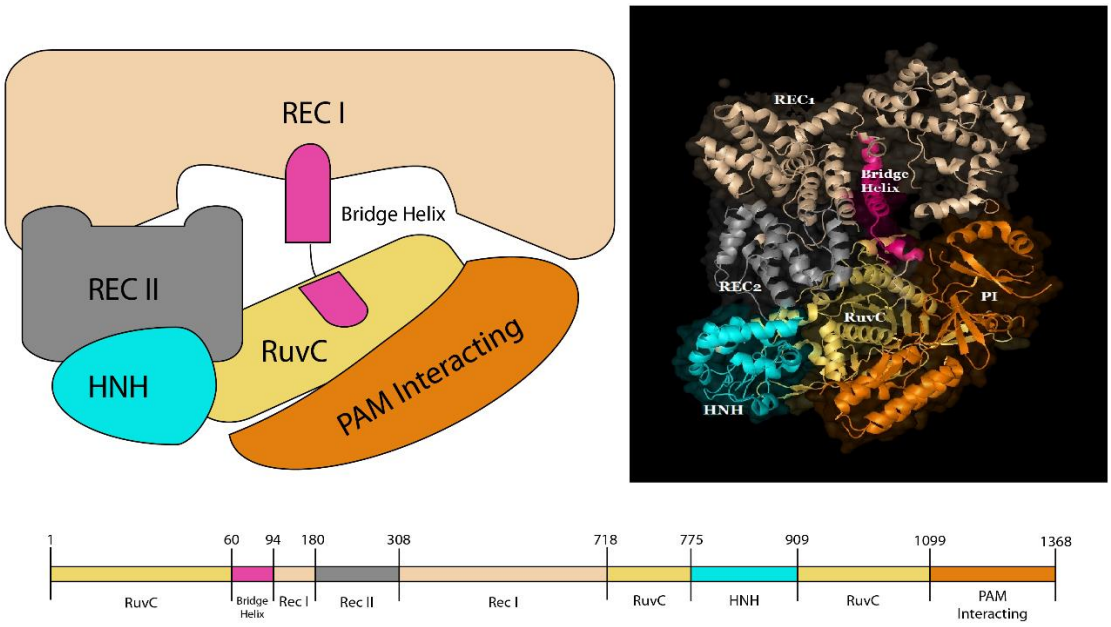



**Figure 55: Schematic and crystal structure of the Cas9 protein.** Cas9 can be divided into six domains: Rec I, Rec II, Bridge Helix, RuvC, HNH, and PAM Interacting. Domains are shown in schematic, crystal, and linear map form. Numbers on the linear map represent the position of aa. Reproduced from Garrity & Cavanagh, 2014 [319].

Nonetheless, a tight repression system would still be a safer strategy to express Target-AID in *C. autoethanogenum* for the first time. An inducible system also has the advantage of discriminating in between the conjugation efficiency of the vector and the toxicity of the protein it expresses. It would give us more flexibility to find the best compromise in between toxicity of the construct and mutagenesis efficiency. After a brief literature review of various inducible systems, two of them were tested in parallel with the constitutive expression of trCas9: lacI, a transcriptional repressor, and RiboCas, a translational repressor.

It is difficult to compare the activity of different promoters from the literature, as they often use different reporter genes and are seldom assayed in the same conditions – or even in the same organisms. Consequently, a qualitative rather than quantitative list of requirements was used to find our ideal Target-AID repression system. It would need to be very tightly repressed, to use a non-toxic, non-metabolisable inducer such as IPTG, and would have already been used in *E. coli* as well as in *C. autoethanogenum* or a closely related species.

Gyulev et al. recently reviewed ten inducible promoters used in *Clostridium acetobutylicum*. Five of them are based on the LacI repressor, and were inducible by IPTG (P*thl-2xlac*, P*thlOid*, P*fdxoid*, P*fac-lac*, P*facOid*); one relies on the expression of TcdR repressor (P*tcdb*); one relies on BgalR and is lactose-inducible (P*bgaL)*; one uses the XylR repressor and is inducible by D-xylose (P*xylA*), another metabolisable sugar; and two more relied on TetR and the tet operator (P*pcm*, P*xyl/tetO*) [85,244,320–325]. Of these promoters, the Pfacoid promoter was the most interesting for having already been used in our lab in the past. Being inducible by IPTG instead of tetracycline (an antibiotic) or a metabolisable sugar such as xylose or lactose was also a great advantage. However, this repression system had only been used to reduce the toxicity of its construct in *E. coli* (it was constitutively expressed in *C. acetobutylicum*) and had never been used in association with Cas9.

Another strong contender for inducible promoter was Pipl12, a tetracycline-inducible promoter derived from Ptet3no which had been successfully used for HDR-mediated Cas9 mutagenesis in *C. autoethanogenum* [84]. The new Pipl12 promoter had been effective at deleting 1kb from the gene CLAU392 (2,3 bdh) with >50% efficiency. It has a higher expression level than Pte3no both in its induced and repressed states, although it has a larger dynamic range. This promoter is interesting because it has been used in a context almost

identical to my purpose. However, it is induced by an antibiotic and its high basal expression might make it incompatible with Target-AID, even if it seemed to work with Cas9.

At last, a brand new repression system had just been developed in our lab by Cañadas et al. It uses riboswitches to repress translation of a target mRNA [326]. Upon induction with theophylline, the mRNA structure of the riboswitch unravels and allows the ribosome to initiate translation. The repression of the riboswitch was shown to be exceptionally tight, as could be expected of a translational repression system [326–331]. The system was specifically optimized to regulate Cas9 expression under the name of RiboCas, and it was tested and characterized in several *Clostridium* species (*C. pasteurianum*, *C. difficile 630* and the group I *C. botulinum* strain) [245]. Additionally, it was the only repression system which did not require the concomitant expression of a repressor protein. This limits the size of the vector and reduces the metabolic load for the host cell, both of which should increase the stability of the vector.

Based on this literature review, the promoter $P_{fdxE}$ from the RiboCas system was deemed the most promising repression system. A repression system based on LacI and Pfacoid was also assembled, should an alternative be necessary.

### V.1.5. Experimental design

This chapter focuses on the characterisation of Target-AID itself; the multiplexing of Target-AID will be addressed in the Chapter VI. Consequently, all the genome editing attempts were made on a single locus: *pyrE*. As previously discussed, *ΔpyrE* mutants are insensitive to FOA. Successful mutagenesis events can thus be easily quantified by counting the number of transconjugant colonies which survive exposure to FOA.

#### *V.1.5.1) Mutagenesis parameters*

Target-AID and its control constructs were compared on basis of three important mutagenesis parameters. The first parameter is the toxicity of the construct. Less toxic constructs will yield more transconjugants upon induction, which is a prerequisite for mutagenesis. Less toxic constructs are also less likely to trigger the SOS response, and should thus have a lower rate of off-target mutations. Toxicity of each construct will be estimated through the percentage of cells which survive induction of the nuclease. A larger proportion of surviving cells implies a lower toxicity of the construct.

However, if a construct is not toxic enough, it might not counter-select the WT cells, resulting in many colonies which survive induction but few of them carrying the desired mutation. Consequently, mutagenesis efficiency, or the percentage of induced colonies with the expected *pyrE(Q130X)* genotype, is also a key parameter.

Lastly, the raw number of transconjugants which survived induction was also considered. It is a trivial but important variable because if too few colonies survive induction, even a high conjugation efficiency might not result in any mutant. Nevertheless, this value has to be handled cautiously because, as was discussed at length in in Chapter IV:, the raw colony count is not a very precise measurement.

#### *V.1.5.2) Chapter outline*

After designing various constructs to help us characterize Target-AID mutagenesis, a bioinformatics tool was written to find gRNA targets which were likely to generate a premature STOP codons in the *pyrE* gene of *C. autoethanogenum* over the course of Target-AID mutagenesis. The sequence of Target-AID was then altered to facilitate its conjugation and its expression in *C. autoethanogenum*. The codon usage of the CDA-half of the protein fusion was optimized to fit the codon usage *of C. autoethanogenum*. Additionally, sequences recognized by the type II RM system of *C. autoethanogenum* were pre-emptively removed from Target-AID and the pMTL backbone.

Once all these technical considerations were dealt with, Target-AID mutagenesis was characterized in *C. autoethanogenum* across two separate experiments (a preliminary and a final characterisation). In between these two major experiments, alternative induction systems were tested and some limitations of the RiboCas repression system were explored.

## V.2. Design of Target-AID constructs

The RiboCas translational repression system was selected to regulate the expression of Target-AID [vFS36_TA_PfdxE_pyrE, cf. Figure 56(a)]. However, a LacI-Pfacoid repression system was also assembled as a backup and was tested alongside it (vFS35_TA_PlacIq_LacI_*pyrE*, cf. Figure 56(b)). Assembly of Target-AID under the control of a constitutive Pfdx promoter was attempted, but it failed despite my best efforts. This apparent toxicity of constitutive expression of Target-AID in *E. coli* highlighted the necessity of a tight repression system in *E. coli*, in agreement with literature [304].

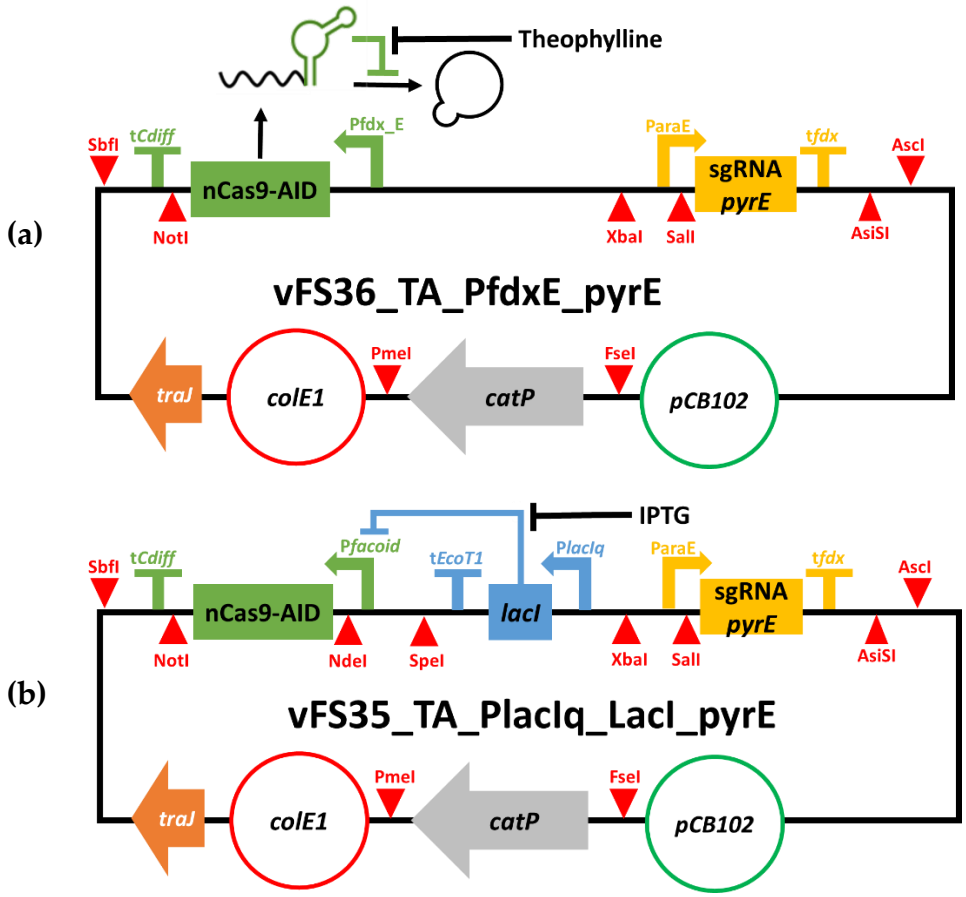



**Figure 56: Plasmid maps of Target-AID vectors under the control of two different repression systems. (a)** In vFS36_TA_PfdxE_pyrE, the expression of Target-AID (nCas9-AID) is regulated at the translational level by a riboswitch which inhibits ribosome binding of the mRNA. When present, a theophylline molecule binds the aptamer which changes the conformation of the riboswitch and enables translation initiation. **(b)** In vFS35_TA_PlacIq_LacI_*pyrE*, the expression of Target-AID is regulated at the transcriptional level by using the strong *E. coli* $P_{lacI}^{Q}$ promoter to express the LacI repressor, and placing Target-AID under the control of Pfacoid. LacI binds the *lacO* operator in Pfacoid, consequently inhibiting transcription of nCas9-AID; IPTG binds LacI, preventing it from binding *lacO* and consequently increasing the transcription rate of nCas9-AID.

Three other control constructs were assembled under the control of the RiboCas repression system to collect more information about the properties of Target-AID:

- vFS37_NosgRNA, a construct identical to vFS36_TA_PfdxE_pyrE but without sgRNA cassette, was assembled to estimate the off-target mutagenesis rate of Target-AID. If the mutagenesis efficiency of vFS37_NosgRNA is close to the one of vFS36_TA_$P_{fdxE}$, then Target-AID is not specific enough in *C. autoethanogenum*.
- vFS39_dTA, a construct identical to vFS36_TA_PfdxE_pyrE but with a catalytically inactive Cas9 (dCas9) instead of a nCas9(D10A), was assembled to verify the claim that a nuclease domain is not necessary for an effective Target-AID mutagenesis in prokaryotes [304]. Without counter-selection, a high number of colonies would be expected to survive induction, but few of these would be mutants and thus survive exposure to FOA – resulting in a low mutagenesis efficiency.
- At last, vFS40_nCas9_PfdxE_HDR, a construct identical to vFS36_TA_PfdxE_pyrE but without a PmCDA1 domain and with a *ΔpyrE* gene editing template, was used to compare the effectiveness of Target-AID with a conventional HDR-mediated Cas9 mutagenesis.

### V.2.1. Conception of a Target-AID gRNA design tool

Through its gRNA component, Target-AID is limited to targeting a 20 bp sequence of genomic DNA upstream of an NGG PAM. Within these 20 bp, only a 3-5 bp window annealing near the 5'-end of the spacer sequence can be mutated. Inside this small editing window, there must be at least one cytosine base in such a position in the codon that changing it to a thymine would produce a STOP codon. At last, this STOP codon has to be relatively early within the coding sequence to have a reasonable chance of inhibiting the function of protein after translation of the CDS. As a result of this very specific set of requirements, most potential gRNAs are not compatible with Target-AID mutagenesis.

In order to find which of the 45 potential gRNAs targets were actually suitable to knock out *pyrE* with Target-AID, a simple Excel table was set up which could receive the gRNA design table from Benchling as an input and flag the subset of gRNAs which could produce a STOP codon as an output. An example of the input and outputs of the Target-AID gRNA design table are given in Table 12 and Table 13.

Two potential gRNA targets were identified among the 45 initial protospacers, each near the end of the *pyrE* locus (in position 388 and 535, out of a 573 bp gene). In order to maximize the probability of knocking-out *pyrE* with the premature STOP codon, the most upstream protospacer (AATTCAGGAATTAGGTGGAG) was selected as gRNA target. It was predicted to the lead to the mutation Q130X.

**Table 12: Raw output from Benchling's gRNA design tool.** Five entries out of 45 are given as an example. The **Position** is the position from the start of the sequence being analysed, which is not necessarily the start of the CDS (as in this example, where the input sequence is the non-coding strand); **Strand** stands for the direction in which the gRNA target is found in the input sequence: 1 is in the same direction as the input, -1 is in the reverse-complementary direction; **Sequence** is the sequence of the gRNA target (or protospacer); **PAM** is the sequence of the PAM, directly downstream of the protospacer; **Specificity** and **efficiency scores** respectively refer to off-target and on-target efficiency scores [cf.Chapters I.5.2.1) and I.5.2.2)]; closer to 100 is better.

| Position | Strand | Sequence | PAM | Specificity Score | Efficiency Score |
|---|---|---|---|---|---|
| **448** | 1 | ATTACATCTTTTGCCCTGTC | AGG | 49.99 | 3.83 |
| **386** | -1 | GTTGACTATGATAAAATAGT | TGG | 49.40 | 23.66 |
| **373** | 1 | ACTAATATCCCCCCCATTGC | AGG | 50.00 | 12.24 |
| **174** | -1 | AATTCAGGAATTAGGTGGAG | AGG | 49.91 | 27.08 |
| **26** | -1 | GGACAAGAATATGTAAAGCC | TGG | 50.00 | 19.22 |

**Table 13: Output of the Target-AID gRNA design Excel table.** The five entries of Table 12 are given as an example. The first three columns ("$C_1$","$C_2$","$C_3$") show the position of the first three cytosines in the gRNA. The rest of the results are only given for the first cytosine ($C_1$). In "$C_1$[16-20]", the position of the cytosine relative to the start of the gene is calculated, unless the C falls outside the predefined editing window of -20 to -16 bp from the PAM. "Rank in Codon" calculates the position of the cytosine within its codon (it is either the first, second or third base of its codon). "Triplet" spells out the sequence of the triplet of the cytosine; the sequence of the triplet can be in the antisense orientation. Finally, "Nonsense?" answers the following question: would mutating this triplet lead to a premature STOP codon?

| Position | $C_1$ | $C_2$ | $C_3$ | $C_1$[16-20] | Rank in codon | Triplet | Nonsense? |
|---|---|---|---|---|---|---|---|
| **448** | 5 | 8 | 14 | 139 | 1 | tac | No |
| **386** | 6 | | | C]16-20[ | | #N/A | |
| **373** | 2 | 9 | 10 | 217 | 1 | #VALUE! | To check |
| **174** | 5 | | | 388 | 1 | cag | Yes |
| **26** | 4 | 19 | 20 | 535 | 1 | caa | Yes |

Despite the effectiveness of my gRNA design table, a more accessible and versatile tool was eventually published and made available online (http://www.rgenome.net/be-designer/) [332]. Accordingly, the genome *C. autoethanogenum* has been submitted to the curators of this online tool, and its use is recommended in the future. For the interested reader, the logic and the code of the Target-AID gRNA design table is detailed in the supplementary materials (cf. Chapter VIII.1.1).

### V.2.2. Optimisation of DNA transfer and Target-AID expression

#### *V.2.2.1) Codon optimization of AID*

Nishida et al. chose cytosine deaminase 1 from *Petromyzon marinus* (PmCDA1) to fuse at the 3'-end of a nCas9(D10A) in order to assemble their base editor. It is a eukaryotic protein which is 208 aa long. They added 5 aa (SRGSG) at its 3'-end, as well as a 104 aa long linker at its 5'-end. This long linker fuses pmCDA1 to nCas9, helps the fusion protein to penetrate the eukaryotic nucleus with a Nuclear Localisation Signal, and provides enough space and flexibility for pmCDA1 to modify the cytosines in the target protospacer. The resulting peptide is 317 aa long with a GC-content of 46%, and a codon usage optimized for expression in human cells.

Theoretically, the general design of Nishida et al. was directly compatible with a prokaryotic expression system. However, to facilitate the expression of the protein in *C. autoethanogenum*, pmCDA1 was codon-optimized.

#### *V.2.2.2) Removal of undesirable restriction sites*

During codon-optimization, the restriction sites which are used in routine assembly with the pMTL8000 vector series were excluded (NotI, SpeI, XbaI, SbfI, PmeI, NdeI, SalI, AsiSI, AscI, FseI). The sequence of the restriction site of a putative endogenous Type II RM system of *C. autoethanogenum* (Cau10061II) was also avoided.

Indeed, a Cau10061II knock-out had been shown to exhibit an increased conjugation efficiency [236], so it was hypothesized that removing the restriction site targeted by Cau10061II would improve the conjugation potential of the Target-AID vector in WT *C. autoethanogenum*. The sequence `GTTAAT` targeted by Cau10061II was identified as a target for the putative methyltransferase and nuclease encoded by CLAU_0514 (http://rebase.neb.com, 04/08/2017) [333,334]. It is the only one of the four putative Type II RM systems to have an associated nuclease domain. Consequently, the sequence `GTTAAT` and its reverse-complement were both avoided during the codon optimization of Target-AID.

Interestingly, this sequence was also found once in the base pMTL vector, three times in nCas9, and once more in the LacI repressor which would potentially be used later in the project. Accordingly, all Cau10061II restriction sites were removed from these parts in the hope of increasing their conjugation yield. Additionally, two NdeI restriction sites were found in Cas9. As this would be problematic for later cloning stages, these restriction sites were

also disrupted in the same step. The locus of each restriction site and their associated silent mutation are summarized in Table 14.

**Table 14: Summary of the disruption of Cau10061II and NdeI restriction sites.** The sequence of each restriction site is underlined. A silent point mutation (capital letter) is used to disrupt each restriction site. The corresponding aa and their positions in the protein sequence are also indicated. "Backbone" indicates that the locus of the restriction site is not in a coding sequence of the vector. nCas9= SpCas9 nickase (D10A); Aa = amino acids.

| Locus | Original sequence | Corrected sequence | Aa | Position |
|---|---|---|---|---|
| ***ncas9*** | attaac | atAaac | IN | 201-202 |
| ***ncas9*** | aagttaatc | aaAttaatc | KLI | 890-892 |
| ***ncas9*** | attaac | atAaac | IN | 978-979 |
| ***ncas9*** | catatg | caCatg | HM | 160-161 |
| ***ncas9*** | acatatgct | acTtatgct | TYA | 638-640 |
| ***Backbone*** | attaac | aGtaac | | |
| ***lacI*** | attaac | atAaac | IN | 124-125 |

Each restriction site was removed by designing primers with a single mismatch which would still anneal with the restriction site and consequently disrupt it through successful PCR amplification. This resulted in the vectors vFS08_pMTL83151_nCas9_trthl_RSKO (for "Restriction Site Knock-Out"), pMTL83151_RSKO, and vFS20_LacI_RSKO – all devoid of undesired NdeI or Cau10061II restriction sites.

#### V.2.2.2)(1) Disruption of Cau10061II does not improve conjugation yield

In order to test whether the removal of Cau10061II actually improved conjugation yield, the restriction sites-free nCas9(D10A) expression cassette of vFS08 was integrated into the restriction sites-free pMTL83151_RSKO backbone, resulting in the vector vFS70_pMTL83151_nCas9_RSKO – without a Cau10061II restriction site. In parallel, the original nCas9(D10A) was also moved into the WT pMTL83151 backbone to form the vector vFS71_pMTL83151_nCas9 – a vector with a total of four Cau10061II restriction sites. The expression cassette of these vectors was not functional because of a truncated RBS, and no sgRNA expression cassette or gene editing template were provided on the vectors.

vFS70_pMTL83151_nCas9_RSKO and vFS71_pMTL83151_nCas9 were conjugated into *C. autoethanogenum* alongside pMTL83151 as a positive control. One colony of each sExpress donor strain was inoculated straight from the conjugation plates. Three times 1mL were taken from each liquid culture to each undergo mating with 200 µL of *C. autoethanogenum*. After 20h of mating on YTF plates without antibiotics, each mating slurry was normalized to 600 µL, of which 200 µL were plated on transconjugant selection plates.

No significant difference could be observed in between the number of transconjugants obtained with vFS70_pMTL83151_nCas9_RSKO and vFS71_pMTL83151_nCas9 (Welch's t-test, n=3, p=0.4) (cf. Figure 57). Consequently, removing the Cau10061II sequence from vectors does not affect their conjugation yield significantly.

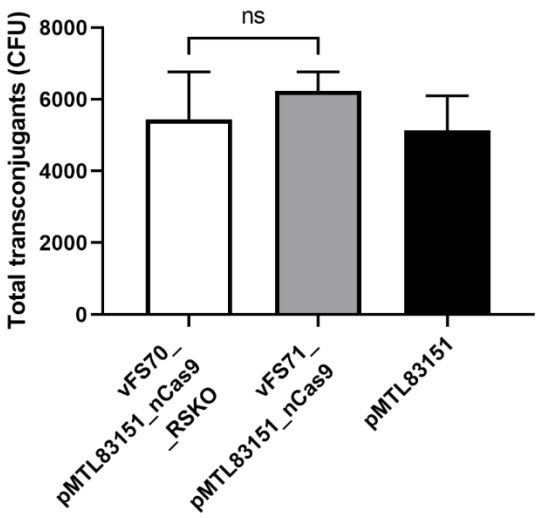



**Figure 57: Impact of the disruption of four Cau10061II restriction sites on conjugation yield.** No significant difference was found in between the number of transconjugants from a vector with four Cau10061II restriction sites (vFS71_pMTL83151) or zero (vFS70_pMTL83151_nCas9_RSKO) (Welch's t-test, n=3, p=0.4).

Interestingly, knocking-out the putative Type II methyltransferase-nuclease which was predicted to target Cau10061II had been shown to increase ten-fold the conjugation efficiency of pMTL83151 [236]. This contradictory result might indicate that the predicted sequence of Cau10061II is incorrect.

Because disrupting the Cau10061II sites did not harm conjugation efficiency, and because the NdeI restriction sites had been removed in the same step, vFS08_pMTL83151_nCas9_trthl_RSKO and the other restriction sites-free parts were kept as a template for later assembly steps involving nCas9(D10A).

### V.2.2.2)(2) Both NdeI restriction sites in ncas9(D10A) were successfully disrupted

To assess whether the disruption of the NdeI restriction sites in the CDS of nCas9(D10A) had been effective, 0.5 µg of vFS04_pMTL83151_nCas9_trthl and vFS08_pMTL83151_nCas9_trthl_RSKO were digested with NdeI at 37°C for 1h and run on a 1% agarose gel. This method was used as a pre-screening immediately after the HiFi assembly of vFS08_pMTL83151_nCas9_trthl_RSKO. The vectors from eight different *E. coli* DH5α colonies transformed with the HiFi mix of vFS08_pMTL83151_nCas9_trthl_RSKO were extracted and digested with NdeI in order to identify insensitive vectors and validate these by Sanger sequencing.

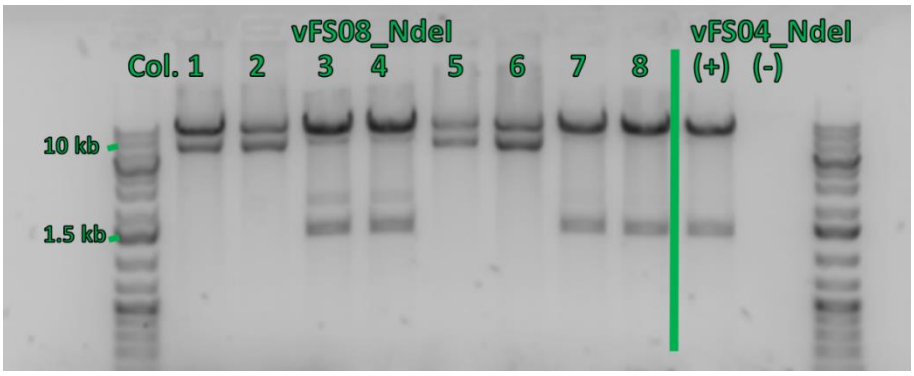



**Figure 58: Pre-screening *E. coli* DH5α vFS08_pMTL83151_nCas9_trthl_RSKO transformants by NdeI digestion.** Col1-8= vectors extracted from eight random *E. coli* DH5α colonies after transformation with vFS08_pMTL83151_nCas9_trthl_RSKO HiFi assembly mix, digested by NdeI; (+) NdeI digested vFS04_pMTL83151_nCas9_trthl; (-) NdeI digestion of water, without DNA template. Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific.

In colonies 1,2,5 and 6, two bands were observed at around 10 kb and 20 kb, typical of the different supercoiled conformations of circular DNA. If only one restriction site had been disrupted, a single band at 10 kb would have been expected, corresponding to the size of linearized vector. In contrast, colonies 3,4,7 and 8 all exhibited a clear band at 1.5 kb, the distance in between both NdeI sites in the CDS of Cas9(D10A), in addition to the 10 kb band of the linearized vector. These same bands band were also observed when the vFS04_pMTL83151_nCas9_trthl, NdeI-sensitive control was treated with NdeI. No band was observed when no plasmid was digested, confirming that none of the bands were the result of contamination.

In conclusion, disrupting NdeI sites with the silent mutations described in Table 14 was sufficient to protect the CDS of nCas9(D10A) and, by extension, the vector vFS08_pMTL83151_nCas9_trthl_RSKO from digestion by NdeI.

## V.3. Preliminary characterisation

This experiment was executed prior to the protocol optimisation described in Chapter IV:. It is indeed the experiment that highlighted some of the weaknesses of my original conjugation protocol. This experiment is thus qualified as “preliminary”, to highlight the fact that it should be reproduced with a more reliable protocol. No technical replicates were made, and subsequent biological replicates will be described separately.

Only 500 µL of PBS was used to resuspend the mating slurry. The mating slurry was not normalized, and 100 µL of the approximately 300 µL mating slurry was spread on each plate (repression or induction plates). On top of all this, the mating slurry was extremely viscous, making the plated volume itself unreliable.

### V.3.1. Experimental setup

The three mutagenesis parameters of Target-AID were characterized in two parallel branches (cf. Figure 59).

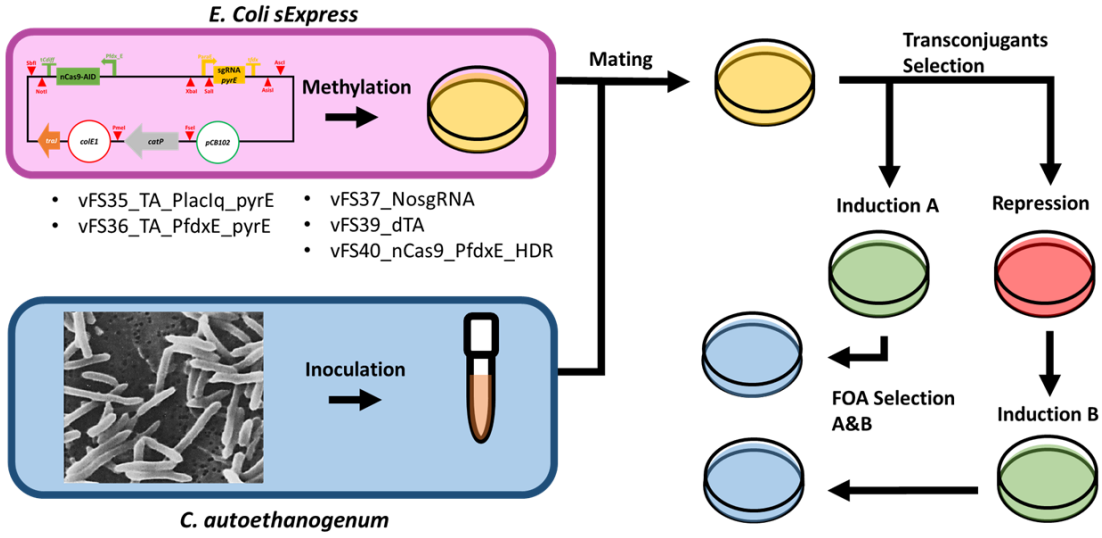



**Figure 59: Schematic of the experimental plan to measure the efficacy of Target-AID.** After conjugation of each *E. coli* sExpress donor strain with an aliquot of *C. autoethanogenum*, the mating slurry is spread in equal volumes on induction and repression plates. After five days, when all colonies have been counted, (a) the induction plates are replicated onto FOA selection plates using velvet stamps; and (b) the repression plates are replicated onto induction plates using velvet stamps; after five more days, the resulting induction plates are replicated onto FOA selection plates using velvet stamps. All plates but the mating plate are also supplemented with D-cycloserine and thiamphenicol. *C. autoethanogenum* Scanning electron microscopy picture reproduced from Abrini et al., 1994 [95].

Survival after induction was measured by spreading each mating mix in equal volume both on a repression plate (i.e., a simple transconjugants selection plate) and on an induction plate (i.e., a transconjugant selection plate supplemented with an inducer: IPTG or theophylline). A low number of transconjugants growing on a repression plate would be symptomatic of a

leaky (and thus toxic) system, while a large proportion of colonies surviving induction directly after mating (Induction A) could be indicative of a low toxicity of the construct.

The survival after induction was also measured by replicating the repression plate onto a second induction plate (Induction B) with the help of a velvet stamp. This was an attempt at mitigating the error on the volume of mating mix which had actually been spread on each plate. Although more precise, this measurement of survival after induction was expected to be less accurate, since the colonies growing on the induction plates B would be the result of the induction of a whole colony instead of a single cell. However, this stage might be necessary if Target-AID turned out to be too toxic to be induced immediately after mating.

Lastly, mutagenesis efficiency was calculated by comparing the number of colonies which survived FOA-selection with the number of colonies which survived on induction plates (without FOA-selection). FOA could potentially interfere with mutagenesis, for example, by killing the cells before they had the opportunity to mutate. Consequently, induction and FOA-selection were done sequentially. Five days after the induction plates had been inoculated with mating slurry (A) or with a velvet stamp from repression plates (B), when individual transconjugant colonies had become visible to the naked eye and new colonies had stopped emerging, they were transferred to a FOA-selection plate by replica-plating with a velvet stamp. The FOA selection plate also contained the inducer molecule to ensure the selective pressure of Target-AID was not lost. The colonies growing on the selection plates were counted four days later.

## V.3.2. Results and discussion

### *V.3.2.1) PlacIq/Pfacoid repression was not effective*

The vector vFS35_TA_PlacIq_LacI_*pyrE* did not yield any transconjugants, be it on repression or on induction plates. This seems to indicate that the $P_{facoid}$/$LacI^{Q}$ system is too leaky.

Upon closer inspection, although the $P_{lacI}^{Q}$ promoter was known to be strongly expressed in *E. coli* [335], it had never been characterized in *C. autoethanogenum*, so it might not be as strongly expressed in *C. autoethanogenum* as in *E. coli*. Additionally, the Target-AID expression cassette of vFS35_TA_PlacIq_LacI_*pyrE* is exposed to transcriptional read-throughs from the LacI expression cassettes, as both were encoded in the same direction. Re-designing the $P_{facoid}$/Laci repression system to express LacI with a strong *C. autoethanogenum* promoter, and changing the orientation of the LacI expression cassette, might consequently improve the repression system and make it compatible with Target-AID.

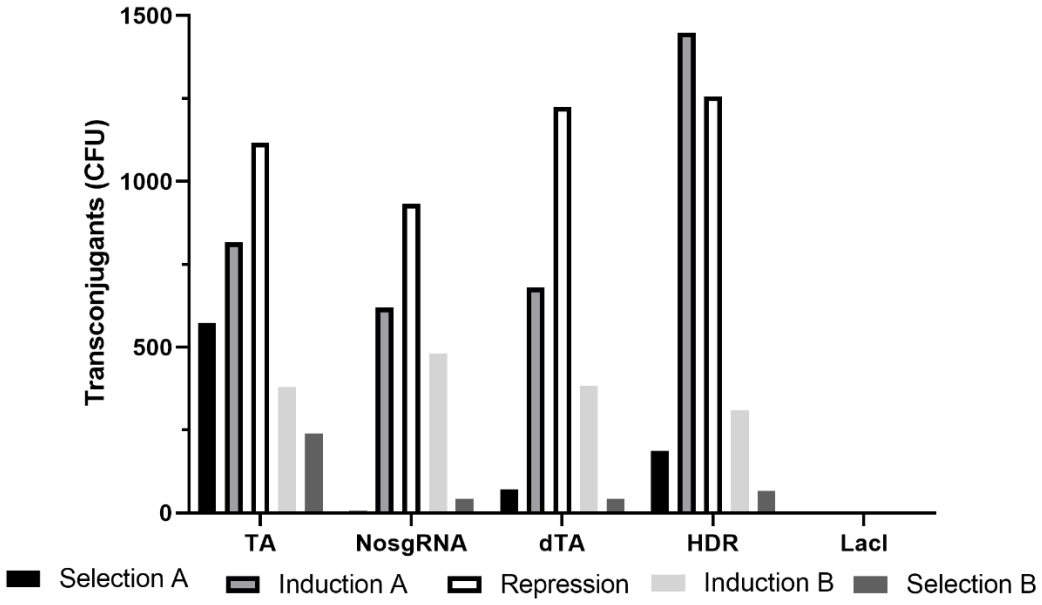



**Figure 60: Raw conjugation yield of Target-AID and variant constructs targeting *pyrE* in *C. autoethanogenum*.** TA = vFS36_TA_PfdxE_pyrE, Target-AID vector with sgRNA cassette targeting *pyrE*; NosgRNA= vFS37_NosgRNA, Target-AID vector without sgRNA cassette; dTA=vFS39_dTA, Target-AID vector with a catalytically inactive Cas9 domain, with sgRNA cassette targeting *pyrE*; HDR= vFS40_nCas9_PfdxE_HDR, Cas9-HDR mutagenesis vector with nCas9, a *ΔpyrE* editing cassette and a sgRNA cassette targeting *pyrE*; LacI= vFS35_TA_PlacIq_LacI_*pyrE*, Target-AID vector under the control of the Pfacoid promoter and a LacI expression cassette. A= induction after mating; B=induction after repression plates.

### *V.3.2.2) $P_{fdxE}$ effectively suppresses the toxicity of Target-AID and the control constructs*

In contrast, all the other constructs yielded about a thousand colonies on repression plates, and about 600 colonies on induction plates (cf. Figure 60). The one exception to this trend is vFS40_nCas9_PfdxE_HDR which actually yielded more transconjugants on induction plates than repression plates (1448 CFU against 1256 CFU, respectively). This resulted in a percentage of survival after induction of 115% (cf. Figure 61), which is of course nonsensical. It was concluded that this difference is not significant, and that both colony counts are probably equivalent.

The fact that all constructs gave relatively equivalent numbers of transconjugants, both on repression and induction plates, indicated that no one construct is more toxic than another when under the control of $P_{fdxE}$. This is good news for Target-AID, considering that at least one construct (vFS37_NosgRNA) should not be toxic at all. In this aspect, it is surprising that vFS40_nCas9_PfdxE_HDR, which expresses a nuclease constitutively, resulted in a higher percentage of survival after induction than vFS37_NosgRNA (which does not express a sgRNA) or vFS39_dTA (which expresses a catalytically inactive Cas9).

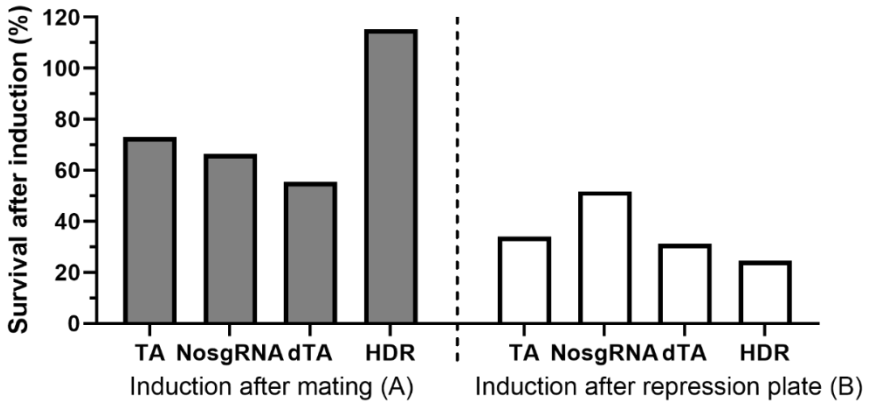



**Figure 61: Percentage of survival after induction of Target-AID vectors and a Cas9-HDR vector targeting *pyrE* in *C. autoethanogenum*.** TA = vFS36_TA_PfdxE_pyrE, Target-AID vector with sgRNA cassette targeting *pyrE*; NosgRNA= vFS37_NosgRNA, Target-AID vector without sgRNA cassette; dTA=vFS39_dTA, Target-AID vector with a catalytically inactive Cas9 domain, with sgRNA cassette targeting *pyrE*; HDR= vFS40_nCas9_PfdxE_HDR, Cas9-HDR mutagenesis vector with nCas9, a *ΔpyrE* editing cassette and a sgRNA cassette targeting *pyrE*.

Indeed, vFS39_dTA obtained the lowest percentage of survival after induction in the branch A (induced right after mating), with only 56%, against 67% for vFS37_NosgRNA and 73% for vFS36_TA_PfdxE_pyrE (cf. Figure 61). It is similarly difficult to explain how expressing a functional Target-AID could be somehow less toxic than expressing a Target-AID without

sgRNA or without functional nickase domain. It was assumed that the colony counts which have led to these results are too imprecise, and thus some of these differences in the branch A of the survival after induction might not be significant.

#### *V.3.2.3) Induction after transconjugants selection on repression plates does not increase the percentage of survival after induction*

In contrast to the above, the survival after induction of colonies replicated from repression plates (branch B) was the highest for vFS37_NosgRNA (52%) and the lowest for vFS40_nCas9_PfdxE_HDR (24%) (cf. Figure 61). This makes more biological sense, when comparing the constructs solely within branch B. However, in direct opposition to the starting hypothesis, survival after induction in branch B was systematically lower than in branch A. Because whole colonies were induced in branch B (instead of single cells in branch A), the probability that at least one cell per colony survived induction and went on to grow into a colony on the induction plate should have been higher than the probability of individual cells surviving induction right after mating. These surprising result could not be explained. Potentially, the physiological state of cells right after mating might facilitate DNA repair (in particular HDR), but that would not explain the increased toxicity of the induction of vFS37_NosgRNA, which should not affect survival at all.

#### *V.3.2.4) Target-AID has the highest mutagenesis efficiency*

Contrarily to what was observed in the survival efficiency, both the absolute number and the proportion of colonies which survived FOA selection were clearly the highest in vFS36_TA_PfdxE_pyrE, in both branches of the experiment (cf. Figure 62). The raw number of FOA-resistant colonies was more important in branch A (574 CFU) than in branch B (240 CFU), but their mutagenesis efficiency was equivalent (70% in branch A, 63% in branch B). The mutagenesis efficiency of all the other constructs was also quite similar in each branch of the experiment, with vFS40_nCas9_PfdxE_HDR reaching the second highest value (13% in branch A, 22% in branch B), and vFS37_NosgRNA the lowest value (1% in branch A, 9% in branch B), with vFS39_dTA in between (10% in branch A, 11% in branch B).

The slight increase in mutagenesis efficiency observed in all constructs but vFS36_TA_PfdxE_pyrE in branch B might be explained by the smaller number of colonies on the induction plates of branch B: less colonies implies that more nutrients and space are available, and thus each colony has potentially more cells in branch B. This increases the chance that at least one colony acquires a mutation which allows it to survive on FOA.

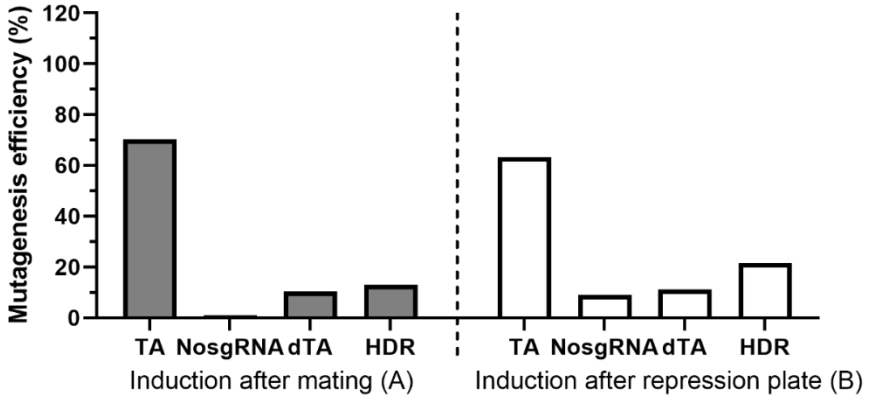



**Figure 62: Mutagenesis efficiency of Target-AID vectors and a Cas9-HDR vector targeting *pyrE* in *C. autoethanogenum*.** TA = vFS36_TA_PfdxE_pyrE, Target-AID vector with sgRNA cassette targeting *pyrE*; NosgRNA= vFS37_NosgRNA, Target-AID vector without sgRNA cassette; dTA=vFS39_dTA, Target-AID vector with a catalytically inactive Cas9 domain, with sgRNA cassette targeting *pyrE*; HDR= vFS40_nCas9_PfdxE_HDR, Cas9-HDR mutagenesis vector with nCas9, a *ΔpyrE* editing cassette and a sgRNA cassette targeting *pyrE*. Calculated from the proportion of colonies growing on FOA after having been induced on theophylline.

#### *V.3.2.5) HDR-mediated mutagenesis was surprisingly ineffective*

At this stage, it is also important to point out the unexpectedly low mutagenesis efficiency of vFS40_nCas9_PfdxE_HDR (13% to 22%). In Chapter III:, a conjugation efficiency of approximately 90% was reached by exploiting the HDR pathway, although this was achieved by expressing a truncated Cas9 constitutively and measuring conjugation efficiency after patching rather than after FOA selection.

This discrepancy might indicate that measuring conjugation efficiency after patching artificially increases conjugation efficiency. On the other hand, it might also indicate that something went wrong in the experimental setup of my characterisation of Target-AID. Perhaps colonies were not properly transferred from the induction plates to the FOA selection plates; maybe the FOA-selection plates of vFS40_nCas9_PfdxE_HDR dried out faster than the other constructs; or perhaps the plasmid somehow mutated in sExpress, prior to the mating.

Lastly, the low mutagenesis efficiency of vFS40_nCas9_PfdxE_HDR compared to pMTL431511_BM4, for example, could also be explained by the difference in expression systems: vFS40_nCas9_PfdxE_HDR was expressing nCas9 under the control of $P_{fdxE}$, while pMTL431511_BM4 was expressing trCas9 under the control of $P_{thl}$ – presumably at very low levels. Interestingly, this implies that swapping the $P_{fdxE}$-nCas9-CDA cassette of vFS36_TA_PfdxE_pyrE for a Pthl-trCas9-CDA cassette might improve mutagenesis efficiency of Target-AID.

### *V.3.2.6) FOA-resistant colonies have the expected genotype*

In order to confirm that the FOA-resistant colonies have the expected genotype, the *pyrE* locus of colonies from all constructs was amplified with oFS105 and oFS106. Twenty-three vFS40_nCas9_PfdxE_HDR colonies were screened by simple electrophoresis gel, while the amplicons of a total of fifteen vFS36_TA_PfdxE_pyrE colonies, three vFS37_NosgRNA colonies and three vFS39_dTA colonies were sent to sequencing.

As already observed in Chapter III.5, colony-PCR of the *pyrE* locus seemed particularly difficult. Of the twenty-three vFS40_nCas9_PfdxE_HDR colonies screened, only fifteen could be amplified. Only eleven of these (73%) had the expected ~500 bp deletion (cf. Figure 63). Consequently, the actual mutagenesis efficiency of vFS40_nCas9_PfdxE_HDR is probably closer to 16% in branch B, instead of the 22% deduced from the number of FOA-resistant colonies alone.

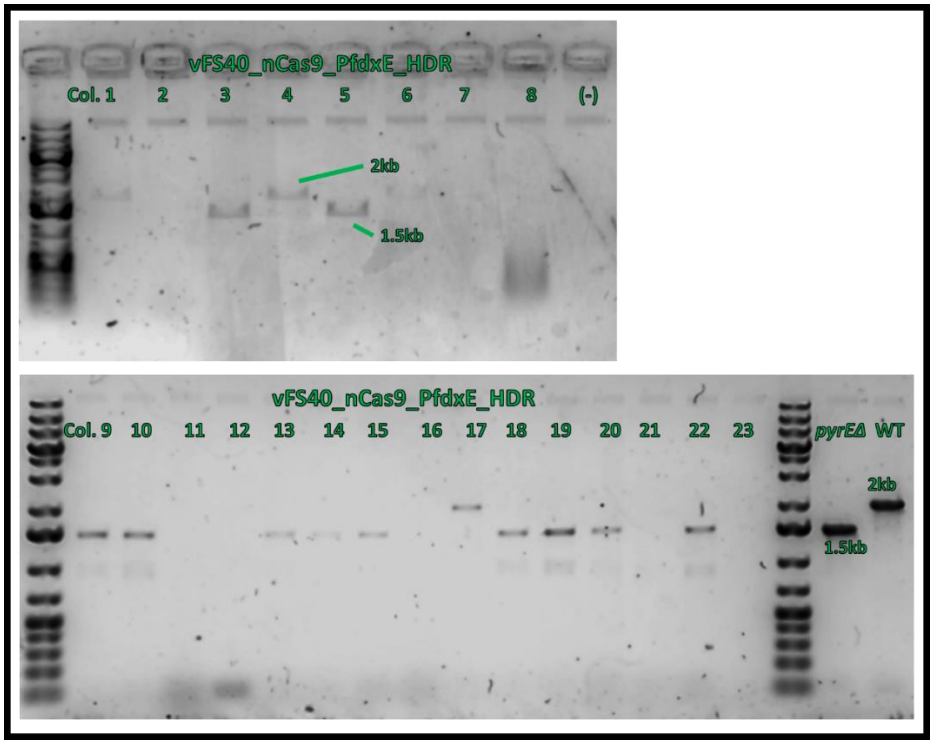



**Figure 63: Gel electrophoresis of the amplification of the *pyrE* locus from 23 *C. autoethanogenum* vFS40_nCas9_PfdxE_HDR colonies collected from FOA-selection plates.** *C. autoethanogenum* WT and *ΔpyrE* strains were also amplified as a positive controls. Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific.

On the other hand, all fifteen vFS36_TA_PfdxE_pyrE and three vFS39_dTA colonies had the intended mutation. Interestingly, all three vFS37_NosgRNA colonies had single base mutations affecting a cytosine elsewhere in the *pyrE* locus (at 210 bp and 434 bp from the start of the CDS, respectively) (cf. Table 15 and Figure 64).

**Table 15: Sequencing of the *pyrE* locus of *C. autoethanogenum* FOA-resistant colonies conjugated with vFS36_TA_PfdxE_pyrE (TA), vFS37_NosgRNA (No sgRNA) or vFS39_dTA (dTA) and induced on theophylline.** The affected codon is highlighted in green, and the mutated base is capitalized. " – " = deletion.

| Position (bp) | Allele | Protospacer sequence | PAM | Mutagenesis efficiency TA | NosgRNA | dTA |
|---|---|---|---|---|---|---|
| 388 | WT | aattcaggaattaggtggag | AGG | | | |
| | A.1 | aattTaggaattaggtggag | | 15/15 | | 3/3 |
| 210 | WT | actaatatccccccattgc | AGG | | | |
| | B.1 | Actaatat-ccccccattgc | | | 1/3 | |
| | B.2 | Actaatatccccccccattgc | | | 1/3 | |
| 434 | WT | tctgtctactatgcaacata | TGC | | | |
| | C.1 | tAtgtctactatgcaacata | | | 1/3 | |

The C210 was deleted, leading to a frameshift mutation which is sure to knock out the gene. A canonical PAM (AGG) was found 11 bp downstream of C210, although, in the absence of cognate sgRNA, this might just be a coincidence. In another colony, the C434 was mutated into an adenine, which is not the main product of Target-AID mutagenesis but is a possible outcome [298,304]. However, no canonical PAM could be found downstream of C434 (NGG, or to a lesser extent NAG and NGA). The number and the proportion of vFS37_NosgRNA FOA-resistant colonies were both small, but the fact that their mutations could potentially be associated with off-target activity of Target-AID is worrying. Unfortunately, in the current experimental setup, there is no way of verifying if these particular mutations would have also happened with the same frequency in the absence of Target-AID. Conjugating an empty vector (pMTL83151) alongside the other controls should enable us to collect that information in the next iterations of this experiment.

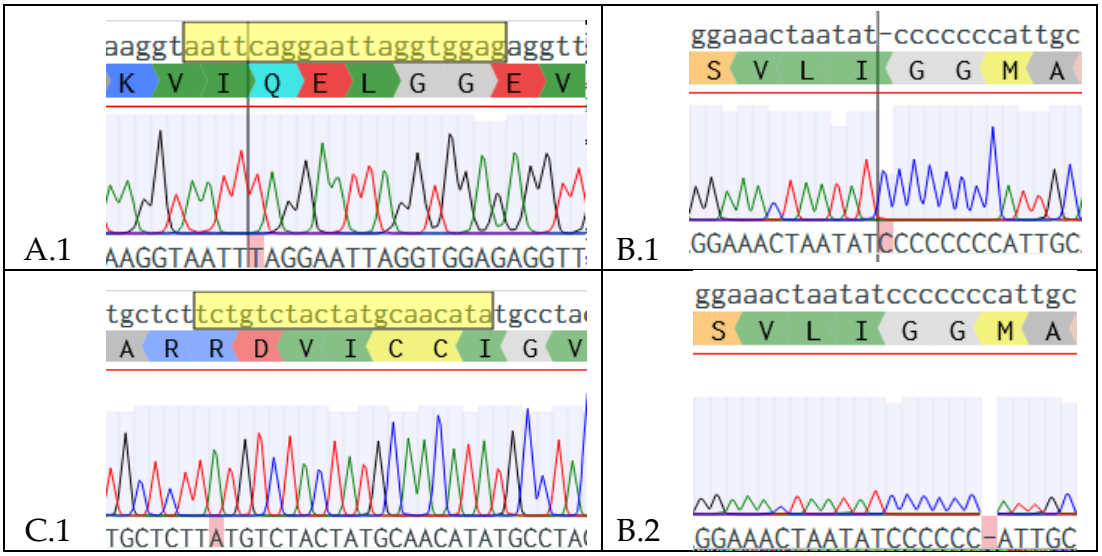



**Figure 64: Representative Sanger sequencing traces of different alleles obtained after conjugation of vFS36_TA_PfdxE_pyrE and vFS39_dTA (A.1), or vFS37_NosgRNA (B.1, B.2, and C.1) in *C. autoethanogenum*.** The sequence of the WT protospacer (or pseudo-protospacer) is highlighted in yellow; the corresponding aa sequence is displayed in a series of consecutive arrows; a vertical bar strikes the start of the targeted codon (in the forward direction); mismatches clear enough to be detected automatically are highlighted in red. Traces and annotations obtained with Benchling.

### *V.3.2.7) Limitations of the study*

The lack of precision of the original conjugation protocol has already been commented on in Chapter IV:. It was optimized over the course of several experiments described in that Chapter. Here it can be further commented that the number of cells plated was excessive: it is laborious and imprecise to count over 1000 colonies per petri dish. Future experiments should try to plate fewer cells, whether by diluting the mating slurry or plating a smaller volume of it.

The decision to plate less cells for more accurate CFU counting unfortunately coincided with the sharp drop in conjugation efficiency which was described in Chapter IV: Once good levels of conjugation efficiency were recovered, it was decided to plate four times fewer cells in the next experiments as in this preliminary experiment (1/12$^{th}$ of the mating mix instead of approximately 1/3$^{rd}$) in order to stay in the vicinity of 250 CFU per plates.

Additionally, it cannot formally be excluded that a construct has mutated while being cloned in the donor strain. To avoid this possibility in later iterations of this experiment, it was decided to prepare cryostocks of donor strains instead of inoculating them for mating straight from their transformation plate. This would allow the extraction of their vector and the verification of its sequence one last time prior to conjugation. However, as described in

Chapter IV.3.2, this turned out to harm conjugation efficiency of all vectors in the long run, so it is a measure which was eventually reversed.

Another major problem was the lack of a pMTL83151 empty vector control. Initially, its role of negative control was supposed to be filled by vFS37_NosgRNA, which was not supposed to be catalytically active in the absence of sgRNA. However, it became apparent during data interpretation that lacking a true negative control in the form of an empty vector was preventing proper interpretation of the data from vFS37_NosgRNA. Were the vFS37_NosgRNA FOA-resistant colonies occurring at a faster rate than WT background mutagenesis rate? Was the nature of their mutations influenced by Target-AID at all? The next iterations of this preliminary experiment would thus involve a pMTL83151 control.

Lastly, the lack any replicate made it hard to identify which difference was significant and which was not. Adding technical replicates would of course increase the precision of the data (especially the raw colony count), but it would also greatly complicate an already complex experiment. Instead, it was decided to just do several biological replicates. The error from biological replicates (executed several weeks apart) should dwarf the technical error and necessarily include it. Some of the strange results described here (such as the low mutagenesis efficiency of vFS40_nCas9_PfdxE_HDR) might be revealed to be outliers in the next few iterations of this experiment.

### V.3.3. Conclusion and perspectives

In summary, Target-AID was shown to be a very effective genome editing tool in *C. autoethanogenum*. It produced over 800 transconjugants after induction on 5 mM theophylline, more than two-third of which harboured the expected mutation. The percentage of survival after induction was also large enough (73%) that using Target-AID for multiplex mutagenesis is justifiable. Indeed, inducing the expression of Target-AID to produce a triple knock-out in one single step would yield $0.70^3*100\%=34\%$ of triple-KO cells, with a survival rate of $0.73^3*100\%= 40\%$ – if each locus is assumed to be mutated independently from the others. In a starting population of 1000 CFU before induction, that would result in 400 CFU growing on the induction plate – 136 of which would be triple-KO. Consequently, design of the multiple mutagenesis strategy was immediately initiated, which is detailed in the next Chapter.

Importantly, however, several incongruities were observed, as well as major flaws in the experimental design. The conjugation protocol itself was imprecise (cf. Chapter IV:), a true negative control was missing, and, of course, the experiment still had to be replicated.

If confirmed, the surprisingly low mutagenesis efficiency of the HDR-mediated mutagenesis control might indicate that theophylline induction or nCas9 could be less effective than constitutive expression of trCas9. It might thus be interesting to test a Pthl-trCas9-CDA construct.

Finally, splitting the experiment into branch A and B turned out to be mostly redundant, as only the repression plate of branch B was necessary to measure the survival after induction in branch A. Even though it yielded surprising results, the survival after induction calculated in branch B has very little biological sense, as it resulted from the induction of whole colonies. This had been done on purpose, as an insurance, just in case too few colonies had survived the induction directly after mating in branch A. However, given that so many cells did survive induction directly after mating, replication of the colonies growing on the repression plate onto an induction plate would no longer be done in the next iterations of this experiment.

## V.4. Follow-up experiments

Several attempts to reproduce the preliminary experiment failed because of the plummeting conjugation yield discussed in Chapter IV:. Too few colonies were growing to obtain meaningful data. However, while I was optimizing the conjugation protocol, different expression systems for Target-AID were explored and valuable insights were acquired into the toxicity of theophylline in *C. autoethanogenum*.

### V.4.1. Test of alternative expression systems

#### *V.4.1.1) Design of alternative constructs*

To confirm that $P_{fdxE}$ was indeed the best repression system for Target-AID, vFS35_TA_PlacIq_LacI_*pyrE* was redesigned. Due to the poor performance of vFS40_nCas9_PfdxE_HDR in the preliminary characterisation of Target-AID, a Target-AID construct based on trCas9 instead of nCas9 was also assembled.

The first LacI-Pfacoid repression system that was assembled to regulate the expression of Target-AID (vFS35_TA_PlacIq_LacI_*pyrE*) did not produce any transconjugants across two conjugation attempts. Suspecting a leaky repression of Target-AID in *C. autoethanogenum*, the $P_{lacI}^{Q}$ *E. coli* promoter, which was initially driving the expression of the LacI repressor, was replaced with $P_{thl}$ from *C. acetobutylicum*, a promoter with a strong expression level both in *E. coli* and *C. autoethanogenum* (cf. Chapter VIII.2.6). The orientation of the LacI expression cassette was also reversed so as to limit the risk of transcriptional read-through onto the Target-AID locus. This resulted in the vector vFS62_TA_Pthl_LacI_pyrE [Figure 65(a)].

Similarly, the vector vFS96_trTA_pyrE was assembled to test if constitutive expression of a trCas9-CDA protein (referred to as trTarget-AID from now on) improved Target-AID mutagenesis [Figure 65.(b)].

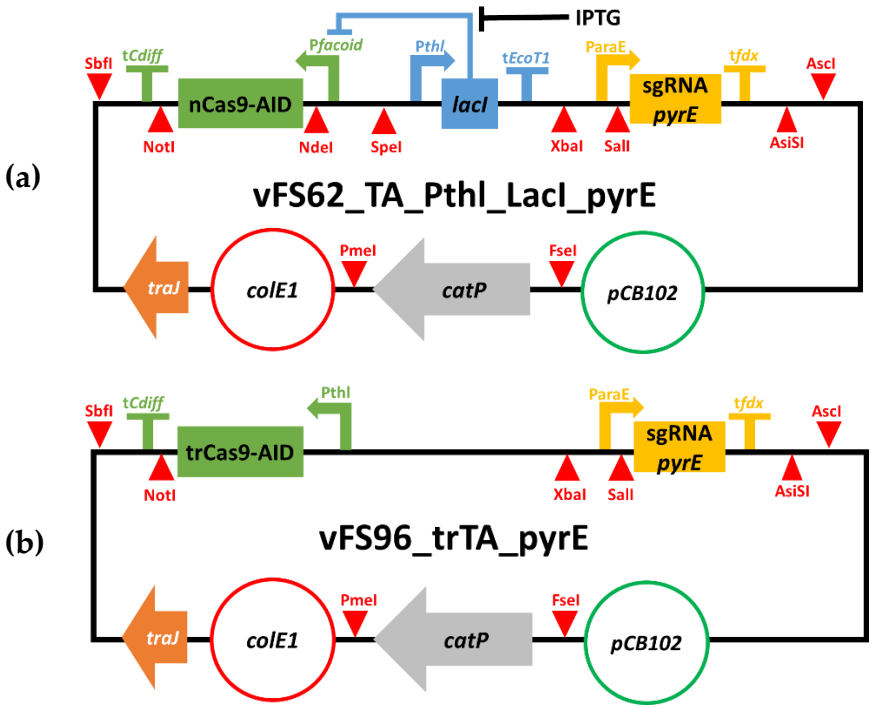



**Figure 65: Plasmid maps of Target-AID vectors with (a) an improved LacI repression system or with (b) a constitutive expression of a trCas9-AID instead of an nCas9-AID.** In vFS62, the $P_{thl}$ promoter from *C. acetobutylicum* is used to express the LacI repressor constitutively instead of $P_{lacI}^{Q}$ in vFS35; the lacI expression cassette is also in reverse orientation compared to vFS35 in order to limit the possibility of transcriptional read-through. In both cases, LacI binds the *lacO* operator in Pfacoid and inhibits transcription of nCas9-AID, while IPTG binds LacI and consequently increases the transcription of nCas9-AID. In vFS96, trCas9-AID is expressed constitutively, but at very low level thanks to the frameshift mutation and the alternative START codon of trCas9.

### *V.4.1.2) Experimental setup*

After improving the conjugation yield of my protocol in Chapter V, it was decided to measure conjugation efficiency with an internal control, namely, the vector pMTL83151. Consequently, all conjugation yields were normalized by the number of colonies obtained through the concomitant conjugation of a pMTL83151 construct. The pMTL83151 empty vector was also used as a negative control to measure mutagenesis.

The performances of the alternative expression systems were compared to the theophylline inducible promoter $P_{fdxE}$ used previously. 600 µL mating slurry was plated in technical triplicates of 25 µL to obtain a more accurate measurement of the colony count. The mating slurry of each construct was plated in parallel on repression and appropriate induction plates (no repression plate for vFS96_trTA_pyrE, induction with 1 mM IPTG for vFS62_TA_PthI_LacI_pyrE and 5 mM theophylline for vFS36_TA_PfdxE_pyrE), then the colonies of the induction plates were transferred to new induction plates complemented with 1 mg/mL of FOA using velvet stamps to select for mutants. The experiment was executed twice several weeks apart

### *V.4.1.1) Different samples failed in each replicate*

The first execution of this experiment failed to obtain any colonies growing on plates complemented with theophylline [cf. Figure 66(a)]. This surprising outcome might have been caused by an extraordinary quick desiccation of the plates. As already mentioned, desiccation regularly disrupts conjugations on its own; however, this time, it seemed to have mainly affected the plates complemented with theophylline, whether their colony was expressing Target-AID or not. This implies that theophylline on its own might be toxic to *C. autoethanogenum* at concentrations close to the 5 mM used during induction of Target-AID.

Theophylline is dissolved into DMSO at close to saturating concentration (250 mM) before being added to the medium; consequently, the observed toxicity of theophylline might also be caused by an excessive DMSO concentration rather than an excessive theophylline concentration. Because the same concentration of theophylline was used during the preliminary experiment, there must be another factor at play; however, it might be worth decreasing the concentration of theophylline and DMSO in future experiments.

In the literature, theophylline has been reported to slow down the growth of *E. coli* and *C. sporogenes* considerably at concentrations of 5 mM and 10 mM, respectively [245,326]. Because the constructs were induced with 5 mM of theophylline, the conjugation rate, the survival rate, or the growth rate of colonies on the transconjugant selection plate might have been affected.

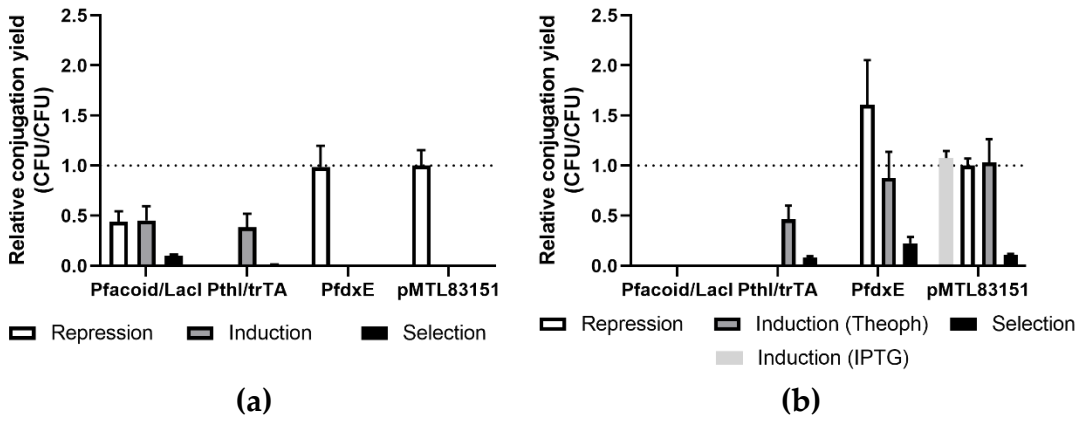



**Figure 66: Conjugation yield of Target-AID vectors using different expression systems targeting *pyrE* in *C. autoethanogenum,* relative to the conjugation yield of pMTL83151, across two replicates.** The first replicate is presented in **(a)**, and the second replicate is presented in **(b)**. The $P_{fdxE}$ expression system is carried by the vector vFS36_TA_PfdxE_pyrE, $P_{facoid}$/LacI by the vector vFS62_TA_PthI_LacI_pyrE and $P_{thl}$/trTA by the vector vFS96_trTA_pyrE. $P_{fdxE}$ was induced with 5 mM theophylline, $P_{facoid}$/LacI with 1 mM IPTG, and $P_{thl}$/trTA is constitutively expressed. pMTL83151 was "induced" with theophylline, even though it does not have an expression cassette. All colonies grew from 25 µL of a 600 µL mating slurry. N=3.

Because so many samples failed to grow, this first replicate was discarded and a second attempt at characterizing alternative expression systems for Target-AID was made [cf. Figure 66(b)]. Unfortunately, no vFS62_TA_Pthl_LacI_pyrE transconjugants could be obtained during this second attempt. Because the pMTL83151 control grew perfectly well on IPTG (1.08±0.07 CFU/CFU relative to the number of pMTL83151 colonies growing on the YTF plate without theophylline), this failure to conjugate cannot be blamed on medium composition. Tentatively, it might thus have been caused by a mutation of the $P_{facoid}$/Laci repression system during methylation in the *E. coli* sExpress strain. Taking into account the leakiness inherent to transcriptional promoters as well as the presence of a native *E. coli lacI* gene in its expression cassette, vFS62_TA_Pthl_LacI_pyrE is indeed at risk of being unstable.

#### *V.4.1.1) $P_{facoid}$/LacI is still leakier than $P_{fdxE}$*

In the first replicate [cf. Figure 66(a)], in the absence of theophylline, vFS36_TA_PfdxE_pyrE yielded as many colonies as pMTL83151 (0.981±0.214 CFU/CFU relative to the number of pMTL83151 colonies), but vFS62_TA_Pthl_LacI_pyrE and vFS96_trTA_pyrE both only yielded about half that amount (0.441±0.104 CFU/CFU and 0.383±0.137 CFU/CFU, respectively). This seems to indicate that the $P_{facoid}$/LacI repression system is leakier than $P_{fdxE}$, maybe even as leaky as the constitutive expression from the $P_{thl}$/trTA expression cassette.

#### *V.4.1.1) $P_{facoid}$/LacI is a suitable expression system for Target-AID*

Perhaps as a consequence of this leakiness, the survival after induction with IPTG of vFS62_TA_Pthl_LacI_pyrE was as maximal (102±23 %) [cf. Figure 67(a)]. This however only translated into a mutagenesis efficiency of 23±7 %, which is double the background noise measured in the pMTL83151 control of the second replicate (11±2%). One single vFS62_TA_Pthl_LacI_pyrE colony growing on FOA was screened by Sanger sequencing, and the expected C388T mutation was observed. This confirmed that $P_{facoid}$/lacI is a suitable expression system for Target-AID, although the resulting vector might be less stable than if $P_{fdxE}$ is used.

At this stage, it is worth highlighting that because no vFS62_TA_Pthl_LacI_pyrE colonies were obtained during the second replicate, the mutagenesis parameters of vFS62_TA_Pthl_LacI_pyrE were calculated from the data obtained during the first replicate. However, because most of the samples in that replicate had to be discarded, the quantitative characterisation of the system should be regarded with suspicion.

### *V.4.1.1) vFS36_TA_PfdxE_pyrE performed worse than expected*

In the second replicate, vFS36_TA_PfdxE_pyrE colonies only reached a survival after induction of 55±4% (instead of 73% during the preliminary experiment) [cf. Figure 67 (a)]. This might be a consequence of the 60% excess of colonies counted on vFS36_TA_PfdxE_pyrE repression plates compared to pMTL83151. vFS36_TA_PfdxE_pyrE could be expected to be more toxic than pMTL83151 even before induction, so this large excess of colonies is disturbing. It might indicate that this transconjugants selection plate was contaminated, or that the antibiotic selection partially failed and allowed some *E. coli* sExpress or WT *C. autoethanogenum* to grow, which would artificially lower the survival after induction.

vFS36_TA_PfdxE_pyrE barely reached the highest mutagenesis efficiency with 25±5% [cf. Figure 67(b)], but this is once again much lower than the 70% obtained during the preliminary characterisation of Target-AID. Even more worrying, no mutation of the *pyrE* locus could be found among 8 randomly selected colonies before or after FOA selection.

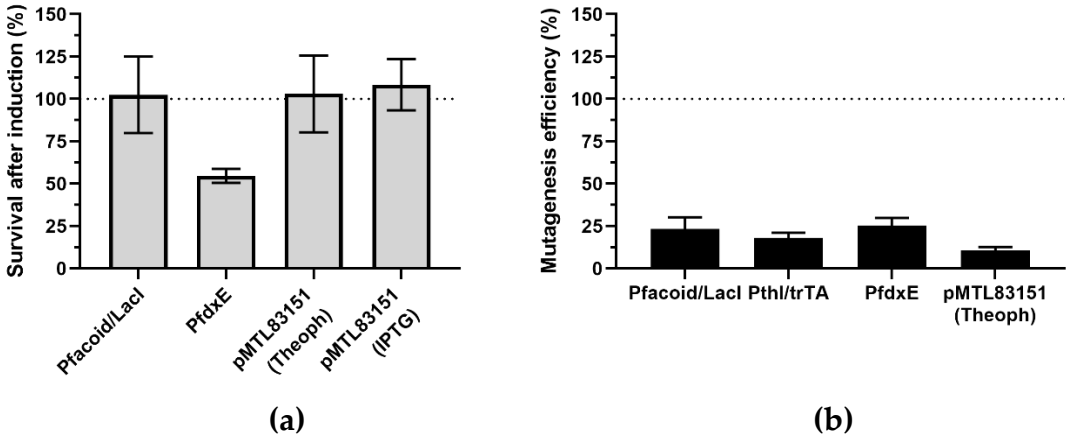



**Figure 67: (a) Percentage of survival after induction and (b) Mutagenesis efficiency of Target-AID vectors using different expression systems targeting *pyrE* in *C. autoethanogenum*.** The $P_{fdxE}$ expression system is carried by the vector vFS36_TA_PfdxE_pyrE, $P_{facoid}$/LacI by the vector vFS62_TA_Pthl_LacI_pyrE and $P_{thl}$/trTA by the vector vFS96_trTA_pyrE. $P_{fdxE}$ was induced with 5 mM theophylline, $P_{facoid}$/LacI with 1 mM IPTG, and $P_{thl}$/trTA is constitutively expressed. pMTL83151 was "induced" with theophylline, even though it does not have an expression cassette. n=2, biological replicates.

### *V.4.1.2) trTarget-AID has a different base editing profile*

Based solely on the data from the second replicate, sequencing of the vFS96_trTA_pyrE both before and after FOA selection revealed a different mutagenesis pattern than those previously observed with vFS36_TA_PfdxE_pyrE (cf. Table 16 and Figure 68). Indeed, only two out of the fourteen colonies sequenced had the intended Q130X mutation. Three more colonies (two from the FOA plate, one from the plate without FOA) saw a deletion of the targeted C388. Practically, this is also an acceptable outcome because such a frameshift mutation is sure to knock out the gene.

However, five more colonies saw mutations in between the positions -4 to -1 from the PAM, or even on the third base of the PAM. This is well outside the expected editing window of -19 to -16 bp from the PAM. Four of these mutations were non-canonical G→T transversions, although three exhibited mixed peaks. The mixed peaks must come from mixed colonies, implying that Target-AID mutagenesis is taking place over time, with some single transconjugant cells initially evading mutagenesis only to see some of their daughter cells mutate. Interestingly, the transversion in position -1 from the PAM also produced a STOP codon (E135X). The G→A transition in position -4 to -3 from the PAM would have been canonical if the PAM had been on the reverse complementary strand (5'CCN in the orientation used in Table 16) At last, one colony which was isolated from the plate without FOA saw an in-frame deletion of 33 bp starting from the position -2 from the PAM.

**Table 16: Sequencing of the *pyrE* locus of *C. autoethanogenum* vFS96_trTA_pyrE colonies on plates with or without FOA.** The targeted codon is highlighted in green, and the mutated bases are capitalized and bolded." – "= deletion.

| Allele | Sequence | | | Mutagenesis efficiency (vFS96) | |
|---|---|---|---|---|---|
| | 5′ | Protospacer (20 nt) | 3′ | FOA | No FOA |
| WT | gtagaggttgctaaggt | aattcaggaattaggtggag | aggtt | 1/7 | 2/7 |
| A.1 | gtagaggttgctaaggt | aatt**T**aggaattaggtggag | aggtt | 2/7 | |
| A.2 | gtagaggttgctaaggt | aatt-aggaattaggtggag | aggtt | 2/7 | 1/7 |
| A.3 | gtagaggttgctaaggt | aattcaggaattaggtgga***T*** | aggtt | | 1/7 |
| A.4 | gtagaggttgctaaggt | aattcaggaattaggtgga$\frac{T}{g}$ | aggtt | 1/7 | |
| A.5 | gtagaggttgctaaggt | aattcaggaattaggtgga$\frac{T}{g}$ | ag$\frac{T}{g}$tt | | 2/7 |
| A.6 | gtagaggttgctaaggt | aattcaggaattaggt**AA**ag | aggtt | 1/7 | |
| A.7 | gta---------- | ----------g | aggtt | | 1/7 |

The fact that roughly half of the observed mutations induced by trTarget-AID were in positions -4 to -1 from the PAM instead of -19 to -16 is probably a direct consequence of the disruption of the RuvC domain of trCas9. Interestingly, but perhaps coincidentally, a GGN sequence can be found directly upstream of the protospacer. Hypothetically, the truncation of RuvC domain could have increased the flexibility of the PAM interacting domain and allowed it to process a PAM of reverse sequence (GGN-protospacer as well as protospacer-NGG). This hypothesis could be easily tested by using trCas9 to target GGN PAMs upstream of protospacers lacking a downstream canonical PAM.

### *V.4.1.3) FOA selection was unreliable*

At least four of the seven vFS96_trTA_pyrE colonies sequenced prior to FOA selection would have survived FOA selection (alleles A.2, A.3 and A.5). This would translate into an actual

mutagenesis efficiency of 57%. Conversely, two of the seven colonies sequenced from the plate with FOA did not exhibit a mutation which should enable them to grow on FOA (alleles WT and A.6). This implies that 29% of the colonies which survived FOA might be false positives, reducing the mutagenesis efficiency from 18±3% to roughly 13% [cf. Figure 67(b)]. At last, one of the colonies surviving on FOA exhibited a mixed peak (A.4), which should not be possible since all the WT cells should have died from the FOA selection.

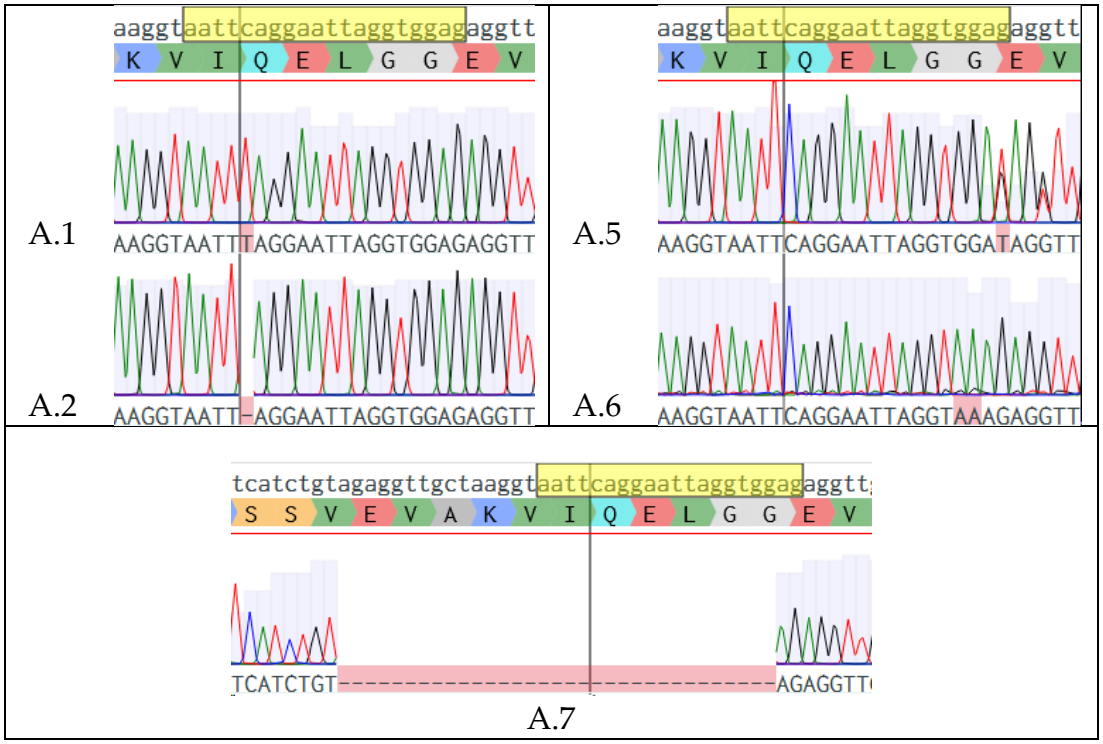



**Figure 68: Representative Sanger sequencing traces of different alleles (A.1, A.2, A.5, A.6 and A.7) obtained after conjugation of vFS96_trTA_pyrE in *C. autoethanogenum*.** The sequence of the WT protospacer is highlighted in yellow; the corresponding aa sequence is displayed in a series of consecutive arrows; a vertical bar strikes the start of the targeted codon (in the forward direction); mismatches clear enough to be detected automatically are highlighted in red. Traces and annotations obtained with Benchling.

The large difference in between the mutagenesis efficiency calculated from sequencing data or from colony counts might just be due to the small sample size of the sequencing data. However, as already mentioned, all eight vFS36_TA_PfdxE_pyrE colonies screened by Sanger sequencing survived on FOA without mutations in the *pyrE* locus. These two datasets together strongly imply that FOA selection was somehow defective in this experiment: many colonies which should have survived died, and many colonies which should have died survived. It is possible that a large proportion of the cells acquired other mutations which knocked out the pyrimidine synthesis pathway, for example within the gene *pyrF* [208,252]. Nonetheless, this would not explain why vFS36_TA_PfdxE_pyrE failed to produce any mutants.

### V.4.2. Theophylline, not DMSO, decreases conjugation yield in *C. autoethanogenum*

In light of the death of all the colonies growing on theophylline induction plates in the first replicate of the previous experiment, the potential toxicity of theophylline and DMSO was investigated in two separate follow-up experiments. Thus, triplicate 50 µL aliquots of pMTL83151 mating slurry were plated on increasing concentrations of theophylline, or on increasing concentrations of DMSO.

Indeed, induction plates with 5 mM of theophylline necessarily also contain 2% (v/v) of DMSO; moreover, because DMSO is also used to dissolve FOA, the concentration of DMSO in selection plates is as high as 3%. The impact of 2% and 3% of DMSO on the number of transconjugants obtained was therefore investigated, without any theophylline or FOA [cf. Figure 69 (a)].

On the other hand, because 5 mM of theophylline had not been so toxic to vFS36_TA_PfdxE_pyrE colonies in the preliminary experiment, it was hypothesized that an experimental error or a particularly high rate of desiccation might have been responsible for increasing the effective concentration of theophylline in the plates. Consequently, the number of pMTL83151 transconjugants were counted when plated on 5 mM of theophylline to confirm the potential toxicity of theophylline at this concentration, and on 10 mM to investigate the consequence of doubling the concentration of theophylline [cf. Figure 69 (b)].

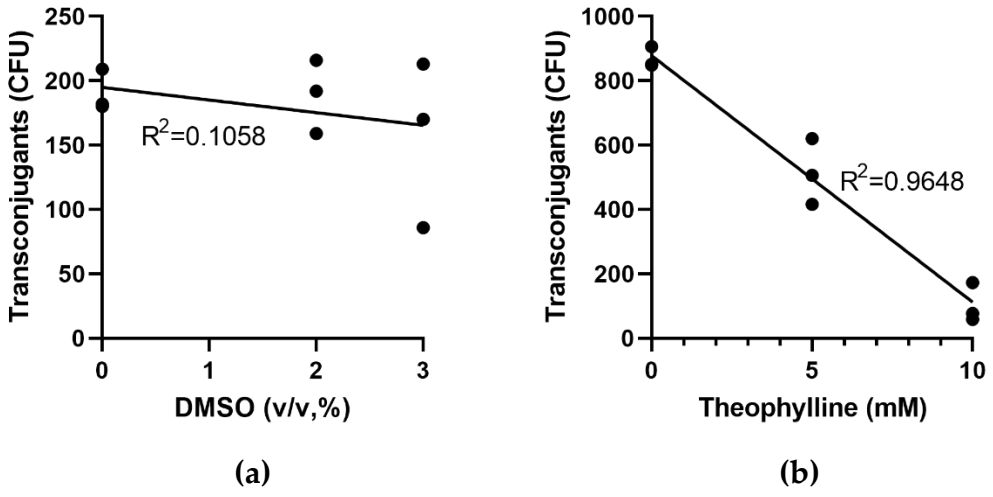
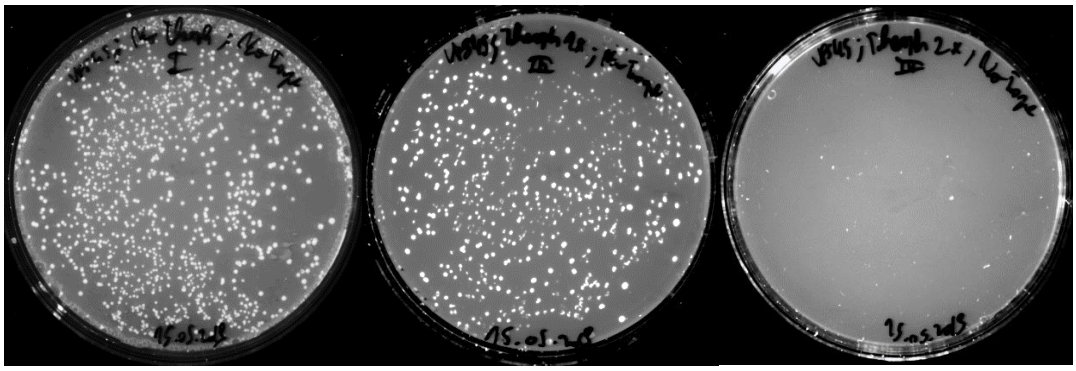



**Figure 69: Impact of increasing (a) DMSO concentration and (b) theophylline concentration on the conjugation yield of *C. autoethanogenum***. v/v, %= concentration calculated as a volume ratio and expressed as a percentage.

A slight decrease in the number of transconjugants was observed at 3% of DMSO, but it is entirely due to one single outlier. Consequently, the concentration of DMSO alone cannot be

considered responsible for the low survival after induction observed during previous experiment. On the other hand, a sharp decrease in the number of transconjugants was observed with increasing theophylline concentrations. 865±30 CFU were counted at 0 mM of theophylline, but only 514±102 CFU (or 59%) grew on 5 mM theophylline. At last, only 103±61 CFU could be counted at 10 mM theophylline, and their colony morphology was small and transparent [cf. Figure 70 (c)].



**(a)** **(b)** **(c)**

**Figure 70: *C. autoethanogenum* colonies conjugated with pMTL83151 and growing on YTF plates supplemented with 250 µg/mL D-cycloserine, 7.5 µg/mL thiamphenicol, and (a) 0 mM (b) 5 mM or (c) 10 mM theophylline.**

Interestingly, this corresponds to a survival after induction of only 59% at 5 mM for pMTL83151 colonies, which do not express any construct and do not even have a PthlE promoter. This is slightly lower than the survival after induction that was measured for vFS36_TA_PfdxE_pyrE during the preliminary experiment (73%), but higher than the 56% obtained by vFS39_dTA. Consequently, it is possible that most or even all of the toxicity that was observed upon induction of Target-AID was actually caused by an excessive concentration of theophylline.

Finally, it is worth mentioning that no toxicity was observed from theophylline alone in the pMTL83151 construct of the previous experiment (cf. Chapter V.4.1). This might indicate that yet another hidden parameter was at play to mitigate the toxicity of theophylline.

### V.4.3. Summary and perspectives

In this section, $P_{facoid}$/LacI repression system was successfully redesigned to be compatible with Target-AID – even if it might still be quite unstable. Target-AID was also combined with trCas9 nuclease in an attempt to make the base editor more convenient and potentially more effective. Each of these two constructs performed much better than the original vFS36_TA_PfdxE_pyrE vector – which perplexingly failed to produce any mutants this time. Indeed, vFS36_TA_PfdxE_pyrE performed much better than them during its preliminary characterization. More data is thus needed to reach a formal conclusion.

It was shown that trTarget-AID had a different base editing profile than the conventional Target-AID, perhaps due to the ability to recognize a PAM upstream of the protospacer. It edited bases in between the positions -4 to +3 of the first base of the targeted PAM – most of which were guanines instead of cytosines – and produced more deletions than the conventional Target-AID – including a 33 bp deletion.

Interestingly, theophylline was found to reduce the number of pMTL83151 transconjugants by as much as 41% when used at the recommended concentration of 5 mM. Moreover, this decrease cannot be attributed to the DMSO used to dissolve theophylline. Although induction at 5 mM of theophylline was previously shown to be effective at expressing Target-AID and produced the expected mutants, reducing this concentration might improve the outcome of the mutagenesis.

This leads us to the first aspect of this whole chapter which could have been improved. A proper characterisation of the strength and dynamic range of each promoter at different inducer concentrations in *C. autoethanogenum* would have identified such a toxicity of theophylline from the start. It would have also enabled interpretation of the different performances of each expression system explicitly in term of transcription levels and dynamic range, instead of estimating these indirectly from the number of colonies growing on repression and induction plates. Rigorously characterizing the $P_{fdxE}$, $P_{facoid}$/LacI, and $P_{thl}$/tr promoters, independently from their interaction with Target-AID, would be a valuable contribution to the synthetic biology toolbox in *C. autoethanogenum*.

## V.5. Final characterisation

Once the conjugation protocol had been optimized (cf. Chapter IV:), and additional insights on the toxicity of theophylline had been gained, improvements to the preliminary characterisation of Target-AID were sought.

### V.5.1. Experimental setup

The experimental design was largely identical to the preliminary characterisation, except that vFS35_TA_PlacIq_LacI_*pyrE* was replaced by a pMTL83151 negative control, the induction and selection steps of branch B were removed, and cells were induced on 2.5 mM theophylline as well as 5 mM theophylline in anticipation of any trade-off in between toxicity of theophylline and expression of Target-AID (cf. Figure 71).

For each construct, 50 µL of the 600 µL mating slurry was thus spread onto one plate without theophylline, one plate with 2.5 mM of theophylline, and one plate with 5 mM of theophylline. As previously, the colonies from the induction plate were transferred to a FOA-selection plate complemented with the appropriate concentration theophylline using a velvet stamp after 5 days of incubation in the anaerobic cabinet. This was done in two biological replicates several months apart, but without technical replicates.

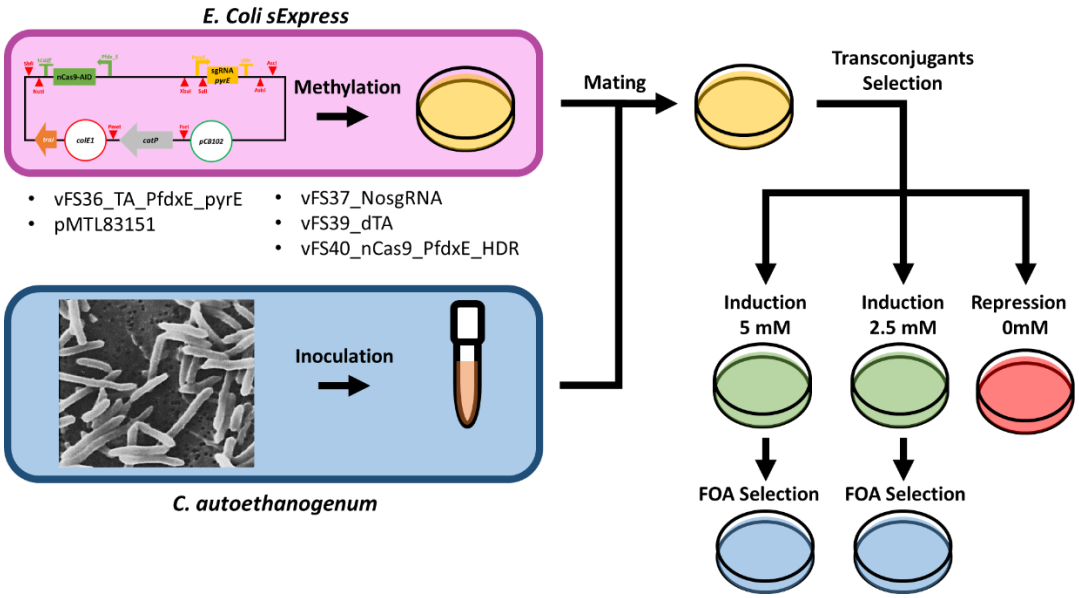



**Figure 71: Schematic of the experimental plan to characterize Target-AID mutagenesis in *C. autoethanogenum*.** After conjugation of each *E. coli* sExpress donor strain with an aliquot of *C. autoethanogenum*, the mating slurry is spread in equal volumes onto three plates with different concentrations of theophylline: 0 mM, 2.5 mM or 5 mM. After five days, when all colonies have been counted, the induction plates are replicated onto FOA selection plates using velvet stamps. All plates but the mating plate are also supplemented with D-cycloserine and thiamphenicol. *C. autoethanogenum* Scanning electron microscopy picture reproduced from from Abrini et al., 1994 [95].

## V.5.2. Results and discussion

The results obtained by inducing the constructs with 5 mM of theophylline will initially be described; the consequences of inducing at 2.5 mM of theophylline will be analysed next.

### *V.5.2.1) Induction with 5 mM of theophylline*

#### V.5.2.1)(1) Relative Conjugation yield on repression plates is highest with pMTL83151

The colony counts of each construct relative to the conjugation yield of the pMTL83151 control construct were surprisingly consistent across treatments, with the exception of the repression of vFS36_TA_PfdxE_pyrE (0.797±0.557 CFU/CFU), and the repression and induction of vFS39_dTA (0.737±0.335 CFU/CFU and 0.548±0.386 CFU/CFU, respectively) (cf. Figure 72). The high standard deviation (SD) of these two constructs prevents any meaningful comparison with remaining constructs. On the other hand, the vFS37_NosgRNA and vFS40_nCas9_PfdxE_HDR colonies growing on repression plate linger at a stable 0.464±0.047 CFU/CFU and 0.358±0.044 CFU/CFU, respectively. This lower yield of transconjugants could be the consequence of some leakiness of the $P_{fdxE}$ promoter. This difference might also be a consequence of the size difference in between pMTL83151 (4.5 kb) and the other vectors (11 kb). However, such an impact of vector size on conjugation efficiency should also have been observed in the experiments described in Chapter V.2.2.2)(1) or Chapter V.4.1.1).

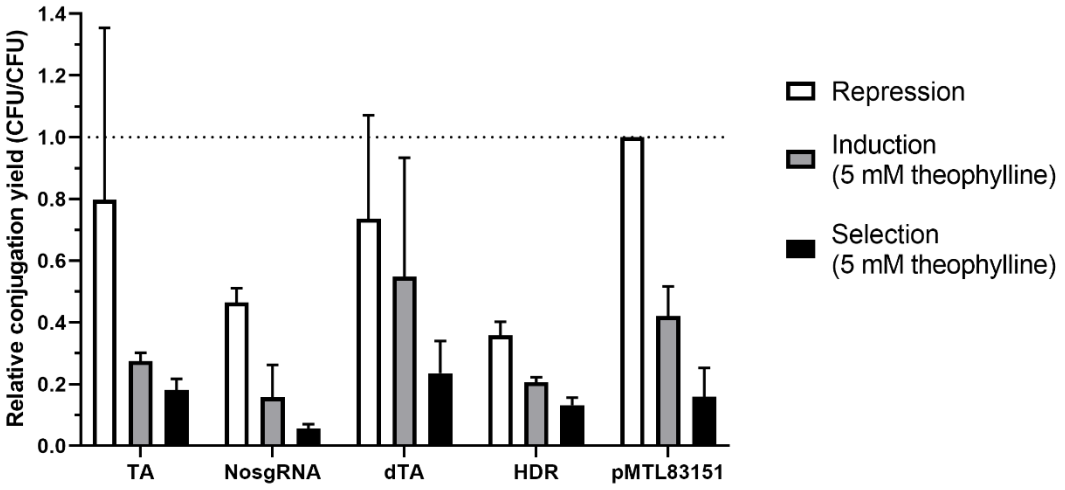



**Figure 72: Relative conjugation yield of Target-AID and variant constructs targeting *pyrE* in *C. autoethanogenum*.** TA = vFS36_TA_PfdxE_pyrE, Target-AID vector with sgRNA cassette targeting *pyrE*; NosgRNA= vFS37_NosgRNA, Target-AID vector without sgRNA cassette; dTA=vFS39_dTA, Target-AID vector with a catalytically inactive Cas9 domain, with sgRNA cassette targeting *pyrE*; HDR= vFS40_nCas9_PfdxE_HDR, Cas9-HDR mutagenesis vector with nCas9, a *ΔpyrE* editing cassette and a sgRNA cassette targeting *pyrE*. All constructs were induced with 5 mM of theophylline. Then, induced colonies were transferred with a velvet stamp to FOA selection plates with the same concentration of theophylline, to select for mutants. Colony count is reported as a proportion of the CFU of the pMTL83151 control construct to facilitate the agglomeration of replicates (n=2).

#### V.5.2.1)(2) Survival after induction is low even in non-mutagenic constructs

As observed previously, even constructs which should not be toxic to the cells (vFS37_NosgRNA and pMTL83151) exhibited a low conjugation yield when plated on 5 mM theophylline (33±20% and 42±9% of survival after induction, respectively) [cf. Figure 73(a)]. A large part of this toxicity could be blamed on theophylline itself (cf. Chapter V.4.2). It is difficult to do any comparison between constructs because of the often large SDs and the low number of replicates, but vFS39_dTA seems to be the least toxic construct after induction (with 70±21% of survival). This might be explained by the absence of functional nuclease domain in this construct. It is unclear why vFS39_dTA and vFS40_nCas9_PfdxE_HDR seemed to be less affected by the toxicity of theophylline than the pMTL83151 control. Perhaps the expression of the theophylline-inducible riboswitch offers some level of protection.

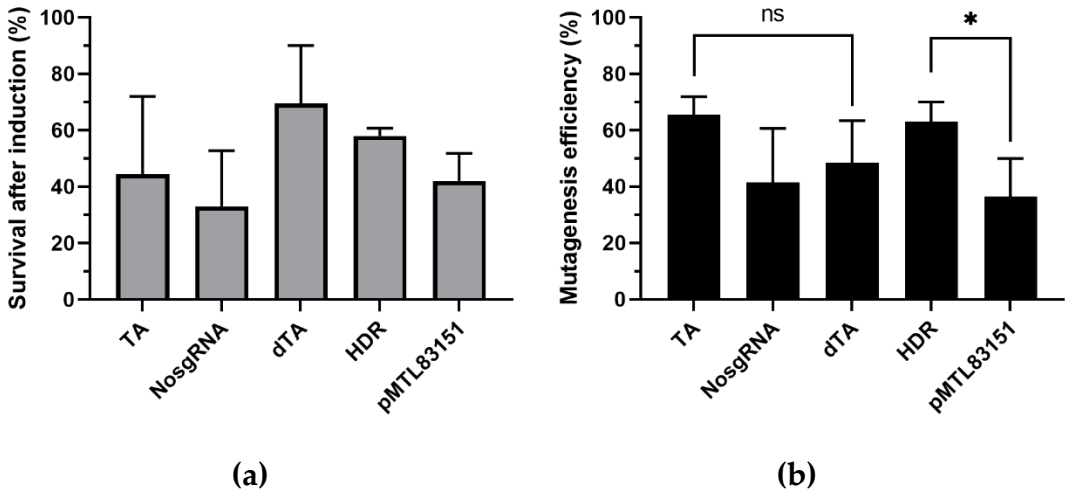



**Figure 73: Percentage of survival after induction (a) and mutagenesis efficiency (b) of Target-AID and variant constructs targeting *pyrE* in *C. autoethanogenum*, after induction at 5 mM of theophylline.** TA = vFS36_TA_PfdxE_pyrE, Target-AID vector with sgRNA cassette targeting *pyrE*; NosgRNA= vFS37_NosgRNA, Target-AID vector without sgRNA cassette; dTA = vFS39_dTA, Target-AID vector with a catalytically inactive Cas9 domain, with sgRNA cassette targeting *pyrE*; HDR = vFS40_nCas9_PfdxE_HDR, Cas9-HDR mutagenesis vector with nCas9, a *ΔpyrE* editing cassette and a sgRNA cassette targeting *pyrE;* ns = non-significant; * = p<0.05. Welch's t-test, n=2.

#### V.5.2.1)(3) The background rate of survival on FOA is surprisingly high

An unexpected 37±14% of pMTL83151 colonies induced on 5 mM of theophylline survived FOA selection [cf. Figure 73(b)]. Because this construct is the empty vector control, which should not be mutagenic at all, this number should represent the background rate of random mutagenesis which allows some cells to survive FOA selection even though they escaped targeted mutagenesis. 37% of background mutagenesis efficiency is far too high to be

plausible, which calls the validity of the experimental setup into question. My FOA stock might have been defective, or perhaps the small biofilm constituted by individual colonies grants them some level of protection from FOA when they were transferred onto FOA selection plates. Similarly, 42±19% of vFS37_NosgRNA colonies induced on theophylline also survived on FOA. This is once again too high to be plausible, but it is reassuring that it is so close to the value obtained with the pMTL83151 control: at face value, it would imply that the rate of off-target mutagenesis induced by Target-AID is low.

#### V.5.2.1)(4) Target-AID seems as effective as conventional HDR-mediated mutagenesis

Nonetheless, vFS36_TA_$P_{fdxE}$ and vFS40_nCas9_PfdxE_HDR both reached the highest mutagenesis efficiency with 66±6% and 63±7%, with a significant (or nearly significant) difference compared to pMTL83151 (Welch's t-test, n=2, p=0.049 and p=0.053, respectively) [cf. Figure 73(b)]. These numbers are encouraging, but they show that Target-AID is unlikely to be more effective than a conventional HDR-mediated mutagenesis strategy. The mutagenesis efficiency of Target-AID with a catalytically inactive Cas9 (vFS39_dTA) seems slightly lower than that of vFS36_TA_$P_{fdxE}$ with only 49±15%, but no statistically significant difference could be found (Welch's t-test, n=2, p=0.571).

### *V.5.2.2) Induction with 2.5 mM of theophylline*

#### V.5.2.2)(1) Variability increases and toxicity decreases with decreasing theophylline concentrations

After induction at 2.5 mM, the raw number of colonies increased across all constructs compared to induction at 5 mM [cf. Figure 74 (a)]. Especially notable are the increase in vFS36_TA_$P_{fdxE}$, vFS37_NosgRNA and pMTL83151, which all gained an average of 0.3 units of conjugation yield, relative to the conjugation yield of pMTL83151 on repression plates. This naturally translated into a sharp increase of survival after induction, with all constructs scoring above 60% and vFS37_NosgRNA going as high as 98±19% [cf. Figure 74 (b)]. Unfortunately, the proportion of colonies which survived FOA selection after being induced at 2.5 mM varied wildly in between each replicate ([cf. Figure 74 (c)], making meaningful comparisons difficult. The mutagenesis efficiency of vFS36_TA_$P_{fdxE}$ and vFS40_nCas9_PfdxE_HDR still hovered around 60 %, but with much greater variation than at 5 mM (going from 6% and 16% of SD at 5 mM of theophylline to 43% and 44% at 2.5 mM).

#### V.5.2.2)(2) Background survival on FOA decreases when theophylline concentration is halved

On the other hand, the levels of background mutagenesis efficiency represented by vFS37_NosgRNA and pMTL83151 has roughly halved from 39±14% at 5 mM to 25±15% at 2.5

mM (paired t-test, n=4, p=0.015) ([cf. Figure 74 (d)]. This could hint at a dose-dependent interference of theophylline with FOA selection, either through a direct increase of background mutagenesis rate or by interfering with the activity of FOA. It is, however, worth mentioning that such an interference was not observed during the preliminary characterisation of Target-AID or the test of alternative expression systems, so further investigation is needed before reaching any conclusion.

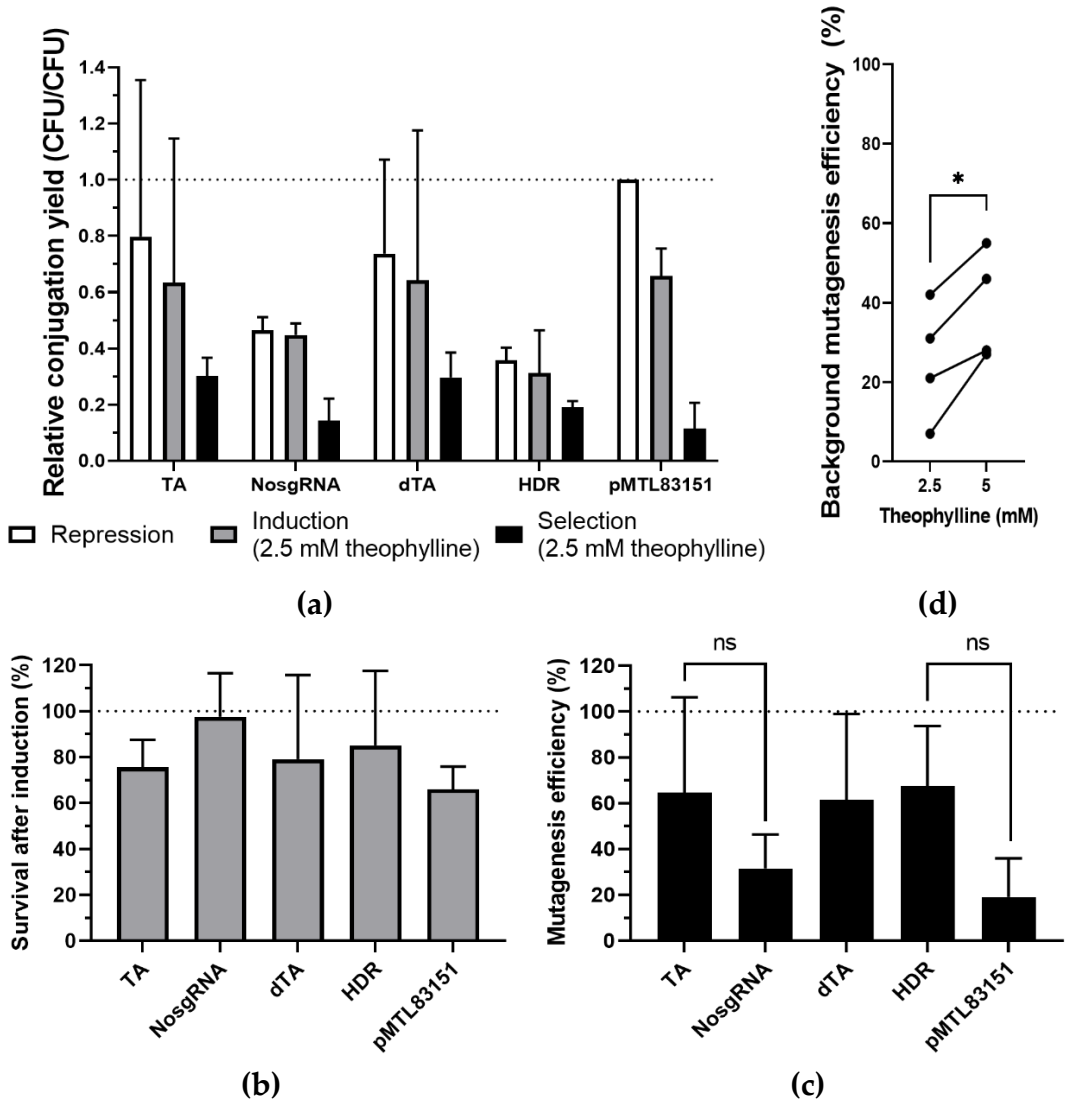
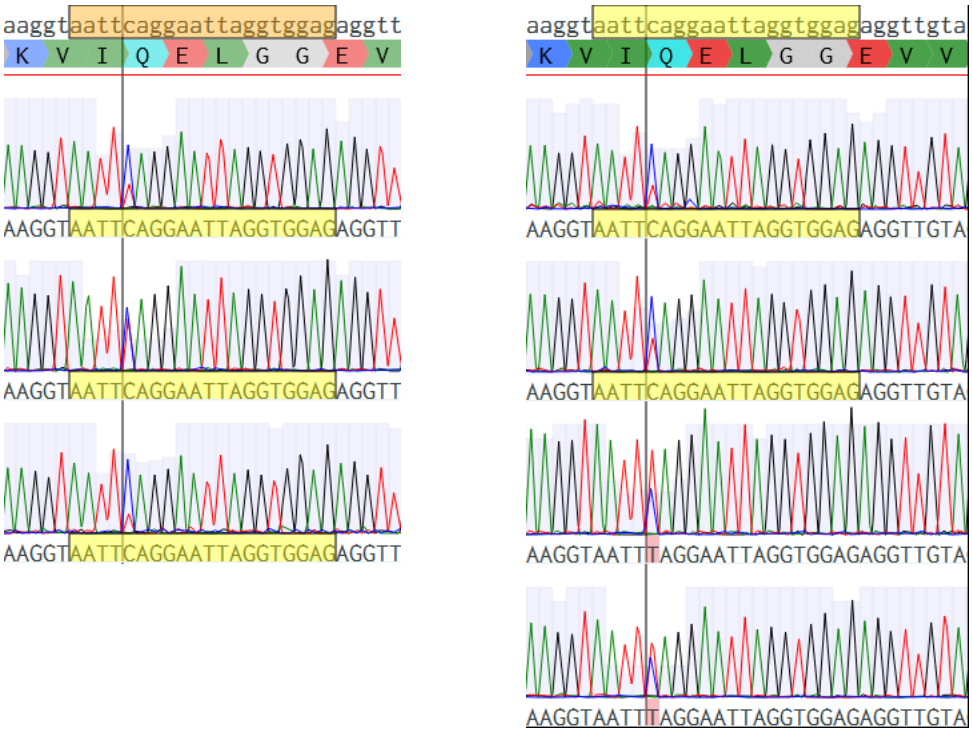



**Figure 74: Characterisation of Target-AID and variant constructs targeting *pyrE* in *C. autoethanogenum* when induced with 2.5 mM of theophylline. (a)** Relative conjugation yield; **(b)** Percentage of survival after induction; **(c)** mutagenesis efficiency; and **(d)** paired comparison of the mutagenesis efficiency of the control constructs vFS37_NosgRNA and pMTL83151 under different theophylline concentrations. TA = vFS36_TA_PfdxE_pyrE, Target-AID vector with sgRNA cassette targeting *pyrE*; NosgRNA= vFS37_NosgRNA, Target-AID vector without sgRNA cassette; dTA=vFS39_dTA, Target-AID vector with a catalytically inactive Cas9 domain, with sgRNA cassette targeting *pyrE*; HDR= vFS40_nCas9_PfdxE_HDR, Cas9-HDR mutagenesis vector with nCas9, a *ΔpyrE* editing cassette and a sgRNA cassette targeting *pyrE*; ns = non-significant; * = p<0.05 (paired t-test, n=4). All constructs were induced with 2.5 mM of theophylline. Then, induced colonies were transferred with a velvet stamp to FOA selection plates with the same concentration of theophylline, to select for mutants. Colony count is reported as a proportion of the CFU of the pMTL83151 control construct to facilitate the agglomeration of replicates (n=2).

### *V.5.2.3) No true mutants occurred before FOA selection*

The sequence of the *pyrE* locus of 8 colonies of each constructs induced on 5 mM theophylline prior to FOA selection were analysed by Sanger sequencing. Once again, inexplicably and in opposition to the observed colony counts and the results of the preliminary experiment, none of the colonies harboured any mutation in the *pyrE* locus.

Once transferred on FOA selection plates, however, three out of the six vFS36_TA_$P_{fdxE}$ colonies sequenced showed clear mixed peaks in the targeted locus [cf. Figure 75(a)]. Two more had mixed peaks which could not be confidently differentiated from noise, and one presented a clean WT peak. Four out of the seven vFS39_dTA colonies also exhibited clearly defined mixed peak at the targeted base [cf. Figure 75(b)]. Mixed peaks should be indicative of mixed colonies, which – according to the literature – is a relatively common occurrence with base editing [318]. However, as described during the preliminary characterisation of Target-AID, this should not be observed under FOA selection since all the WT cells of the colonies should die. No mutation could be detected by Sanger sequencing in the *pyrE* locus of vFS37_NosgRNA and pMTL83151 colonies growing on FOA.




**(a) vFS36_TA_PfdxE** **(b) vFS39_dTA**

**Figure 75: Sanger sequencing traces of the *pyrE* locus of mutated *C. autoethanogenum* colonies conjugated with two different Target-AID constructs induced at 5 mM of theophylline and growing on FOA.** Each row represents the sequence obtained from one individual colony. **(a)** vFS36_TA_$P_{fdxE}$ colonies **(b)** vFS39_dTA colonies. The sequence of the WT protospacer is highlighted in yellow; the corresponding aa sequence is displayed in a series of consecutive arrows; a vertical bar strikes the start of the targeted codon (in the forward direction). Traces and annotations obtained with Benchling.

If these results are to be trusted, mutations only started occurring after FOA selection, and not after induction with theophylline. Even then, mutagenesis was incomplete, yielding mixed peaks.

Lastly, despite my best efforts, the *pyrE* locus of vFS40_nCas9_PfdxE_HDR could not be amplified with either oFS105 & oFS106, oFS313 & oFS314, oFS318 & oFS319, or oFS278 & oFS106; using either Dreamtaq, Onetaq or Q5 polymerase from NEB; or at annealing temperatures in between 50°C and 72°C. Consequently, vFS40_nCas9_PfdxE_HDR colonies induced at 2.5 mM or 5 mM of theophylline could not be genotyped by gel electrophoresis as originally intended.

#### *V.5.2.4) Comparison with preliminary experiment*

Several samples gave conflicting results in between the preliminary and the final experiment. The mutagenesis efficiency of vFS40_nCas9_PfdxE_HDR soared in the final experiment, but this was expected: it was abnormally low during the preliminary experiment. However, this could not be confirmed by PCR amplification of the *pyrE* locus.

More importantly, many more colonies survived FOA selection in the final experiment than in the preliminary experiment. In combination with the mixed peaks obtained during the sequencing of the colonies growing on FOA, this increased survival on FOA raises serious questions about the integrity of the FOA stock used during this final characterisation of Target-AID. Alternatively, it could also be the consequence of mutations in the *pyrF* locus in a large proportion of the *C. autoethanogenum* cells. The colonies of both experiments could be considered biofilms, so hypothesis that biofilms could protect colonies from FOA cannot readily explain the increase in background FOA resistance in the final experiment.

Finally, none of the off-target mutations observed with vFS37_NosgRNA during the preliminary characterisation were observed in the final characterisation, although it might not mean much given the apparent failure of FOA to be selective over the course of the final characterisation.

#### *V.5.2.5) Limitations of the experiment*

A glaring weakness of the experiment was the difficulty in interpreting the sequencing data. The *pyrE* locus of the vFS40_nCas9_PfdxE_HDR colonies could simply not be amplified. The absence of mutants produced by vFS36_TA_$P_{fdxE}$ and vFS39_dTA prior to FOA selection is hard to reconcile with the 100% of mutants identified by Sanger sequencing after FOA selection during the preliminary characterisation, or even with the 100% of mutagenesis

efficiency often obtained without any selection in Chapter VI:. For this reason alone, the best course of action may be to simply discard this whole experiment.

An additional limitation was the fact that only two replicates of the experiment were performed. The two replicates undertaken with induction by 5 mM of theophylline exhibit good agreement with one another, but in contrast the results obtained with 2.5 mM theophylline showed poor agreement, leading to high SD of all measured parameters. More replicates would have helped identify which of the two replicates with 2.5 mM theophylline was the outlier, and determine if induction of Target-AID at 2.5 mM of theophylline would bring any benefit.

A further limitation was the observation that FOA selection appeared ineffective. It did kill many colonies, but many of the surviving colonies did not show any mutation within their *pyrE* gene – which defeats the purpose of the selection. This made the estimation of off-target mutagenesis from Target-AID impossible. Several explanations come to mind: the stock of FOA may have been defective; a spontaneous *pyrF* mutation could have occurred early in the incubation of WT *C. autoethanogenum*; small biofilms may protect colonies from FOA; or theophylline might be interfering with the toxicity of FOA. If one of these hypotheses is correct, the mutagenesis efficiency calculated from the ratio of surviving colonies is meaningless.

Even if FOA selection had been working as intended, transferring cells to a FOA selection plate by using velvet stamps introduces an obvious bias. It is a quick way to screen a large number of colonies, but each colony is composed of thousands of individual cells, which have each a given probability of undergoing independent mutagenesis events in the pyrimidine synthesis pathway which turns FOA into a toxic molecule. As such, this screening method could artificially increase the number FOA-resistant colonies, producing false-positives.

Screening by PCR amplification of colonies growing on induction plate could be more accurate, but it is also much less precise due to the lower number of colonies screened. Alternatively, the need for replica-plating could be dispensed with by inducing Target-AID in liquid culture. This would allow the plating of single colonies on induction plates and FOA-selection at the same time, much like the mating slurry was plated in parallel on the repression and induction plates in the current setup. Such a change would require consequent protocol optimisation and would be less representative of routine mutagenesis experiments, but it would give a more accurate measurement of the mutagenesis efficiency.

Based on the extensive difficulties encountered during these successive characterisation attempts, it may be more appropriate to do away with FOA selection entirely and quantify the mutagenesis efficiency of Target-AID by next-generation sequencing. It is an expensive and complicated method which is even less representative of routine mutagenesis conditions, but it should be exposed to much fewer experimental mishaps.

### V.5.3. Conclusion

In isolation, this last attempt at characterizing Target-AID by knocking-out the *pyrE* locus should invalidate the use of the tool in *C. autoethanogenum*: indeed, no mutants could be detected without selection. Even with selection, barely half of the colonies showed any sign of targeted mutagenesis and no pure mutants could be isolated.

However, these sequencing results not only conflict with what could have been expected from the improved FOA survival rate of the Target-AID constructs, they also conflict with the results of the preliminary characterisation and of the many mutagenesis's carried out without any selection marker in the next chapter. Consequently, it is reasonable to assume that this final characterisation is flawed and should be discarded and repeated – or, preferably, redesigned, for example by making use of next-generation sequencing.

Other questions remain unanswered after this last characterisation attempt. It is still unclear whether inducing Target-AID at 2.5 mM would improve the outcome of Target-AID mutagenesis. It did seem to reduce the rate of false positives surviving FOA selection but the two replicates conflicted regarding its ability to produce true positives. Until better data is collected, induction of Target-AID using 5 mM theophylline remains the preferred option.

It is similarly unclear whether Target-AID performs better than conventional HDR-mediated mutagenesis – the colony count seem to indicate that both tactics are equivalent, but the *pyrE* locus of colonies which were conjugated with the HDR-mutagenesis vector could not be amplified to verify that they carried the intended mutation.

## V.6. Summary and perspectives

In this varied chapter, attempts were made to characterize a new, state-of-the-art genetic tool in *C. autoethanogenum*. Target-AID, a base editor, is thought to induce mutations by hijacking the MMR pathway, when traditional Cas9-mediated mutagenesis exploits the NHEJ and/or the HDR pathway. Target-AID was a candidate to compensate the lack of NHEJ pathway *in C. autoethanogenum* and other *Clostridium* species by providing a similar way to quickly knock out genes with no need for cumbersome homology arms. Additionally, it was hoped that it would perform better than HDR-mediated Cas9 mutagenesis, opening the way to multiplex genome editing.

After designing an in-house bioinformatics tool to design appropriate sgRNAs, it was shown that the removal of the predicted sequence of the endogenous Cau10061II restriction site from the vector did not increase conjugation efficiency in *C. autoethanogenum*.

A preliminary characterisation of Target-AID resulted in hundreds of transconjugants, over 70% of which were the intended mutants when induced with 5 mM of theophylline.

The comparison of a LacI-based repressor, a theophylline-inducible riboswitch, and the constitutive expression of a truncated Target-AID, was largely inconclusive. Nonetheless, it did establish that the LacI-based repressor might be unstable and that the trTarget-AID construct produced a different profile of base editing. Either the base editing window shifted from the 5’ end of the protospacer to the 3’ end, or trCas9 gained the ability to bind PAMs upstream of the protospacer as well as downstream; furthermore, more deletions and non-canonical transversions and transitions were observed with trTarget-AID than conventional Target-AID.

It was also observed that theophylline can be toxic – and, potentially, mutagenic – within the recommended induction range of 5 mM of theophylline.

Finally, two separate characterisations of Target-AID were attempted, sometimes with conflicting results. Many problems were apparent during that final characterisation of Target-AID: most notably, an ineffective FOA selection and confusing sequencing results. Consequently, this characterisation attempt should be repeated several more times or even completely redesigned.

In the meantime, the preliminary characterisation of Target-AID had already given the confidence to move on and use Target-AID to develop a multiplex mutagenesis strategy detailed in the next Chapter.

Base editing is a very active field of research. Several critical developments in different organisms were published while the experiments described here were still ongoing. Some of them, like base editors with an expanded targeting space, will be developed in the next Chapter. Others, like prime editing [336], could not be acted upon, but constitute promising areas of future research.

Prime editing closely resembles base editing, but it can produce any mutation within a single or a few DNA bases (transition and transversion, as well as short deletions or insertions). Instead of fusing Cas9 with a CDA, it fuses it with a reverse-transcriptase and adapts the sgRNA into a so-called pegRNA (prime editing gRNA) with an extended 3'-end encoding the editing template. After nicking the strand, the reverse transcriptase extends the 3'-end of the nick using the pegRNA as a template. Ultimately, the 5'-flap left-over from the original DNA strand, now conflicting with the reverse-transcribed section of the edited strand, is removed by an endogenous exonuclease and the edited strand can be ligated. To facilitate the subsequent mutation of the non-edited strand through endogenous DNA repair pathways (cf. Chapter I.3), an extra sgRNA can be added to target and nick the non-edited strand close to (~50 bp) or even within the edited locus. To the best of my knowledge, this tool has still not yet been tested in prokaryotes and although more complicated to implement than base editors, it holds great potential [337].

# Chapter VI: Multiplex genome editing of *C. autoethanogenum*

## VI.1. Introduction

The ability to edit several genes at once is highly desirable for the metabolic engineering of industrial strains. It drastically cuts down the complexity, duration and cost of a metabolic engineering project, and accelerates the Design-Build-Test-Learn cycle that forms the basis of modern synthetic biology projects [338]. In this chapter, Target-AID is used to test several multiplexing strategies based on CRISPR-Cas9. This has the dual objective of demonstrating the suitability of Target-AID in multiplexing strategies, and finding the best multiplexing strategy for Cas9-mediated mutagenesis in *C. autoethanogenum*.

### VI.1.1. Multiplexing strategies

In this section, some of the most common and relevant multiplexing methods for CRISPR-mediated genome editing are highlighted. For a more comprehensive review, the reader is directed to Adiego-Pérez et al., 2019 [339].

#### *VI.1.1.1) Array of individual sgRNA cassettes*

Perhaps the most obvious way of targeting several protospacers is simply to use an array of sgRNAs expression cassettes [181,340–342]. This method is bulky and demanding in cloning steps, but it is straightforward. The main limitations are the sheer size of so many expression cassettes (each coming with their own promoter, sgRNA, and terminator) and the inconvenience of having to assemble so many parts. Because it is good practice to use different promoters and terminators in the same construct to maximise the stability of the vector (and make homology-based assembly easier) [343], the limited number of characterised promoters and terminators in *Clostridium* is also a significant burden [85].

To make this strategy more convenient, Golden-Gate assembly pipelines have sometimes been set-up, allowing to build an array of as many as seven sgRNA expression cassettes (albeit using the same promoters and terminators) in less than a week [344]. In organisms such as *S. cerevisiae*, where homologous recombination is especially effective, *in vivo* assembly of several sgRNA cassette has also been achieved [345].

#### *VI.1.1.2) Polycistronic sgRNA cassette*

Another strategy to achieve multiplexing still relies on individual sgRNAs, but it expresses them all into a single synthetic operon instead of transcribing them all individually. Each sgRNA is flanked by specific sequences which, when transcribed into RNA, either self-cleave

or constitute a target for endogenous or exogenous endonucleases. This is approach is more complex and must be carefully designed, but it usually results in much shorter vectors than expressing each sgRNA individually, especially for long arrays of sgRNAs.

#### VI.1.1.2)(1) Csy4

The Csy4 multiplexing strategy is inspired by the Type III native CRISPR-Cas systems: Csy4 is an endonuclease used by *Pseudomonas aeruginosa* to process its crRNA array. It specifically recognizes and cleaves the RNA transcript of the 28bp sequence "GTTCACTGCCGTATAGGCAGCTAAGAAA", allowing to cut out any sgRNA flanked by this sequence from the transcript of the sgRNA operon as long as the Csy4 endoribonuclease is also expressed [341,346,347].

#### VI.1.1.2)(2) tRNA array

Multiplexing CRISPR using tRNAs is a widespread strategy in yeast and other eukaryotic cells [348–350]. Indeed, eukaryotic tRNA processing is straightforward (RNAse P and Z cut each extremity of the pre-tRNA before further processing [351–353]), enabling the use of tRNAs as a divider in between two sgRNAs. This way, when the endogenous tRNA-processing machinery of the host cell processes the pre-tRNA into tRNA, it simultaneously frees individual sgRNAs. Eukaryotic tRNAs also double as promoter for RNA polymerase, increasing even more sgRNA transcription levels [354–356]. Lastly, different tRNA sequences can easily be picked, allowing the avoidance of repetitive sequences which can threaten plasmid stability [343].

To the best of my knowledge, no such strategy has been adapted to a prokaryotic system to date. Indeed, the tRNA maturation process is different and less amenable in prokaryotes. The first step of *E. coli* monocystronic tRNA maturation process is thought to be the endonucleolytic cleavage of an AU-rich region at the 3'-end of the pre-tRNA by RNase E. The exact site is thought to be recognized through an ill-defined secondary structure downstream of the mature tRNA sequence, making it difficult to isolate *in silico* [357–360]. The 3' nucleotides in excess are then trimmed down by a cohort of exonucleases. The second step is the endonucleolytic cleavage immediately upstream of the 5'nucleotide of the mature tRNA by RNAse P [361–363]. Recognition of the cutting site is thought to rely on overall tRNA structure, and thus should not be overly dependent on the RNA sequences flanking the pre-tRNA [351,364].

Interestingly, it is unclear whether *C. autoethanogenum* has an RNAse E. Its closest protein is encoded by the gene CLAU_2749 and is about half the size of *E. coli* RNAse E. Instead, this locus might encode a protein closer to *E. coli* RNAse G, a smaller RNAse with a more limited substrate range which might not be involved in tRNA processing [365]. In the absence of RNAse E, *C. autoethanogenum* might thus rely on RNAse Z to cut right after or close to the 3' terminal CCA of the pre-tRNA, as it has been observed in *Thermotoga maritima* [366]. In support for this hypothesis, key features of *C. autoethanogenum* RNAse Z are closer to the RNAse Z of *T. maritima* than *E. coli* (cf. Figure 76).

```
T. maritima             24-74 FDAGEGVSTTL---GSKVYAFKYVFLTHGHVDHIAGLWGVVNIRNNGMGDREK
C. autoethanogenum      34-85 IDCGEGTQVSLKILGCKIKNIDVILFTHFHADHIAGLPGLLLTIA--NSGRNM
B. subtilis             36-87 FDCGEATQHQILHTTIKPRKIEKIFITHLHGDHVYGLPGLLGSRS--FQGGEE
E. coli                 37-88 FDCGEGTQHQLLHTAFNPGKLDKIFISHLHGDHLFGLPGLLCSRS--MSGIIQ
```

**Figure 76: Alignment of a key domain of RNAse Z from selected species with Clustal Omega** [367]**.** Reproduced from Minagwana et. al, 2014 [366] to include *C. autoethanogenum*. Key features associated with cut-site determination and endonuclease activity are highlighted by bold letters (double glutamine) or a red frame (histidine-domain, with similarities in between *T. maritima* and *C. autoethanogenum* bolded).

Assuming that the tRNA maturation process of *C autoethanogenum* does not deviate too far from the *E. coli* or *T. maritima* models, tRNAs could thus also be exploited to separate sgRNAs in *C. autoethanogenum*.

#### VI.1.1.2)(3) Self-cleaving ribozymes

In a different take on the problem, Ryan et al. proposed to flank each sgRNA with self-processing ribozymes, protecting the RNA from degradation by 5' endonuclease in the same move (cf. Figure 77) [368,369]. Gao and Zhao pushed the logic further by flanking each sgRNA with two self-processing ribozymes: the hammerhead (HH) and hepatitis delta virus (HDV) ribozymes (cf. Figure 78) [356,370].

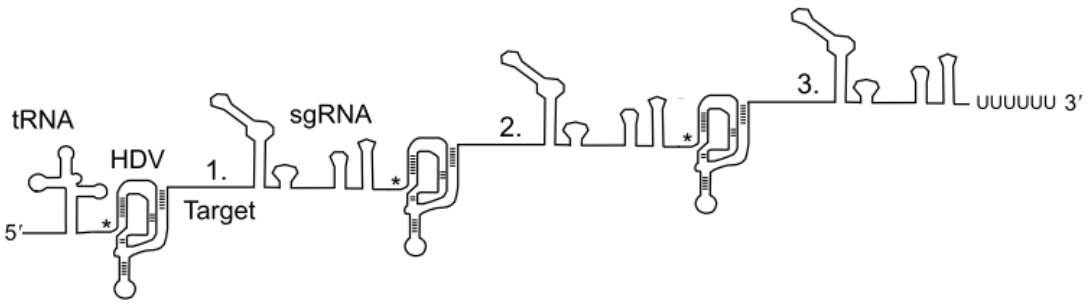



**Figure 77: HDV-sgRNA array proposed to multiplex sgRNA expression by Ryan et al.** Reproduced from Ryan & Cate, 2014 [369].

Applications of this mode of multiplexing, however, are few and far between. To the best of my knowledge, ribozymes have never been used to process a sgRNA array in prokaryotes [371,372].

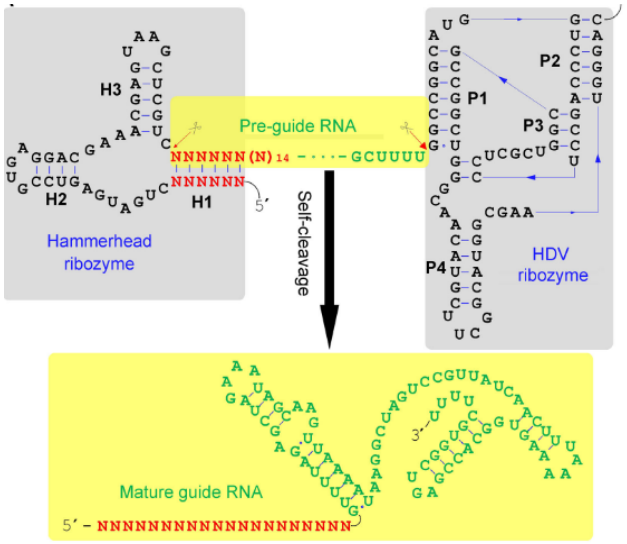



**Figure 78: Ribozyme-Guide-Ribozyme (RGR) model of a modular, self-splicing sgRNA unit.** Adapted from Gao, Y., & Zhao, Y., 2014 [370].

### *VI.1.1.3) CRISPR arrays*

Finally, the last general strategy simply mimics WT CRISPR systems. Indeed, at its core, CRISPR is already a multiplexed genome editing system [197,349,373,374]. The chimeric sgRNA developed by Jinek et al. [179] simplified the genome editing of a single target by reducing the CRISPR-Cas9 to a two-components system (sgRNA and Cas9), but by doing so, it made its multiplexing less straightforward. An obvious way to achieve multiplexing with a CRISPR system is thus to simply replace the sgRNA cassette by a CRISPR array. The following sections describe four kinds of CRISPR arrays which have been shown to achieve multiplexed genome editing.

#### VI.1.1.3)(1) SpCas9

The first, most obvious strategy is to simply reproduce the WT SpCas9 CRISPR system on the shuttle vector. In the SpCas9 CRISPR system, after transcription of the CRISPR array as a single pre-crRNAs molecule, a tracrRNA binds each direct repeat (DR) within the pre-crRNA. This creates a dsRNA substrate for RNAse III [375], which cuts the pre-crRNA into individual crRNAs in the presence of Cas9. Each crRNA-tracrRNA-Cas9 complex is then able to target a different locus in parallel.

RNAse III is ubiquitous in bacteria [376]. In fact, the RNAse III of *C. autoethanogenum* is very similar to the RNAse III of *S. pyogenes* and *E. coli* (86% and 55% of coverage with 51% and

37% of identity, respectively). Consequently, it is sufficient to express a tracrRNA and to flank each spacer sequence with the DR of *S. pyogenes* to reproduce a functional multiplex CRISPR system [182,223,377].

#### VI.1.1.3)(2) Cas12a

As briefly discussed, Cas12a is part of another Type II CRISPR system which has many advantages over Cas9, especially in a *Clostridium* chassis [204–207,378]. Unlike the blunt-ended double-strand break produced by Cas9, the double-strand break induced by Cas12a is staggered (it produces 5' and 3' ssDNA), which should be easier to repair by homologous recombination. It also contains its own RNAse domain which enables it to process pre-crRNA into crRNA completely on its own, without requiring the help of a tracrRNA. This last particularity makes the whole CRISPR genome editing system much shorter and simpler [379–384].

Over the course of this project, Cas12a was successfully applied to targeted mutagenesis in *Clostridium* by a different group [385]. Eventually, it was even tested within a base editor with a catalytically inactive Cas12a (although fused to APOBEC1 rather than PmCDA1) [315]. However, even though its `TTTV` PAM is `AT`-rich, it is also one (and one fourth) base longer. This drastically increases the requirement for a PAM and consequently reduces the number of potential targets in the genome of *C. autoethanogenum*. Taking the 573 bp-long *pyrE* gene as an example, the traditional `NGG` PAM of Cas9 leads to 45 protospacers, while the `TTTV` PAM of Cas12a only yields 16 protospacers.

#### VI.1.1.3)(3) Endogenous CRISPR system

A more subtle strategy for genome editing in prokaryote is to exploit the native CRISPR-Cas system of the host cell, instead of expressing a foreign Cas *in trans* [208]. This necessitates an exhaustive bioinformatics analysis to identify the Cas, DR and PAM associated with each endogenous CRISPR-system [386,387], but it reduces considerably the size and the complexity of the vector used for the targeted mutagenesis. Indeed, all the Cas are already expressed in the host cell. The vector only needs to express a CRISPR array [388]. This method is not compatible with base editing, but a large part of the necessary bioinformatics work had already been published for *C. autoethanogenum* [209,285].

As a side-project, the validity of this bioinformatics analysis was tested by identifying a 36 bp long target protospacer in *C. autoethanogenum* $\Delta pyrE::ACE$ downstream of a `NAA` PAM. A successful HDR mutagenesis would have restored the WT *pyrE* allele. However, desired HDR-

mediated genomic insertion could not be obtained (results not shown), even when the synthetic CRISPR array was placed under the control of the $P_{facoid}$ promoter and the LacI repressor was being expressed constitutively by $P_{thl}$ from *C. acetobutylicum* on the same vector as the synthetic CRISPR array. This seems to indicate that the biofinformatics analysis is not correct, or at best that the endogenous CRISPR system of *C. autoethanogenum* is not suitable for HDR-mediated mutagenesis.

### VI.1.2. Experimental design

In this chapter, the Design-Build-Test-Learn process is followed, whereby after identifying appropriate targets for a multiplex genome editing attempt, three consecutive cycles of multiplex mutagenesis, each instructed by the results of the previous round, were implemented. The initial multiplexing design was tested in the first cycle, refined it in the second cycle, and, in the final cycle, the capabilities of Target-AID itself was greatly expanded in order to finally obtain a triple knock-out. For each mutagenesis, the transconjugants were plated directly onto induction plates, and the screening was done solely through colony PCR and Sanger sequencing.

## VI.2. Target identification

To illustrate the multiplexing abilities of Target-AID, only three genes were targeted at once. According to the preliminary characterisation of Target-AID (cf. Chapter V.3.3), this would still result in about a third of transconjugants exhibiting all three mutations. Such a mutagenesis efficiency should be high enough that triple mutants could be identified with a simple screen by colony PCR. Indeed, in order to illustrate the use of Target-AID in conditions representative of a metabolic engineering experiment, it was decided to target three genes which did not lend themselves to a phenotypic screen and presented some interest to the metabolic engineering community.

After deliberation with Dr. Thomas Millat, it was decided to knock-out three NADPH-dependant alcohol dehydrogenases which he had identified as being overexpressed at different times when the carbon source of *C. autoethanogenum* was progressively being shifted from $CO_2$ to CO in order to test the prediction of our metabolic model (unpublished). The three genes are *CLAU_0532* (*adh1*), *CLAU_0534* (*adh3*) and *CLAU_1794*. According to the genome-wide transposon mutagenesis of *C. autoethanogenum* carried out by Dr. Craig Woods [389], none of them are essential, at least individually, whether the Carbon source is fructose or carbon monoxide. Consequently, it should be possible to knock them out. It was anticipated that such a triple knock-out should severely decrease ethanol production and drastically lower the growth rate of *C. autoethanogenum* under autotrophic growth on CO, when cells shift to NADPH. The change in the concentration of other metabolites (mainly 2,3 butanediol and acetate) would allow Dr. Millat to test the predictions of an in-house metabolic model of *C. autoethanogenum* [390].

Using the developed Target-AID sgRNA design tool (cf. Chapter V.2.1), three protospacers compatible with Target-AID were identified in each gene (Table 17).

**Table 17: Protospacers suitable to introduce a STOP codon in *C. autoethanogenum CLAU_0532, CLAU_0534 and CLAU_1794* after a C→ T mutation within an editing window of -19 to -16 bp from the PAM.** The **position** of the target C is given as a fraction of the whole gene. **Strand** indicates whether the protospacer is on the sense (1) or antisense (-1) strand; **Sequence** is the sequence of the sgRNA target (or protospacer), with the codon to be mutated represented in capital letters; **PAM** is the sequence of the PAM, directly downstream of the protospacer; off-target and on-target scores respectively refer to the specificity and efficiency of the protospacer [cf.Chapters I.5.2.1) and I.5.2.2)] – closer to 100 is better.

| Gene | Label | Sequence | PAM | Strand | Off-target score (%) | On-target score (%) | Position (bp) |
|---|---|---|---|---|---|---|---|
| *CLAU_0532* | A | ctCCAgtcag gtgttgtgca | tgg | -1 | 50 | 9 | 272/1056 |
| | B | tgttCAAtgg ggctgcggca | tgg | 1 | 50 | 32 | 838/1056 |
| | C | ccCCAttgaa cacgaggtat | tgg | -1 | 50 | 13 | 845/1056 |
| *CLAU_0534* | A | agccCAAtgt ctagctggga | tgg | 1 | 50 | 4 | 727/1167 |
| *CLAU_1794* | A | aaaCAAgcaa ttgttccgtt | tgg | 1 | 50 | 17 | 370/1194 |
| | B | atcaCAAtgt ttagcaggta | tgg | 1 | 50 | 1.5 | 754/1194 |

Whenever possible, the protospacer the most upstream was selected to maximize the impact of a potential nonsense mutation within the protospacer sequence. This resulted in the protospacers CLAU532A, CLAU534A and CLAU1794A.

## VI.3. First round of multiplex genome editing with Target-AID

Because it was the most scalable and elegant solution, it was decided to multiplex Target-AID with a *S. pyogenes* CRISPR array. An array of individual sgRNAs expression cassettes was also assembled as a backup. This was chosen on the basis of it not relying on any additional molecular machinery. Lastly, in order to explore a more innovative multiplexing method, it was opted to investigate the potential of tRNAs to separate a polycistronic sgRNA array.

### VI.3.1. Design of initial multiplexing constructs

#### *VI.3.1.1) sgRNAs array*

In order to multiplex the vFS36_TA_PfdxE_pyrE Target-AID vector with additional sgRNA expression cassettes, only two additional strong promoters and terminators were needed. The selected promoters were from *Clostridium perfringens* ($P_{cpthl}$ and $P_{cpfdx}$) (cf. Chapter VIII.2.6.1) [246]. Because the sgRNAs do not need to be translated, no RBS were added downstream of the promoters. Notably, $P_{araE}$ still contains its original RBS. It was not removed, since $P_{araE}$ had already demonstrated its ability to express sgRNAs successfully. For the terminators, the *E. coli rrnB* terminator T1 loop (t*ecT1*) and the *B. subtilis tyrS* tRNA terminator (t*ectyrS*) [244] were selected. The sequence of all promoters and terminators can be found in the supplementary materials (Chapter VIII.2.6). The resulting array of sgRNA expression cassettes – illustrated in Figure 79 – replaced the sgRNA cassette of vFS36_TA_PfdxE_pyrE to form the plasmid vFS50_TA_msgRNA.

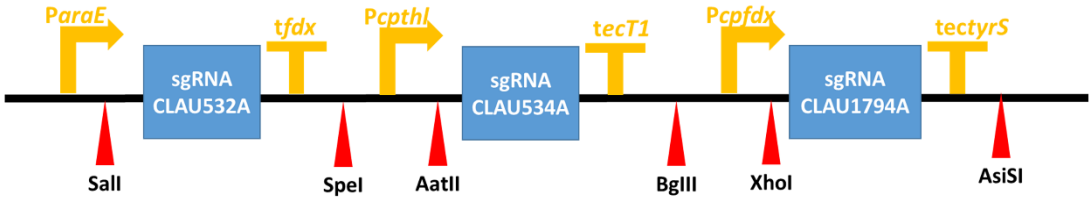



**Figure 79: Schematic of the sgRNA array of vFS50_TA_msgRNA.**

### *VI.3.1.2) SpCRISPR array*

According to the literature, the native CRISPR system from *S. pyogenes* uses spacers of ~30 bp, of which only 20 bp are actually needed to target the protospacer sequence [179,194,197,377,391]. Accordingly, a random stretch of 4 bp and a 6 bp restriction site were added upstream of each 20 bp protospacer sequence to reconstitute a whole 30 bp spacer. Each spacer is flanked by the DR of SpCas9 CRISPR locus and a tracrRNA expression cassette is added to the Target-AID vector (cf. Table S. 4) – once again, under the control of $P_{cpthl}$ (without RBS) and $t_{ecT1}$. The resulting SpCRISPR array – illustrated in Figure 80 – replaced the sgRNA cassette of vFS36_TA_PfdxE_pyrE to form the plasmid vFS51_TA_mCRISPR.

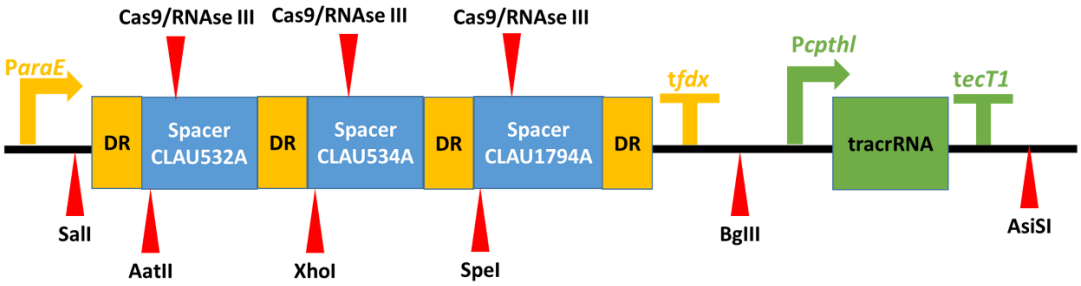
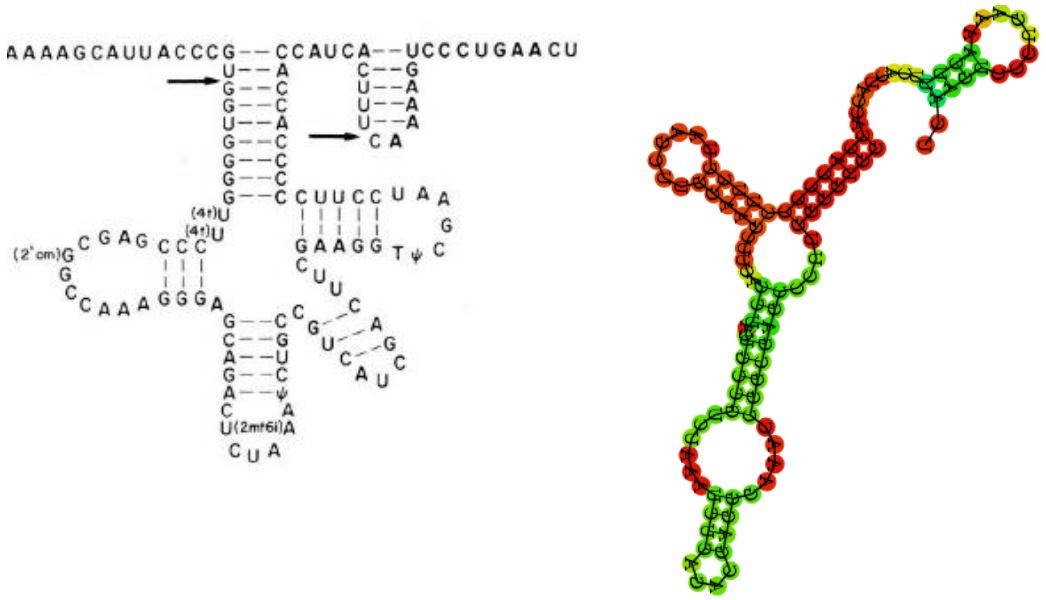



**Figure 80: Schematic of the SpCRISPR array of vFS51_TA_mCRISPR.** DR= Direct repeat; tracrRNA= trans-activating CRISPR RNA.

### *VI.3.1.3) tRNA-sgRNA array*

In an attempt to reduce cross-talks and increase the stability of the multiplex tRNA construct in *C. autoethanogenum*, tRNAs from a different *Clostridium* species, *C. pasteurianum*, were selected. Without experimental data about tRNA maturation in *C. autoethanogenum* to guide the design, it was decided to carefully select pre-tRNAs which shared structural features with a pre-tRNA whose maturation had been well described in the literature: the $tRNA^{Tyr}$ from *E. coli* described by Sekiya et al. (cf. Figure 81) [357]. The sequence of each tRNA was extracted and complemented with ~20 bp of nucleotides immediately downstream from their mature tRNA sequence using the GtRNAdb tRNA database [392,393]. The resulting RNA structures were predicted using the online RNA structure prediction tool RNAfold [394].



(a) (b)

**Figure 81: Comparison between (a) the structures of the pre-$tRNA^{Tyr}$ of *E coli* as described by Sekiya et al., and (b) the structure of the same pre-tRNA as modelled by RNAfold.** Particular attention is paid to the 3' end of the tRNA, as RNAse P is supposed to process the 5'-end solely based on the tRNA structure, while the processing of the 3' might be more dependent on the pre-tRNA structure. Modified from Sekiya et al., 1979 [357].

A larger set of qualitative criteria were defined to select the tRNAs which would be processed in the same manner as *E. coli* $tRNA^{Tyr}$, without interfering with cloning (cf. Table 18).

**Table 18: Criteria used to select an appropriate pre-tRNA to separate the sgRNA polycistron.**

| Criterion | Priority | Explanation | Design tool |
|---|---|---|---|
| Weak A-U rich stem-loop (~16 bp from CCA-3') | High | Emulation of $tRNA^{Tyr}$ processing from *E. coli*. | RNAfold |
| Absence of 3' poly-U tail. | High | Avoid accidental ρ-independent terminator. | RNAfold |
| Compatibility with sgRNA | High | Pre-tRNA cannot form strong secondary structure with sgRNA design primers | mfold |
| High codon usage in host. | Low | Avoid disturbing the host's tRNA relative abundance. | CUSP |
| Low sequence homology with native host genome. | Low | Avoid homologous recombination between construct and host genome. | BLAST |

The full list of the tRNAs which have been considered can be found in Table S. 3 with their predicted structure and the associated decision matrix.

Two tRNAs (tRNA-Thr-TGT-1-1 and tRNA-fMet-CAT-1-1) were eventually selected to separate the three sgRNAs under the control of a single promoter (cf. Table S. 5). The resulting tRNA-sgRNA array (cf. Figure 82) replaced the sgRNA cassette of vFS36_TA_$P_{fdxE}$, to create a new vector called vFS48_TA_mtRNA.

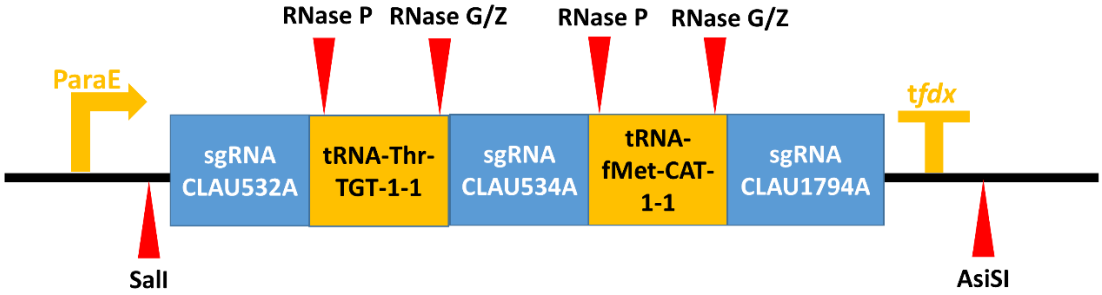



**Figure 82: Schematic of the sgRNA array and associated tRNAs in vFS48_TA_mtRNA.**

### *VI.3.1.4) Monoplex controls*

In addition to the multiplex constructs, three standard Target-AID vectors (vFS52_TA_CA532A, vFS53_TA_CA534A, and vFS54_TA_CA1794A) – each with a single sgRNA expression cassette under the control of $P_{araE}$ and $t_{fdx}$ – were assembled to knock-out each gene separately. These controls ensure that the mutagenesis potential of each protospacer is assessed independently from each multiplexing strategy.

### VI.3.2. Multiplex KO of *CLAU_0532, CLAU_0534, and CLAU_1794*

All three multiplex Target-AID constructs (vFS50_TA_msgRNA, vFS51_TA_mCRISPR, and vFS48_TA_mtRNA), as well as all three individual sgRNA control constructs (vFS52_TA_CA532A, vFS53_TA_CA534A, and vFS54_TA_CA1794A) and pMTL83151 as the positive control were conjugated into WT *C. autoethanogenum* and directly plated onto YTF plates complemented with appropriate antibiotics and 5 mM of theophylline to induce each construct. Because a much lower survival rate in constructs which target several genomic loci was expected, 150 µL of a 600 µL mating slurry were plated for each construct. The resulting colony count is illustrated in Figure 84. Once grown, eight colonies of each construct besides pMTL83151 were patched, amplified with the appropriate primers (oFS173-174 for CLAU_0532, oFS175-176 for CLAU_0534 and oFS177-178 for CLAU_1794), and the amplicons from five colonies were sequenced by Sanger sequencing. The sequencing results are summarized in Table 19.

#### *VI.3.2.1) CLAU1794A is not an appropriate protospacer*

The sequencing of the CLAU1794 protospacer in all colonies of all constructs (including the vFS54_TA_CA1794A control which only targets this particular protospacer) failed to find any mutated allele (cf. Table 19). Consequently, the failure of my multiplex constructs to target this protospacers is probably due to the protospacer itself, rather than any multiplexing strategy.

#### *VI.3.2.2) Most mutated bases are not pure*

The other two control constructs successfully mutated their protospacer: vFS52_TA_CA532A yielded five mutant colonies out of five, and vFS53_TA_CA534A three out of five. Unfortunately, all but one of these mutant colonies exhibited mixed bases in the target locus (cf. Figure 83). This is indicative of mixed colonies: all the cells in the colony do not share the same mutation, and some of them might even be WT. Similar observations can be made for all constructs: of all twenty-three mutated loci, only two had a pure mutated base. This might indicate that WT cells are still able to divide, or already host several copies of their chromosome, when Target-AID is expressed. This would allow the daughter cells to mutate differently or even escape mutagenesis altogether.

**Table 19: Sequencing of five *C. autoethanogenum* colonies conjugated with different Target-AID constructs and induced on theophylline. vFS48=vFS48_TA_mtRNA; vFS50=vFS50_TA_msgRNA; vFS51=vFS51_TA_mCRISPR; vFS52=vFS52_TA_CA532A; vFS53=vFS53_TA_CA534A; and vFS54=vFS54_TA_CA1794A.** The targeted codon is highlighted in green, and mutated bases are capitalized and bolded. Fractions indicate mixed reads for this base.

| Protospacer | Allele | Sequence | Mutagenesis efficiency | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Multiplex | | | Controls | | |
| | | | vFS 48 | vFS 50 | vFS 51 | vFS 52 | vFS 53 | vFS 54 |
| CLAU 532A | WT | ctccagtcaggtgttgtgca | 0 | 1/5 | 4/5 | 0 | | |
| | A.1 | ct**TT**agtcaggtgttgtgca | 0 | 1/5 | 0 | 0 | | |
| | A.2 | ct**T T / c c**agtcaggtgttgtgca | 0 | 3/5 | 0 | 5/5 | | |
| | A.3 | ct**T / c**cagtcaggtgttgtgca | 5/5 | 0 | 1/5 | 0 | | |
| CLAU 534A | WT | agcccaatgtctagctggga | 5/5 | 0 | 5/5 | | 2/5 | |
| | A.1 | ag**T**ccaatgtctagctggga | 0 | 0 | 0 | | 1/5 | |
| | A.2 | ag**T / c**caatgtctagctggga | 0 | 0 | 0 | | 2/5 | |
| | A.3 | ag**T T T / c c c**aatgtctagctggga | 0 | 5/5 | 0 | | 0 | |
| CLAU 1794A | WT | aaacaagcaattgttccgtt | 5/5 | 5/5 | 5/5 | | | 5/5 |
| | A.1 | aaa**T**aagcaattgttccgtt | 0 | 0 | 0 | | | 0 |

### *VI.3.2.3) The msgRNA array was the only successful multiplexing strategy*

vFS50_TA_msgRNA was the only construct which yielded colonies with more than one mutation. Indeed, four out of five colonies had mutated bases both in CLAU532A and CLAU534A. The remaining colony did not show any mutated base in CLAU532A. Lastly, one of the mutated colonies (col.2) had a pure mutated CLAU532 locus which produced a STOP codon (TAA). On the other hand, only one of the three mixed bases of CLAU534A would lead to a STOP codon.

While vFS48_TA_mtRNA was about as effective as vFS50_TA_msgRNA in mutating CLAU532A, it completely failed to target CLAU534A. Because the sequence upstream of a sgRNA should have little impact on its effectiveness [391], but the 3' extremity of the sgRNA might interact with Cas9 [395], it was hypothesized that the first tRNA of the tRNA array ((tRNA-Thr-TGT-1-1) might be processed more effectively than the second tRNA (tRNA-fMet-CAT-1-1). This would have resulted in a poorly matured CLAU534A sgRNA, leading to a poor mutagenesis efficiency.

Interestingly, vFS51_TA_mCRISPR yielded only a single (mixed) mutant allele out of the 10 it could have been expected to edit. Tentatively, residual RNA (i.e., promoter sequence downstream of the Transcription start site, or leftover terminator sequence) from the

expression cassette of the tracrRNA could have affected its activity and hindered pre-crRNA maturation by RNAse III or interactions with Cas9.

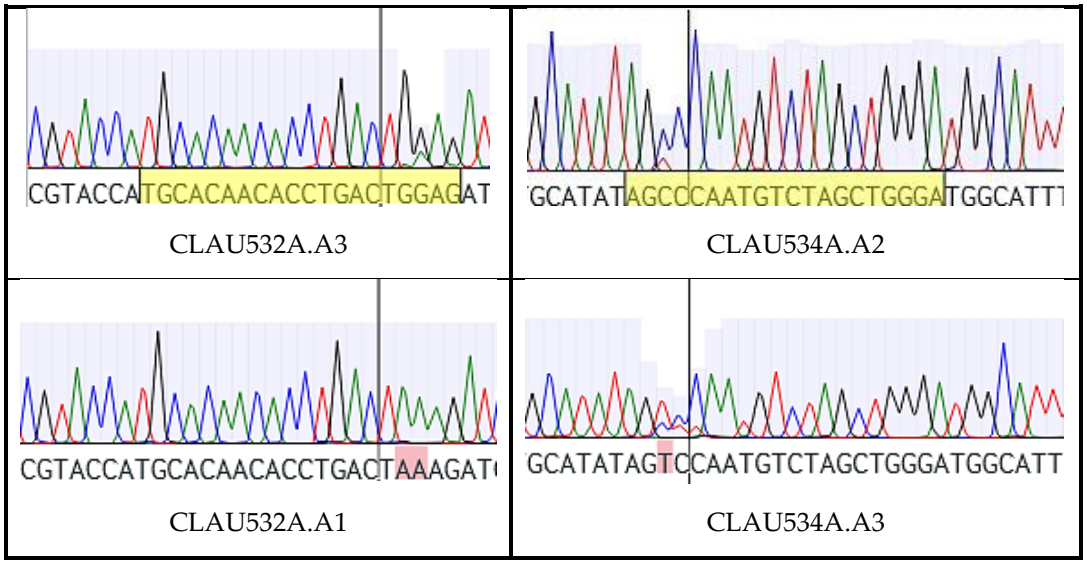



**Figure 83: Representative Sanger sequencing traces of different alleles of the protospacers CLAU532 and CLAU534.** The sequence of the WT protospacer is highlighted in yellow; a vertical bar strikes the start of the targeted codon (in the forward direction); mismatches clear enough to be detected automatically are highlighted in red. Traces and annotations obtained with Benchling.

#### *VI.3.2.4) One streaking step is sufficient to obtain pure colonies*

In an attempt to isolate a strain with a pure nonsense mutation both in CLAU_0532 and CLAU_0534, col.2 of vFS50_TA_msgRNA was carefully re-streaked on theophylline induction plates. During the second round of sequencing, three colonies were shown to have only the first two most upstream cytosines mutated, while one colony (col.2.4) had all three cytosine of the editing window mutated. This resulted in the expected STOP codon in CLAU534A. The last colony could not be sequenced. After one round of re-streak, all colonies screened were thus pure mutants. The colony with a nonsense mutation in both CLAU0532 and CLAU0534 was re-streaked on YTF plates without any antibiotics in order to lose the vFS50_TA_msgRNA vector. Once the vector was lost, that strain was conserved in a cryostock under the label cFS04_CA_Δ532Δ534.

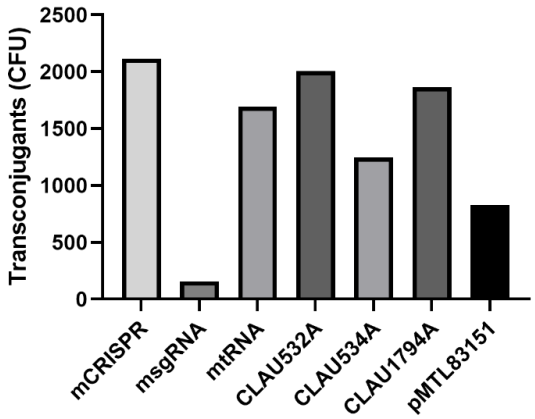



**Figure 84: Number of *C. autoethanogenum* transconjugant colonies of different Target-AID constructs.** All constructs were plated on YTF with antibiotics and 5 mM of theophylline. mtRNA = vFS48_TA_mtRNA; msgRNA = vFS50_TA_msgRNA; mCRISPR = vFS51_TA_mCRISPR; CLAU532 = vFS52_TA_CA532A; CLAU534 = vFS53_TA_CA534A; and CLAU1794 = vFS54_TA_CA1794A.

#### *VI.3.2.5) msgRNA produces the fewest transconjugants*

Strikingly, the only construct which produced less transconjugants than pMTL83151 (828 CFU) was vFS50_TA_msgRNA (157 CFU) (cf. Figure 84). All other constructs yielded from 1244 CFU (vFS53_TA_CA534A) to 2006 CFU (vFS51_TA_mCRISPR). Assuming that all construct would have had the same number of colonies before induction, these colony counts after induction hint once again at an important toxicity of theophylline on the pMTL83151 colonies. Somehow, such toxicity was not observed in the other constructs.

At first approximation, if an average of 2000 CFU is assumed for the other non-mutagenic constructs (vFS51_TA_mCRISPR and vFS54_TA_CA179A), vFS50_TA_msgRNA obtained a percentage of survival after induction of only 8% for the mutagenesis of only two genes. That is equivalent to a survival after induction of 27% per gene; about half of the 60-70% that was measured when only *pyrE* was targeted. If confirmed, this result would indicate that survival after induction worsens significantly for each additional gene being targeted.

### VI.3.3. Conclusion and perspectives

In summary, after the first round of multiplex mutagenesis, only one effective multiplexing strategy (the array of individual sgRNAs) was identified which resulted in the isolation of a *C. autoethanogenum adh1(Q243X), adh3(W90X)* strain. The protospacer used to target CLAU1794 seemed to be dysfunctional, and a different one should thus be used in the next attempt.

The tRNA and CRISPR arrays failed to edit several protospacers at once – the CRISPR array was not capable of targeting even a single gene with significant efficiency. These two constructs might be improved by using twice the tRNA associated with the only protospacer which could be targeted by vFS48_TA_mtRNA; and by removing all superfluous RNA bases from the tracrRNA, respectively.

## VI.4. Second round of multiplex genome editing with Target-AID

The purpose of this second attempt is to mutate CLAU_1794 and troubleshoot the mCRISPR and mtRNA systems. Additionally, a *C. autoethanogenum* strain which had a truncated *pyrE* locus (*ΔpyrE::ACE*) that had been designed by Dr. Christopher Humphreys to facilitate a subsequent complementation using ACE was employed. As explained in Chapter I.4.3, ACE allows to complement large stretches of DNA next to the *pyrE* locus, but it requires a truncated *pyrE* locus. At last, in an attempt to reduce the number of mixed colonies, it was decided to add one step of streaking of the transconjugant colonies on appropriate medium in order to isolate pure colonies before patching them and sending them to sequencing.

### VI.4.1. Re-design of multiplexing constructs

To circumvent the failure of the CLAU1794A protospacer, it was replaced in all constructs by the CLAU1794B protospacer.

#### *VI.4.1.1) sgRNAs array*

No other modifications were made. vFS50_TA_msgRNA was replaced by vFS57_TA_msgRNA_CA1794B (cf. Figure 85).

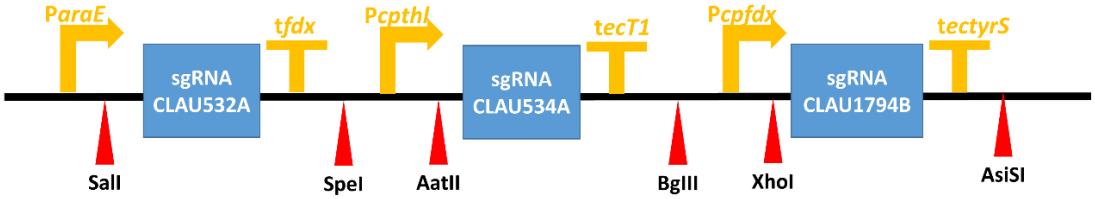



**Figure 85: Schematic of the sgRNA array of vFS57_TA_msgRNA_CA1794B.**

### *VI.4.1.2) SpCRISPR array*

In an attempt to remove a potential interference from extra RNA bases transcribed upstream and downstream of the tracrRNA, it was flanked by the self-cleaving HH and HDV ribozymes [cf. Chapter VI.1.1.2)(3) and Table S. 4]. This led to the assembly of the vector vFS60_TA_mCRISPR_HHDV (cf. Figure 86).

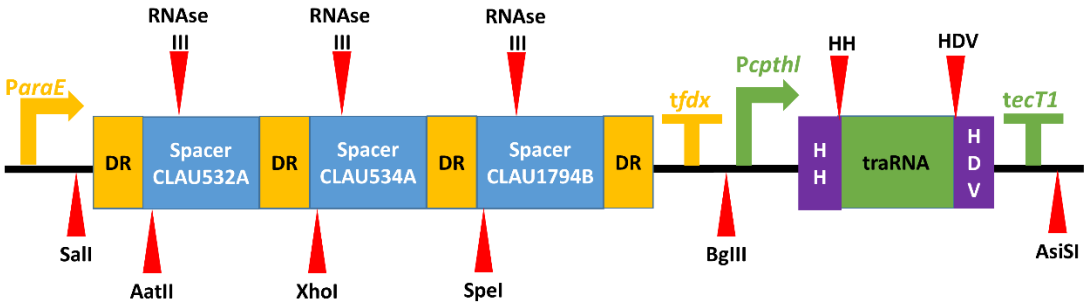



**Figure 86: Schematic of the SpCRISPR array of vFS60_TA_mCRISPR_HHDV.** DR= Direct repeat; tracrRNA= trans-activating CRISPR RNA; HH= hammerhead ribozyme; HDV= hepatitis delta virus ribozyme.

### *VI.4.1.3) tRNAs array*

Because the first sgRNA of vFS48_TA_mtRNA was successful in the previous experiment, its associated tRNA (tRNA-Thr-TGT-1-1) was also used to separate the sgRNAs targeting CLAU534A and CLAU1794B. It resulted in the new vector vFS56_TA_mTHRtRNA (cf. Figure 87).

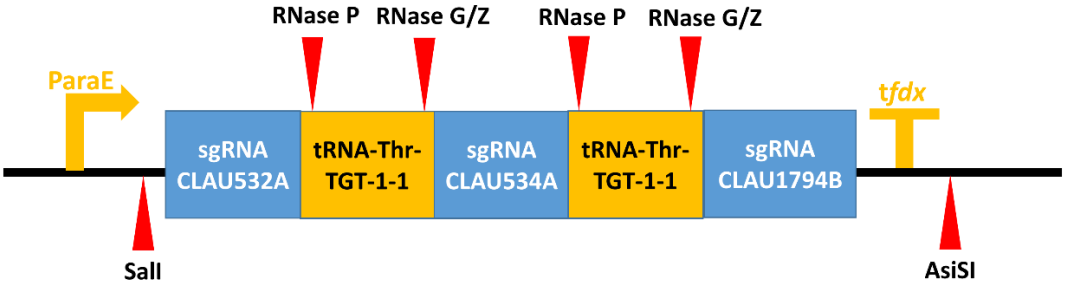



**Figure 87: Schematic of the sgRNA array and associated tRNAs in vFS56_TA_mTHRtRNA_1794B.**

### *VI.4.1.1) Monoplex controls*

Since CLAU532A and CLAU534A had been shown to be targetable, vFS52_TA_CA532A and vFS53_TA_CA534A were not needed anymore. vFS54_TA_CA1794A, on the other hand, had to be modified to target the new protospacer and was replaced by vFS58_TA_CA1794B.

### VI.4.2. Multiplex KO of *CLAU_0532, CLAU_0534, and CLAU_1794*

All three new multiplexed vectors vFS56_TA_mTHRtRNA, vFS57_TA_msgRNA_CA1794B, vFS60_TA_mCRISPR_HHDV, only one sgRNA control vector (vFS58_TA_CA1794B) and pMTL83151 were conjugated into *C. autoethanogenum ΔpyrE::ACE.* This time, it was decided to measure the colony count more accurately; consequently, 3x50 µL of each 400 µL mating mix was plated on induction plates (with 5 mM theophylline) with the appropriate antibiotics to select transconjugants. Eight colonies of each construct were streaked onto YTF medium with antibiotics and theophylline in an attempt to isolate pure mutants. One single colony from each resulting streak was then patched onto the same medium to accumulate biomass. Five colonies from each construct were finally sent to sequencing (cf. Table 20).

**Table 20: Sequencing of five *C. autoethanogenum* colonies conjugated with different Target-AID constructs and induced on theophylline. vFS56= vFS56_TA_mTHRtRNA ; vFS57= vFS57_TA_msgRNA_CA1794B; vFS60= vFS60_TA_mCRISPR_HHDV; and vFS58= vFS58_TA_CA1794B.** The targeted codon is highlighted in green, and mutated bases are capitalized and bolded. Fractions indicate mixed reads for this base.

| | | | Mutagenesis efficiency | | | |
|---|---|---|---|---|---|---|
| | | | Multiplex | | | Control |
| Protospacer | Allele | Sequence | vFS56 | vFS57 | vFS60 | vFS58 |
| CLAU 532A | WT | ct**cca**gtcaggtgttgtgca | 0 | 0 | 0 | |
| | A.1 | ct**TTa**gtcaggtgttgtgca | 0 | 1/5 | 0 | |
| | A.3 | ct**T/c**-cagtcaggtgttgtgca | 3/5 | 3/5 | 3/5 | |
| | A.4 | ct**T**cagtcaggtgttgtgca | 2/5 | 0 | 0 | |
| | A.5 | ct**T**-**T/c**-agtcaggtgttgtgca | 0 | 1/5 | 0 | |
| | A.6 | **T**t**TT**agtcaggtgttgtgca | 0 | 0 | 1/5 | |
| | A.7 | **T/c**-t-**T/c**-cagtcaggtgttgtgca | 0 | 0 | 1/5 | |
| CLAU 534A | WT | agcc**caa**tgtctagctggga | 3/5 | 0 | 5/5 | |
| | A.1 | ag**T**c**caa**tgtctagctggga | 0 | 1/5 | 0 | |
| | A.2 | ag-**T/c**-c**caa**tgtctagctggga | 2/5 | 2/5 | 0 | |
| | A.4 | ag-**T/c**-**T/c**-**caa**tgtctagctggga | 0 | 1/5 | 0 | |
| | A.5 | ag**TT**caatgtctagctggga | 0 | 1/5 | 0 | |
| CLAU 1794B | WT | atca**caa**tgtttagcaggta | 3/5 | 3/5 | 0 | 1/5 |
| | A.1 | at**T**a**caa**tgtttagcaggta | 0 | 1/5 | 1/5 | 2/5 |
| | A.2 | at-**T/c**-a**caa**tgtttagcaggta | 2/5 | 1/5 | 0 | 2/5 |

#### *VI.4.2.1) Streaking transconjugants before patching did not suffice to isolate pure mutants*

Even after one round of streaking, 21 out of 50 reads (or ~40% of reads) still exhibited some mixed bases. This points towards an important temporal dimension to Target-AID mutagenesis: mutations occur late, even after five days of induction. It also implies that Target-AID does not inhibit the growth of WT cells while it is being expressed. Increasing the time before screening the cells or going through several streaking rounds could thus be a potential venue to increase the mutagenesis efficiency of Target-AID.

#### *VI.4.2.2) CLAU1794B is a more effective protospacer, but produced no knock-out*

80% of the colonies conjugated with vFS58_TA_CA1794B exhibited a C→T mutation within the editing window (cf. Figure 88). However, the targeted cytosine at -16 bp from the PAM failed to mutate; instead, it was systematically the cytosine at -18 bp from the PAM which was edited. The same phenomenon was observed in CLAU532A and CLAU534A, and is in agreement with literature: the optimal mutagenesis efficiency is observed around the base -18 from the PAM. Tentatively, the mutagenesis efficiency at a C in position -16 would be higher in the absence of an upstream C. As a result, no triple KO colony could be isolated in this round of mutagenesis either.

The mutagenesis efficiency of CLAU1794B was lower in the multiplex vectors (from 20% to 40%), but CLAU1794B could be targeted by all multiplex vectors. This confirms that the failure to mutate CLAU1794A in the previous attempt was due to CLAU1794A itself, and not due to deficient multiplexing systems.

#### *VI.4.2.3) vFS56_TA_mTHRtRNA is more effective than vFS48_TA_mtRNA*

The multiplex tRNA construct, vFS56_TA_mTHRtRNA, performed better than the previous construct, vFS48_TA_mtRNA. Not only did it manage to mutate 40% of the CLAU534A loci screened when vFS48_TA_mtRNA failed to mutate any, it also mutated 40% of the CLAU1794B protospacers. This is as much as vFS57_TA_msgRNA_CA1794B, could achieve on the same target. Because the mutagenesis efficiency of vFS57_TA_msgRNA_CA1794B is still higher when targeting CLAU534 (100%), however, vFS57_TA_msgRNA_CA1794B remains the best multiplexing construct. Nonetheless, it would appear that tRNA machinery has been successfully hijacked to process a synthetic CRISPR array for the first time in prokaryotes.

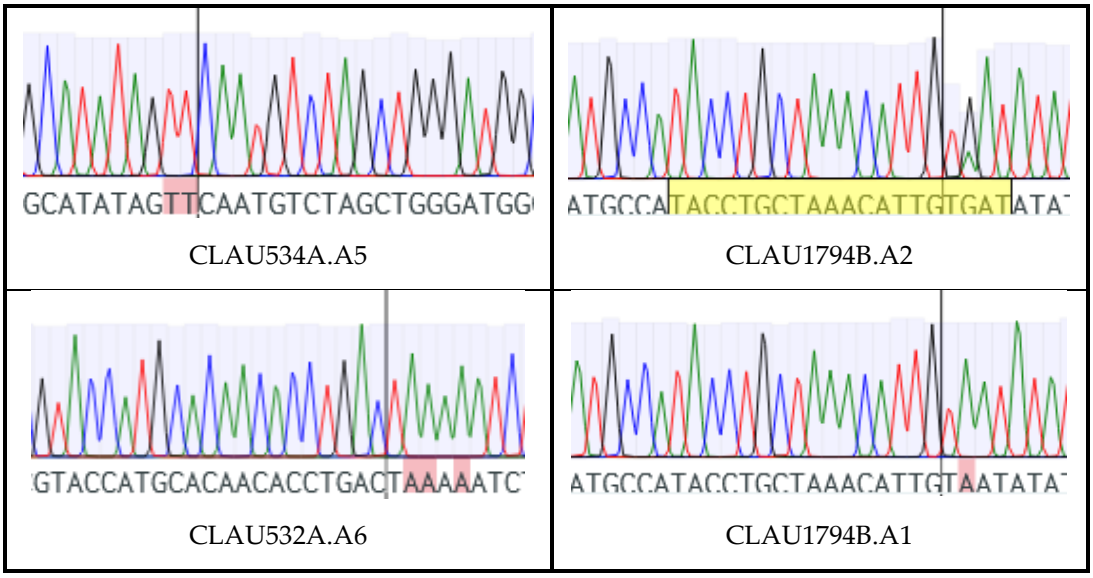



**Figure 88: Representative Sanger sequencing traces of different alleles of the protospacers CLAU532A, CLAU534A and CLAU1794B.** The sequence of the WT protospacer is highlighted in yellow; a vertical bar strikes the start of the targeted codon (in the forward direction for CLAU532A and CLAU534A, but reverse complementary direction for CLAU1794B); mismatches clear enough to be detected automatically are highlighted in red. Traces and annotations obtained with Benchling.

#### *VI.4.2.4) vFS60_TA_mCRISPR_HHDV is more effective than vFS51_TA_mCRISPR, but is still the least effective strategy*

vFS51_TA_mCRISPR only managed to partially mutate one single CLAU532A protospacer; in contrast, vFS60_TA_mCRISPR_HHDV successfully mutated all five of the CLAU532A protospacers screened, and even one CLAU1794B protospacer. Unfortunately, none of the CLAU534 protospacers were successfully targeted by this construct. Flanking the tracrRNA with HH and HDV ribozymes thus seemed to modestly improve the functionality of the CRISPR array, but the system must be enhanced further before it can be used for multiplex mutagenesis.

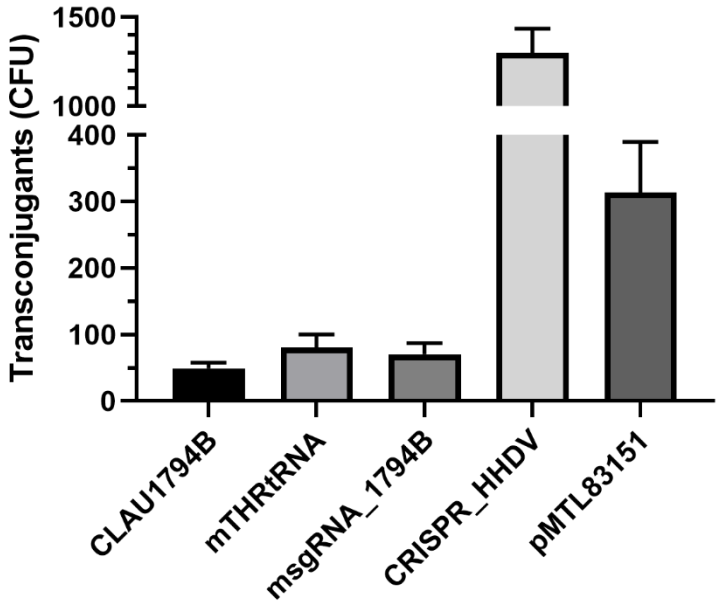



**Figure 89: Number of *C. autoethanogenum ΔpyrE::ACE* transconjugant colonies of different Target-AID constructs.** All constructs were plated on YTF with antibiotics and 5 mM of theophylline. mTHRtRNA = vFS56_TA_mTHRtRNA; msgRNA_1794B = vFS57_TA_msgRNA_CA1794B; CRISPR_HHDV = vFS60_TA_mCRISPR_HHDV; and CLAU1794B=vFS58_TA_CA1794B.

#### *VI.4.2.5) The number of transconjugants is not coherent across constructs*

Surprisingly, all constructs had a very low conjugation yield on induction plates (in between 49±9 CFU for vFS58_TA_CA1794B and 80±20 CFU for vFS56_TA_mTHRtRNA) except for vFS60_TA_mCRISPR_HHDV, which reached 1302±133 CFU when even pMTL83151 only reached 314±76 CFU (cf. Figure 89). On top of this aberrant disproportion in colony count, all multiplex constructs reached a higher number of transconjugants than vFS58_TA_CA1794B, which only targets a single gene. The opposite would have been expected, and was observed in the previous attempt with the other constructs which only targeted one gene.

### VI.4.1. Conclusion and perspectives

The main achievement of this experiment is to have successfully demonstrated the use of tRNAs as sgRNA processing tools for the first time in prokaryotes. Further studies are needed to find even better suited tRNAs and to find direct evidence that the synthetic CRISPR array was indeed fragmented as a consequence of tRNA maturation, and not some other coincidental mechanism.

Similarly, the effectiveness of the deployed SpCRISPR array was successfully improved by flanking the tracrRNA with HH and HDV ribozymes. However, the improvement was modest and further efforts are needed to make it an operational multiplexing strategy. After carefully comparing the SpCRISPR array used here with published arrays which have been exploited to multiplex Cas9 mutagenesis in prokaryotes [182,223,377], the conclusion was reached that integrating a 6 bp-restriction site and four random bases at the 5'-end of the spacer might have had deleterious consequences. In theory, these 10 extra base-pairs are supposed to be trimmed away during pre-crRNA processing, and thus their sequence should not matter. In practice, leftover bases might create mismatches with the extended protospacer sequence, or the 6-bp restriction site might create a secondary structure which disturbs the annealing of the tracrRNA to the DR. Consequently, it is probably advisable that the entirety of the 30 bp spacer matches the sequence of its genomic target, and not only the 20 downstream base-pairs which are supposed to be maintained after crRNA maturation [179]. Although the modest improvement observed after flanking the tracrRNA with HH and HDV could be due to the additional streaking step (and thus longer duration of mutagenesis), it makes the prospects of exploiting the same ribozymes to multiplex sgRNA array more encouraging.

Unfortunately, no triple-KO colony could be isolated, and both CLAU534A and CLAU1794B were shown to mostly induce mutations which do not lead to a STOP codon. Because there were no more alternative protospacers available for these genes, it would be necessary to either increase drastically the mutagenesis efficiency of the existing protospacers, or increase the targeting space of Target-AID.

## VI.5. Expansion of the targeting space of Target-AID

The fact that already one of the only five protospacers tried with Target-AID did not lead to any mutation, and the lack of alternative protospacers, both raised the question of how relevant Target-AID might be over the scale of the whole genome of *C. autoethanogenum*. A total of 3964 CDS's have been predicted in *C. autoethanogenum*, which is far too large a number to screen manually with the sgRNA design table designed in Chapter V.2.1, or even for the more convenient web-based tool which eventually replaced it [332].

A collaboration was therefore established with Claudio Tomi Andrino from the School of Pharmacy at the University of Nottingham to design a new genome-wide tool based on his expertise in the Matlab programming language. What was needed was an estimation of the proportion of the CDS's of *C. autoethanogenum* that actually had a potential Target-AID sgRNA target early enough in their sequence that it was likely to knock-out the gene. The editing window of Target-AID was restricted to cytosines within the bases -19 to -16 from each PAM, and the locus of a potential target was limited to 75% of the CDS's length. The last instruction assumes that removing the last 25% of any given protein is likely to eliminate function. The term "base editing coverage" was adopted to describe the proportion of genes which could be knocked-out with Target-AID under these two criteria. A description of the algorithm is available on page 270.

### VI.5.1. Target-AID has poor base editing coverage in *C. autoethanogenum*

According to the analysis undertaken using the algorithm, only 51.61% of the 3964 CDS's of *C. autoethanogenum* are targetable by Target-AID, or just over a half. This disappointing value is mainly attributable to the low GC-content of *C. autoethanogenum*'s genome (31%). Indeed, ultimately, the coverage of Target-AID is depending twice on the abundance of G or C bases: once because it can only target protospacers immediately downstream of a NGG PAM; the second time because it can only mutate a C within the editing window of these protospacers.

Several strategies to improve the targeting space of Cas9 in general, or base editors in particular, have already been explored in the literature. Enlarging the editing window [259] is not desirable, because that would also increase the possibility of editing a base which does not produce a STOP codon instead of the targeted C. It also does not address the fundamental limitation of the reliance on G or C bases.

Developing an adenine base editor [308] instead of a cytosine base editor would reduce some of the reliance on GC-content, but it is impossible to create new STOP codons by changing a A to a G or a T to a C; as such, an adenine base editor would not be suited to purpose.

Instead, the most promising option is to relax the reliance on a GC-rich PAM. Several Cas9 proteins which recognize alternative, more AT-rich PAMs have been developed in recent years. The VQR and EQR Cas9 variants target NGA and NGAG PAMs, respectively [198,396]; the iSpymac variant targets NAA PAMs [200,309]; finally, xCas9 and Cas9−NG both target NG PAMs [260]. Near the end of this project, two additional Cas9 variants were published: SpG and SpRY, respectively targeting NG and NH (if not any) PAMs (H=A, C, G). The variants and their PAMs are summarized in Table 21.

**Table 21: Some of the published Cas9 variants and their associated PAMs.**

| Cas9 variant | PAM | Reference |
|---|---|---|
| **VQR** | NGA | [198,396] |
| **EQR** | NGAG | [198,396] |
| **QQR1** | NAAG | [310] |
| **iSpymac** | NAA | [200,309] |
| **xCas9** | NG/NGD | [312,397] |
| **Cas9−NG** | NG | [260] |
| **SpG** | NG | [314] |
| **SpRY** | N/NH | [314] |

Importantly, the NH-PAM of SpRY was not included in the bioinformatics analysis. The initial reason for this is that the description of SpG and SpRY was published after the analysis had already been completed [314]. Nonetheless, attempts to analyse it retroactively failed because the algorithm was too computationally intensive to process sequence requirements as permissive as "NH". It means that more than 3/4$^{th}$ of genomic bases are potential PAMs for this Cas9 variant. Even the requirement of NH is an oversimplification: SpRY was shown to target NG−PAMs quite effectively – just slightly less effectively than SpG, which was itself reported to be only marginally more effective than Cas9−NG. Consequently, the base editing coverage of SpRY can be assumed to be approximately 100%.

On their own, the SpG and SpRY variants do lead to a higher rate of off-target mutagenesis, but they are compatible with other mutations which drastically reduce the rate of off-targets [201]. It is unclear to us how the authors prevented SpRY from targeting the vector expressing its associated sgRNA. Such an activity was already measured when SpRY was applied in rice genome editing, although it was reduced when SRY was used within a base

editor [388,398]. In summary, these two variants are promising but need more characterisation and could not be tested within this project.

QQR1 was also removed from the analysis because it had been poorly characterized and its nuclease activity was reported as much lower than WT Cas9 [310].

Incidentally, a strategy which was overlooked here is the total replacement of Cas9 with Cas12a, which recognizes TTTV PAMs (V=A,C,G) upstream of a protospacer, instead of a NGG PAMs downstream of a protospacer. As mentioned previously, although it has been shown to work, this approach would have led to even less genomic targets because of the larger size of the PAM, and so was not tested here.

### VI.5.2. NG and NAA are the two most promising alternative PAMs

As illustrated in Figure 90, the alternative PAMs identified in the previous section provide a significant improvement of base editing coverage over the canonical NGG PAM. While using NGG PAMs could only hope to knock-out 51.64% of the genes of *C. autoethanogenum*, this value raises to 85.32% for NG, 82.37% for NGD and 81.81% for NAA. Combining the base editing coverage of NGG and NAA results in 85.80%, and targeting both NG and NAA PAMs would increase the base editing coverage to 91.78% of the genome.

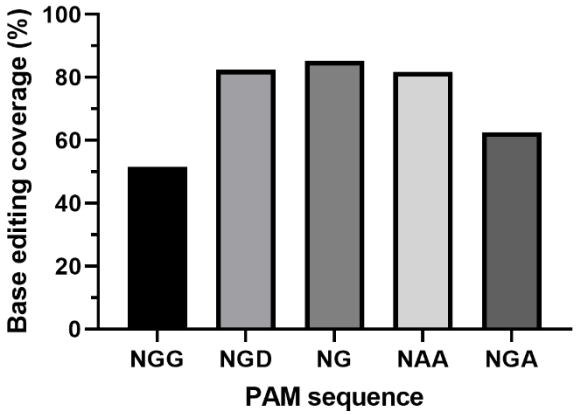



**Figure 90: Base editing coverage of Target-AID in *C. autoethanogenum* as a function of different PAM sequences**. Base editing coverage is defined here as the proportion of CDS's in the genome which have at least one protospacer within the first 75% of the gene which can produce a STOP codon when the cytosines in between the bases -19 to -16 from the PAM are mutated into thymines.

Cas9−NG has been reported as more effective than xCas9 [260,314]. Consequently, Cas9−NG and Cas9-iSpymac were chosen to continue optimizing Target-AID for applications in *C. autoethanogenum.*

### VI.5.3. Identification of alternative protospacers

Incidentally, the base editing coverage algorithm generated a database of all protospacers targetable with Target-AID and the Cas9 variants to produce STOP codons. This database was thus used to design the sgRNAs of the next round of multiplex mutagenesis (cf. Table 22).

**Table 22: Protospacers suitable to introduce a STOP codon in *C.autoethanogenum CLAU_0532, CLAU_0534 and CLAU_1794* after a C→ T mutation within an editing window of -19 to -16 bp from the PAM.** The **position** of a given PAM is given as a percentage of the whole gene. **Strand** indicates whether the protospacer is on the sense (1) or antisense (-1) strand; **Sequence** is the sequence of the sgRNA target (or protospacer); **PAM** is the sequence of the PAM, directly downstream of the protospacer.

| Gene | Label | Sequence | PAM | Type | Strand | Position (%) |
|---|---|---|---|---|---|---|
| *CLAU_0532* | A | ctccagtcaggtgttgtgca | tgg | NG | -1 | 24 |
| | D | tccagtcaggtgttgtgcat | ggt | NG | -1 | 24 |
| | E | agtccaagctggttttcagc | agc | NG | 1 | 29 |
| | F | tcagcagcattcaaacggta | tgc | NG | 1 | 30 |
| | G | cttccatcctgcaagcatac | cgt | NG | -1 | 30 |
| *CLAU_0534* | B | aggcagaaggcacaatttgt | tgc | NG | 1 | 34 |
| | A | agcccaatgtctagctggga | tgg | NG | 1 | 64 |
| | C | gcccaatgtctagctgggat | ggc | NG | 1 | 64 |
| *CLAU_1794* | C | atccaatctggcccaaattc | tgt | NG | -1 | 22 |
| | A | aaacaagcaattgttccgtt | tgg | NG | 1 | 32 |
| | D | agacaaaaagctaaatttgt | agc | NG | 1 | 35 |
| | E | ccagacaatgccacctaaat | taa | NAA | 1 | 49 |
| | B | atcacaatgtttagcaggta | tgg | NG | 1 | 64 |
| | F | tcacaatgtttagcaggtat | ggc | NG | 1 | 65 |
| | G | gccatacagctcctgtttta | tgt | NG | -1 | 68 |

Surprisingly, Cas9-iSpymac could only target one of the three genes, and with only one protospacer (CLAU1794E). As such, even with a genome coverage comparable to Cas9−NG, Cas9-iSpymac would not be suitable for the current experimental setup. As expected, the protospacers CLAU532A, CLAU534A, CLAU1794A and CLAU1794B were identified once again, but this time as targets for Cas9−NG. Each of these protospacers systematically produced two targets for Cas9−NG: one for each G of the NGG−PAM.

CLAU532A had been found to be a suitable target in both mutagenesis attempt, so was kept as a target in the next attempt. However, CLAU_0534 and CLAU_1794 needed different protospacers.

CLAU534C is almost identical to CLAU534A: it is just shifted from 1 bp. CLAU534B, on the other hand, had only one single C within the editing window, and would knock out a much large portion of the protein as it occurred earlier in the gene. CLAU534B was thus selected as the next target for *CLAU_0534*.

Similarly, CLAU1794C occurs the earliest in the gene. It has two cytosines in the editing window, but editing of any or both of them would lead to a STOP codon anyways. CLAU1794C was thus selected as the main target protospacer for *CLAU_1794*.

Because many more protospacers were still available to target *CLAU_1794*, it was opted to test three other protospacers targeting this gene in a separate construct. CLAU1794A and CLAU1794B had already been found lacking, and CLAU1794E could not be targeted with Cas9−NG. The three remaining protospacers, CLAU1794D, CLAU1794F, and CLAU1794G were thus chosen to assemble a backup multiplex construct which would allow them to be tested simultaneously.

### VI.5.4. Structural alignment with known alcohol dehydrogenases

A successful mutagenesis of protospacer CLAU1794G should reduce the size of the protein encoded by CLAU1794 by 32%. This should be sufficient to hinder the protein function, but could be confirmed by modelling the impact of the nonsense mutation on the 3D structure of the protein. To this end, the aa sequence of CLAU1794 was aligned with the 3D model of the alcohol dehydrogenase from *Thermococcus thioreducens* (SMTL ID : 6c75.1, NCBI ID: ASJ12775) [399], using SPDBviewer v.4.10 [400]. This alcohol dehydrogenase is substantially similar to CLAU1794 (37.94% of sequence identity) and, like CLAU1794, is NADP dependant. It was crystallized with an iron atom ($Fe^{3+}$) and an NADP molecule, facilitating the visualisation of the active sites. After alignment [cf. Figure 91(a)], it became apparent that the deleted domain [cf. Figure 91(b)] could only possibly interact directly with the iron ion. The four aa interacting directly with this iron ion [399] were highlighted in both structures [cf. Figure 91(c) and (d)], and one of them was shown to be downstream of the W278X nonsense mutation induced by CLAU1794G. Thanks to this direct interference in the active site of the protein, it is thus likely that a successful *CLAU_1794(S251L, W278X)* mutation would severely disrupt the activity of the protein. The protospacer CLAU1794G was thus an appropriate target for Target-AID, in spite of its late occurrence in the CDS of CLAU_1794.

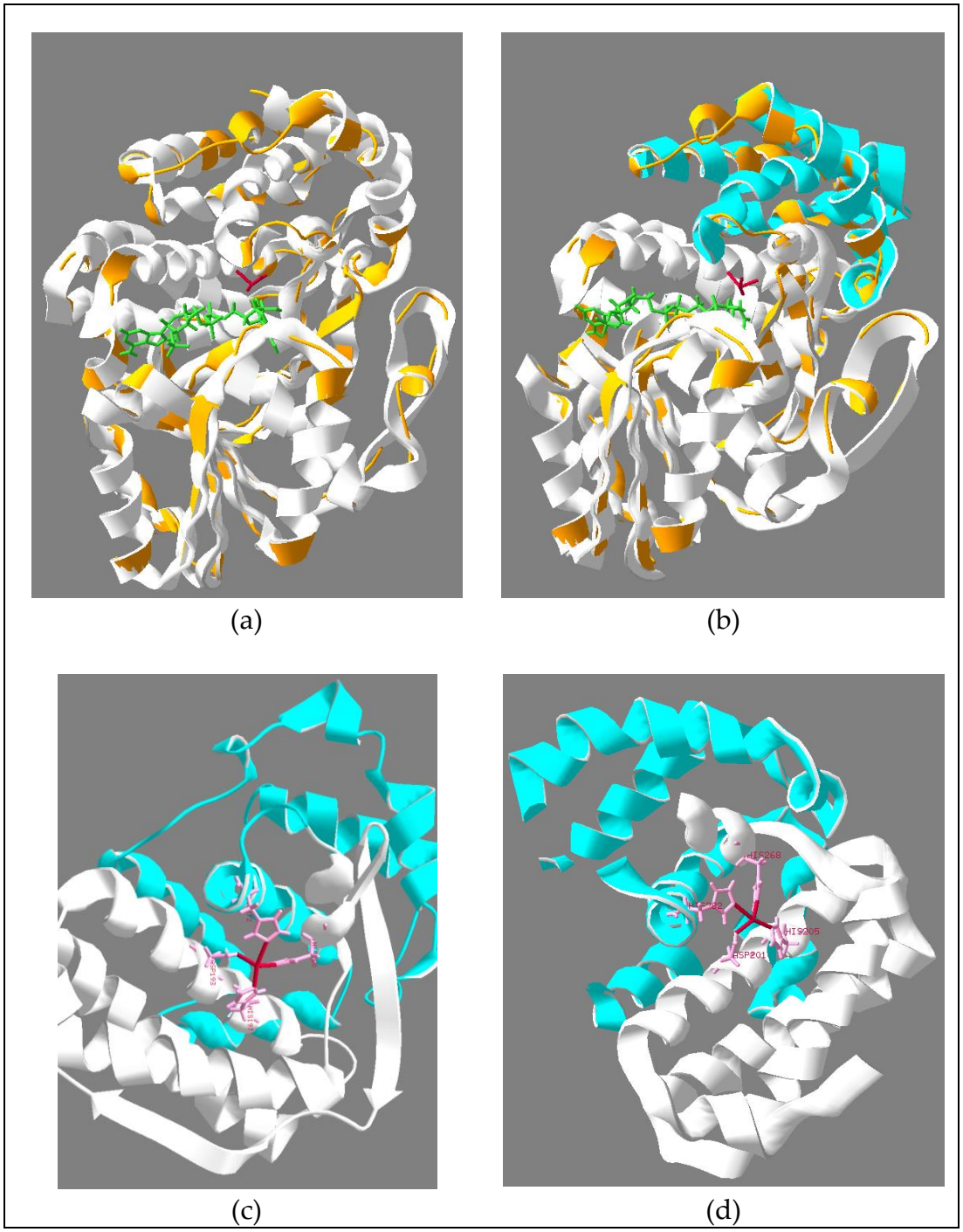



**Figure 91: Structural model of CLAU1794 by alignment with an ADH from *T. thioreducens* (6c75.1) with SPDBviewer v4.10. (a)** Superposition of 6c75.1 (yellow) with CLAU1794 (white). The ligands are represented in green (NADP) and red ($Fe^{3+}$) **(b)** Highlight of the aa lost after the W278X mutation (turquoise). **(c)** Close-up of the Fe-binding domain of 6c75.1 (white), with key aa (His197, His260, His272 and Asp193) in pink and the structure lost after the equivalent to the CLAU_1794 (S251L, W278X) mutation in turquoise. **(d)** Close up on the Fe-binding domain of CLAU1794 (white), with homologous key aa (His205, His268, His282, and Asp201) in pink and the structure lost after the W278X mutation in turquoise.

### VI.5.5. Characterisation of TA-iSpymac

Although Cas9-iSpymac was not suitable to target CLAU532 and CLAU534, its availability would still improve the range of genes targetable by Target-AID. A Target-AID construct using a Cas9-iSpymac nickase (TA-iSpymac) was thus assembled in order to test its functionality in *C. autoethanogenum*. The PAM-interacting domain of Target-AID (V1100 to D1368) was replaced with the codon-optimized PAM-interacting domain of the Cas9 from *Streptococcus macacae* (I1078 to D1338) and the cutting efficiency of the system was improved by implementing the R221K and N394K mutations. The sgRNA cassette of the Target-AID vector under the expression of $P_{fdxE}$ was also changed to target a NAA PAM within the *pyrE* gene (GCAAAATGGAAATATGACTA, with a TAA PAM, knocking out 52% of *pyrE*). This resulted in the vector vFS83_TA-spymac_pyrE (cf. Figure 92). The accuracy of Cas9-iSpymac was later reported to be further improved with the R691A mutation, but this last mutation is not part of my design.

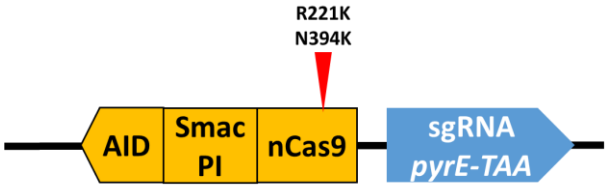



**Figure 92: Schematic of the genome editing cassette of vFS83_TA-spymac_pyrE. Smac PI = Pam interacting domain from *S. macacae*.**

vFS83_TA-spymac_pyrE was then conjugated into *C. autoethanogenum* alongside the constructs of the second replicate of the characterisation of alternative Target-AID expression systems in Chapter V.4.1.

#### *VI.5.5.1) TA-iSpymac performed better than Target-AID*

Even though vFS83_TA-spymac_pyrE yielded fewer colonies than vFS36_TA_PfdxE_pyrE across the board, it reached both a higher survival after induction and a higher mutagenesis efficiency (81±19% and 35±2% versus 55±5% and 25±4%, respectively) [cf. Figure 93]. Only the difference in mutagenesis efficiency was statistically significant (Welch's t-test, p=0.04, n=3).

Seven colonies from the induction plates before and after FOA selection were sent to Sanger sequencing. As previously described for vFS36_TA_PfdxE_pyrE (cf. Chapter V.4.1), no mutation could be detected in any of the colonies growing in the absence of FOA. However, contrarily to what was observed with vFS36_TA_PfdxE_pyrE, all seven colonies growing on FOA harboured the desired mutation, with only one of them still showing signs of a slightly mixed WT peak.

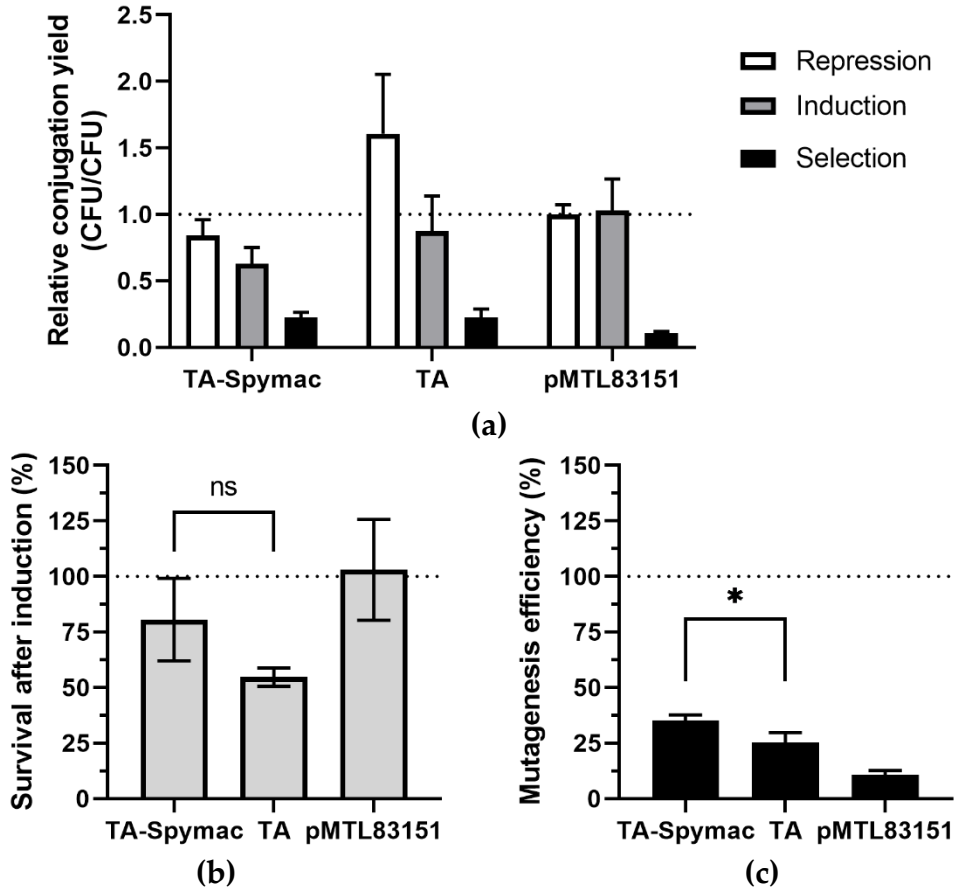



**Figure 93: Characterisation of TA-iSpymac targeting *pyrE* in *C. autoethanogenum*. (a)** Relative conjugation yield; **(b)** Percentage of survival after induction; and **(c)** mutagenesis efficiency. TA = vFS36_TA_PfdxE_pyrE, Target-AID vector with sgRNA cassette targeting *pyrE*; TA-Spymac= vFS83_TA-spymac_pyrE. All constructs were induced with 5 mM of theophylline. Then, induced colonies were transferred with a velvet stamp to FOA selection plates with the same concentration of theophylline, to select for mutants. Colony count is reported as a proportion of the CFU of the pMTL83151 control construct (Welch's t-test, n=3).


**Figure 94: Representative Sanger sequencing traces of different alleles (A.1, A.2) obtained after conjugation of vFS83_TA-spymac_pyrE in *C. autoethanogenum growing* on 5 mM FOA.** The sequence of the WT protospacer is highlighted in yellow; the corresponding aa sequence is displayed in a series of consecutive arrows; a vertical bar strikes the start of the targeted codon (in the forward direction); mismatches clear enough to be detected automatically are highlighted in red. Six colonies out of seven carried the allele A.1, and the remaining one had the allele A.2. Traces and annotations obtained with Benchling.

#### *VI.5.5.1) TA-iSpymac should be an effective targeted genome editing tool*

In the narrow context of the previous chapter, the relative success of TA-iSpymac would not be sufficient to make it a convenient targeted mutagenesis tool. After all, it still underperforms what has been observed during the preliminary characterisation of Target-AID, and real knock-out experiments rarely show a selectable mutant phenotype. Given how Target-AID has shown to be much more effective throughout this chapter than during its formal characterisation in the previous chapter, it is nonetheless reasonable to assume that TA-iSpymac would also be an effective tool in the absence of FOA selection. Under this hypothesis, the failure to isolate mutated colonies from the plates without theophylline in the previous chapter might have been caused by manipulation mistakes or other freak incident.

If, however, the data collected throughout this experiment can be trusted, it would be unclear which of the PAM-interacting domain of *S. macacae*, the different protospacer, or the R221K and N394K mutations best explain the superior performance of TA-iSpymac relative to the canonical Target-AID. In the latter case, these mutations could be directly transposed to Target-AID in order to improve its performance.

## VI.6. Third round of multiplex genome editing with Target-AID

Satisfied with the performance of the multiplexing systems, attention was focused on strategies to finally obtain a triple knock-out strain. As described in the previous section, this was achieved by increasing the number of potential targets for Target-AID, but also by exploring ways of increasing conjugation efficiency. This time, only the best best performing multiplexing strategy was used, namely, the array of individual sgRNA cassettes.

Additionally, constructs were plated on repression plates in addition to induction plates so as to measure explicitly the survival rate of the multiplex constructs.

### VI.6.1. Final redesign of the multiplexing constructs

On top of modifying Target-AID to recognize alternative PAMs, several strategies were explored to improve mutagenesis efficiency. The first method was to fuse Target-AID with a third protein, UGI, described in the previous chapter (cf. Chapter V.1.2). It had been shown to be toxic and increase off-target mutagenesis in *E. coli*; as such, an LVA degradation tag had been added to UGI in an attempt to mitigate these undesirable properties in *E. coli* [304]. The published design was adhered to in the generation of this construct, retaining the LVA degradation tag. The second method consisted in shortening the protospacers in order to shift the editing window of Target-AID [260,298] closer to the position -16 from the PAM, where sits the targeted C of CLAU534A and CLAU1794B.

#### *VI.6.1.1) Target-AID−NG*

Modifying the PAM recognition domain of SpCas9 only requires seven point mutations: R1335V, L1111R, D1135V, G1218R, E1219F, A1322R, and T1337R. These mutations were brought onto the vFS57_TA_msgRNA_CA1794B vector, and its last two protospacers were changed to CLAU534B and CLAU1794C. This resulted in the vector vFS72_mTA−NG (CF. Figure 95).

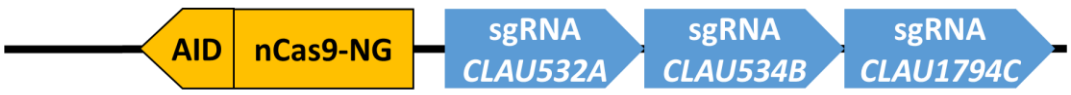



**Figure 95: Schematic of the genome editing cassette of vFS72_mTA−NG.**

### *VI.6.1.2) Target-AID-UGILVA*

UGI and its LVA degradation tag were added to the end of Target-AID in the vFS57_TA_msgRNA_CA1794B vector to create the vector vFS75_mTA-UGILVA (CF. Figure 96).

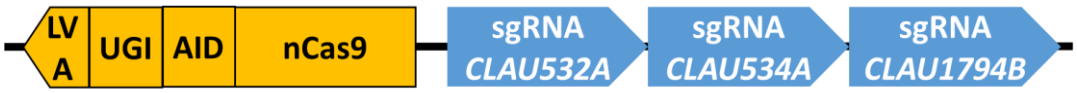



**Figure 96: Schematic of the genome editing cassette of vFS75_mTA-UGILVA.**

### *VI.6.1.3) Shorter protospacers*

The sequence of the protospacers CLAU534A and CLAU1794B from vFS57_TA_msgRNA_CA1794B were shortened from 20 bp to 18 bp, resulting in the truncated protospacers trCLAU534A and trCLAU1794B as described in Table 23. This resulted in the vector vFS94_mTA_trsgRNA.

**Table 23: Truncation of protospacers to shift the editing window of Target-AID.**

| Gene | Protospacer | Sequence |
|---|---|---|
| ***CLAU_0534*** | CLAU534A | agccCAAtgtctagctggga |
| | trCLAU534A | ccCAAtgtctagctggga |
| ***CLAU_1794*** | CLAU1794B | atcaCAAtgtttagcaggta |
| | trCLAU1794B | caCAAtgtttagcaggta |

### *VI.6.1.4) Multiplex CLAU_1794 knock-out*

In a separate experiment, an attempt was made to knock out CLAU1794 from *C. autoethanogenum adh1(Q243X), adh3(W90X)* strain using the three remaining protospacers (CLAU1794D, CLAU1794F, and CLAU1794G). vFS72_mTA−NG was thus modified to target only CLAU_1794 but with three different protospacers, resulting in the vector vFS74_mTA−NG_CA1794 (cf. Figure 97).


**Figure 97: Schematic of the genome editing cassette of vFS74_mTA−NG_CA1794.**

### VI.6.2. Multiplex KO of CLAU_0532, CLAU_0534, CLAU_1794

The three new multiplexing vectors vFS72_mTA−NG, vFS75_mTA-UGILVA, vFS94_mTA_trsgRNA and pMTL83151 were conjugated into *C. autoethanogenum ΔpyrE::ACE;* vFS57_TA_msgRNA_CA1794B was also conjugated as a control. Each slurry was normalized to 650 µL and 100 µL were plated in triplicate both on induction and repression plates. Eight colonies of each constructs were patched on induction plates, and five of these were amplified and sequenced by Sanger sequencing (cf. Table 24).

**Table 24: Sequencing of five *C. autoethanogenum* colonies conjugated with different Target-AID constructs and induced on theophylline. vFS72 = vFS72_mTA−NG; vFS75 = vFS75_mTA-UGILVA; vFS94= vFS94_mTA_trsgRNA; and vFS57= vFS57_TA_msgRNA_CA1794B.** The targeted codon is highlighted in green, and mutated bases are capitalized and bolded. Fractions indicate mixed reads for this base.

| | | | Mutagenesis efficiency | | | |
|---|---|---|---|---|---|---|
| | | | Redesigned | | | Control |
| Protospacer | Allele | Sequence | vFS72 | vFS75 | vFS94 | vFS57 |
| CLAU532A | WT | ct**cca**gtcaggtgttgtgca | 0 | 4/4 | 0 | 0/5 |
| | A.2 | ct**T/C T/C**agtcaggtgttgtgca | 0 | 0 | 0 | 5/5 |
| | A.3 | ct**T/C**cagtcaggtgttgtgca | 3/3 | 0 | 3/4 | 0 |
| | A.8 | **T**t**T/C T/C**agtcaggtgttgtgca | 0 | 0 | 1/4 | 0 |
| CLAU534A | WT | agc**caa**tgtctagctggga | | 3/5 | 5/5 | 0 |
| | A.1 | ag**T**c**caa**tgtctagctggga | | 2/5 | 0 | 0 |
| | A.4 | ag**T/C T/C**caatgtctagctggga | | 0 | 0 | 1/5 |
| | A.6 | ag**T**t**T/C**caatgtctagctggga | | 0 | 0 | 3/5 |
| | A.7 | ag**T**t**T/C T/C**aatgtctagctggga | | 0 | 0 | 1/5 |
| CLAU534B | WT | **agg**cag**aaggcacaatttgt** | 3/4 | | | |
| | A.1 | **agg**t**T/C**ag**aaggcacaatttgt** | 1/4 | | | |
| CLAU1794B | WT | atc**aca**at**gtttagcaggta** | | 5/5 | 5/5 | 1/5 |
| | A.2 | at**T/C**a**caa**t**gtttagcaggta** | | 0 | 0 | 4/5 |
| CLAU1794C | WT | at**cca**atctggcccaaattc | 2/5 | | | |
| | A.1 | at**T/C**caatctggcccaaattc | 3/5 | | | |

#### *VI.6.2.1) Cas9−NG is functional, but not efficient enough*

vFS72_mTA−NG yielded mutants in all three genes, but all of them were mixed with WT. If 100% of the colonies exhibited mixed reads for CLAU532A, only 25% did in CLAU534B and 60% in CLAU1794C. None of the colonies were mutated in all three genes at once. However, the fact that some mutants were obtained shows that Cas9−NG may be suitable to use in *C. autoethanogenum* outside of a base editor, or only to target one target at a time.

### *VI.6.2.2) UGI and truncated protospacers harm conjugation efficiency*

Contrarily to what had been described in the literature, neither UGI nor shorter sgRNAs increase the mutagenesis efficiency of CLAU534A and CLAU1794B. In fact, only the WT sequences of these two loci were observed with vFS94_mTA_trsgRNA, and vFS75_mTA-UGILVA only managed to mutate 40% of the colonies screened in CLAU534A. In comparison, 100% of CLAU534A and 80% of CLAU1794B were mutated with mixed reads in the vFS57_TA_msgRNA_CA1794B control.

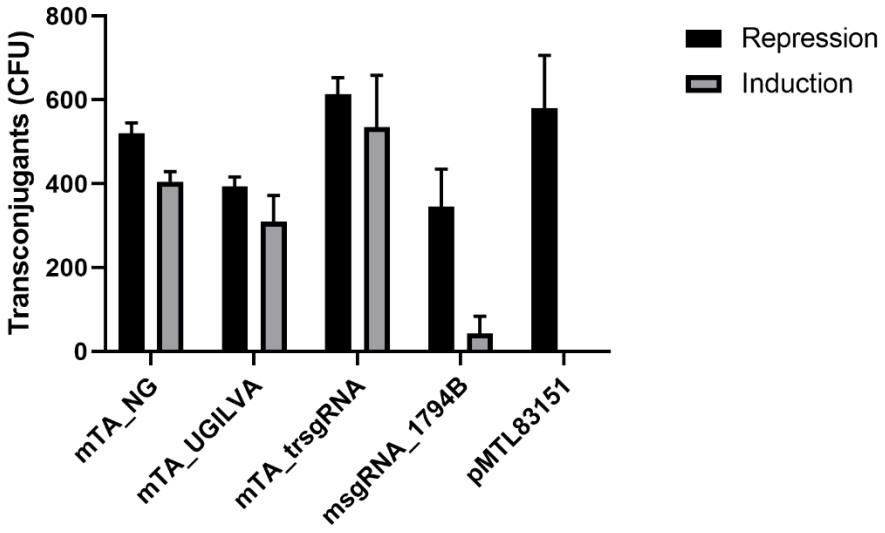



**Figure 98: Number of *C. autoethanogenum* *ΔpyrE::ACE* transconjugant colonies of different Target-AID constructs.** All constructs were plated on YTF with antibiotics and 5 mM of theophylline (induction) or without theophylline (repression). mTA_NG = vFS72_mTA−NG; mTA_UGILVA = vFS75_mTA-UGILVA; mTA_trsgRNA = vFS94_mTA_trsgRNA, msgRNA_1794B = vFS57_TA_msgRNA_CA1794B.

### *VI.6.2.3) Empty vector colonies did not survive theophylline induction*

The number of transconjugants mostly followed the expected trend, with vFS57_TA_msgRNA_CA1794B – the most effective construct – also being the most toxic after induction (Cf Figure 98). It went from 346±89 CFU to 42±42 CFU after induction. This is equivalent to a survival rate of 50% per targeted gene. The conjugation yield of other multiplex constructs, however, barely changed value after induction. This is somewhat surprising, due to each construct producing at least some mixed colonies, but it makes sense that their survival after induction is higher than that of vFS57_TA_msgRNA_CA1794B – which produced almost 100% of mixed colonies. Interestingly, the same cannot be said of the empty vector control, pMTL83151, which, this time, did not survive induction at all. As discussed in the previous chapter, theophylline toxicity might be partially to blame, but something else must have impacted the plate to produce such a drastic result.

### *VI.6.3.* Multiplex KO of CLAU_1794 in *C. autoethanogenum adh1(Q243X), adh3(W90X)*

In parallel, vFS74_mTA-NG_CA1794 and pMTL83151 were conjugated into *C. autoethanogenum adh1(Q243X), adh3(W90X).* This was meant to test the remaining protospacers targeting CLAU1794, as well as a last attempt to obtain the triple knock-out. Each slurry was normalized to 600 µL, and 100 µL were plated on both induction and repression plates. Ten colonies were patched and five were sent to Sanger sequencing.

**Table 25: Sequencing of five *C. autoethanogenum* colonies conjugated with different Target-AID constructs and induced on theophylline.** The targeted codon is highlighted in green, and mutated bases are capitalized and bolded. Fractions indicate mixed reads for this base.

| Protospacer | Allele | Sequence | Mutagenesis efficiency vFS74_mTA-NG_CA1794 |
|---|---|---|---|
| CLAU1794D | WT | Agacaaaaaagctaaatttgt | 5/5 |
| | A.1 | Aga**T**aaaaagctaaatttgt | 0/5 |
| CLAU1794F | WT | tcacaatgtttagcaggtat | 0 |
| | A.2 | t**T**acaatgtttagcaggtat | 5/5 |
| CLAU1794G | WT | gccatacagctcctgtttta | 0 |
| | A.1 | g**TT**atacagctcctgtttta | 1/5 |
| | A.2 | g**T**catacagctcctgtttta | 1/5 |
| | A.2 | g**T/C**catacagctcctgtttta | 3/5 |

#### *VI.6.3.1) Two out of three protospacers had 100% mutagenesis efficiency*

The protospacers CLAU1794F and CLAU1794G were mutated with maximal efficiency (cf. Table 25); in fact, all colonies were pure mutants in CLAU1794F, and 40% of CLAU1794G loci were also pure. CLAU1794F, however, failed to mutate the cytosine which would have produced a STOP codon; in contrast, all colonies exhibited a nonsense mutation in CLAU1794G. After streaking the colony with the CLAU1794G.A.1 allele on YTF without antibiotics, a pure *C. autoethanogenum adh1(Q243X), adh3(W90X) CLAU_1794(S251L, W278X)* double mutant was isolated, and labelled cFS05_CA_Δ532Δ534Δ1794.

#### *VI.6.3.2) The downstream sgRNA cassettes outperformed most upstream one*

Interestingly, the sgRNAs targeting these two protospacers are in the last positions of the sgRNA array. These sgRNA cassettes seemed to be underperforming in the two previous attempts, but this does not seem to be the case here. Their low conjugation efficiency in the previous experiments was thus probably due to sub-optimal protospacers rather than poor expression cassettes.

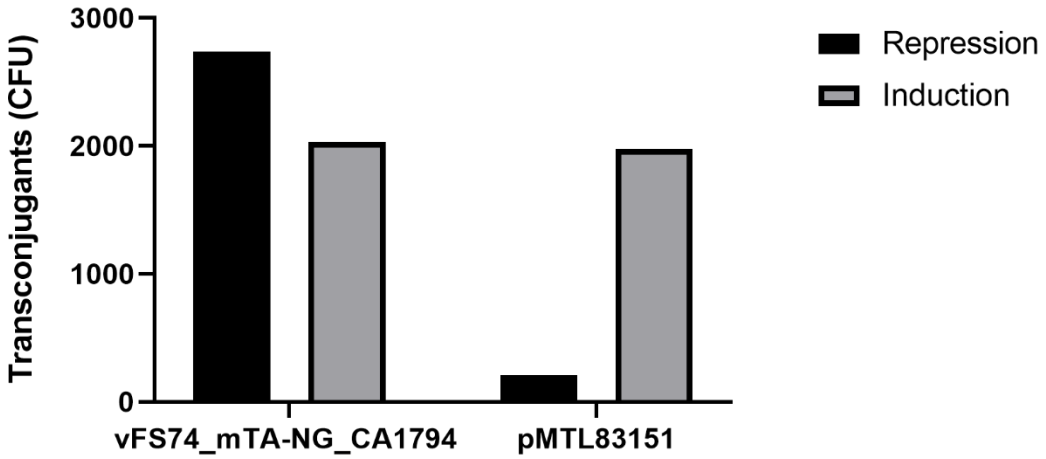



**Figure 99: Number of *C. autoethanogenum adh1(Q243X), adh3(W90X)* transconjugant colonies with different Target-AID constructs.** All constructs were plated on YTF with antibiotics and 5 mM of theophylline (induction) or without theophylline (repression).

### *VI.6.3.3) The conjugation yield was exceptionally high*

Conjugating vFS74_mTA−NG_CA1794 and pMTL83151 into *C. autoethanogenum adh1(Q243X), adh3(W90X)* resulted in over 2000 CFU in only one sixth of the mating mix (cf. Figure 99); or four times as much as the conjugation yield previously obtained with *C. autoethanogenum ΔpyrE::ACE*. As discussed in Chapter IV:, page 123, this might be a consequence of the recent loss of the vector vFS57_TA_msgRNA_CA1794B from *C. autoethanogenum adh1(Q243X), adh3(W90X)*. Indeed, all the *C. autoethanogenum adh1(Q243X), adh3(W90X)* cells are the progeny of a cell which was more receptive than most to conjugation with a pMTL vector; such a property might have been inherited (genetically or epigenetically) by its progeny, facilitating subsequent conjugations

Such a high CFU makes any colony count imprecise; nonetheless, considering that only two genes were successfully targeted, it would imply a survival after induction of vFS74_mTA−NG_CA1794 of about 37% per gene.

### *VI.6.3.4) The pMTL83151 control behaves aberrantly*

As in the previous experiment, it is hard to make sense of the behaviour of the pMTL83151 vector. This time, colonies fared much worse without theophylline (210 CFU) than with it (1976 CFU). No explanation can be found within the context of this experiment, so it must be assumed that an external element has disturbed the setup, at least for the repression plate of pMTL83151 (for example, quick dessication of that plate or exposure to oxygen).

### VI.6.4. Conclusion and perspectives

In this last round of multiplex mutagenesis, a Target-AID construct with an expanded PAM recognition potential was successfully implemented. However, Target-AID−NG, like the conventional Target-AID, was not effective enough to produce a triple knock-out in a single step. Instead, a second round of mutagenesis was necessary in order to obtain a triple knock out strain [*C. autoethanogenum adh1(Q243X), adh3(W90X) CLAU_1794(S251L, W278X)]*.

Two common strategies to increase the efficiency of Target-AID [260,298,304] were also implemented, but without success: its fusion with a UGI domain complemented with a LVA-tag to alleviate the potential toxicity of UGI, and shorter sgRNAs in order to target more effectively the bases at the downstream end of the editing window. On the contrary, these devices considerably reduced mutagenesis efficiency. Tentatively, the LVA-tag decreased the intracellular concentration of Target-AID-UGILVA excessively and removing it would allow UGI to increase the mutagenesis efficiency as originally intended.

Lastly, a strange behaviours of the control vector pMTL83151 on theophylline was encountered; these events encouraged the testing of a lower concentration of theophylline in the last experiment of the previous chapter (cf. Chapter V.5).

## VI.7. Summary and perspectives

In this ambitious chapter, Target-AID was used to knock out three genes in *C. autoethanogenum*. Three Cas9-mediated mutagenesis multiplexing systems were explored and improved upon: an array of individual sgRNA expression cassettes, a synthetic CRISPR array, and an operon of sgRNAs interspersed by pre-tRNA. All methods worked, and although the array of individual sgRNAs produced the most reliable results, decent performances were observed in the other two systems, with ample room for improvement. Notably, the synthetic CRISPR array would likely benefit from using 30 bp spacers which fully anneal to their protospacers and do not include any restriction site. On the other hand, more tRNAs could be screened in order to find the ones which help processing the sgRNA array the most effectively. A proper characterisation of the processing of the tRNA array is recommended to gain more insights into the actual molecular mechanisms involved. To the best of my knowledge, it is the first time that tRNA maturation is exploited to process an array of sgRNA in prokaryotes.

Two published SpCas9 variants (Cas9-NG and Cas9-iSpyMac) were also reproduced in order to expand the targeting space of Target-AID. This strategy was effective, but insufficient to reliably obtain triple knock-outs in one single step. Other tactics were attempted to improve the conjugation efficiency of Target-AID (notably, using shorter spacer sequences and fusing Target-AID to UGI with an LVA-tag), without success. A Target-AID-UGI protein fusion without LVA-tag might yield better results, and it might be improved further with the R221K and N394K mutations from Cas9-iSpyMac.

Adequate protospacer design was shown to be critical: when more than one cytosine was present within the editing window, the closest cytosine from to the position -18 was the most frequently mutated, even if it did not produce a STOP codon. Whenever possible, only protospacers where any and all cytosines within the -20 to -16 editing window would lead to a STOP codon if changed to a thymine should be picked.

Unfortunately, time constraints meant that there was no opportunity to characterise the phenotype of the *C. autoethanogenum adh1(Q243X), adh3(W90X), CLAU_1794(S251L, W278X)* strain in serum bottles and bioreactors or to sequence its whole genome to look for potential off-target mutations. This would be the natural continuation of this work.

With the hindsight gained in the Chapter IV about the potential impact of excessive concentration of theophylline on the survival after induction and the conjugation efficiency,

reproducing these experiments at 2.5 mM instead of 5 mM of theophylline might lead to a more successful outcome. Prolonging the duration of mutagenesis might also increase the chances of obtaining a triple knock-out strain.

Nonetheless, the simultaneous targeting mutation of two different loci was achieved on several occasions without relying on the HDR machinery or selective pressure. With its extended PAM recognition, Target-AID−NG was thus proven to be a valuable and promising tool for genome editing in *Clostridium*, as should be Target-AID-iSpymac.

# Chapter VII: General conclusion and future work

As the global average temperature keep rising and the 6$^{th}$ mass extinction follows its course, drastic changes in the way humanity processes and utilizes its limited resources have become necessary. If used wisely, industrial biotechnology – and in particular gas fermentation – can facilitate these changes by reducing the carbon and wider environmental footprints of many existing products – from fuels to food and chemicals. It is also critical to building resilience and self-sufficiency into our energy and material supply chains in a post-oil age. In addition to the existing pool of biodiversity, industrial biotechnology relies on the domestication of microorganisms, which demands a suite of effective, precise and well characterized synthetic biology tools applicable in each organism.

Over the course of this thesis, a set of new genome editing tools have been proven compatible with *C. autoethanogenum* – a gas fermenting organism – and most likely many other species.

The described studies began with the design of Cas9-targetable genomic bookmarks: short sequences of 24 nt which can be used as placeholder in knocked-out loci to facilitate subsequent complementation while minimizing the chances of polar effects. Bookmarks also have the potential to facilitate the assembly of genetic constructs directly into the genome, for example by incremental rounds of insertion and targeting of bookmarks alongside the knock-in of genes involved in heterologous metabolic pathways. In another possible application, inserting a bookmark upstream of every genomic heterologous promoter would offer a convenient way to swap each promoter during fine-tuning of an engineered metabolic pathway.

Attention then turned to improving and understanding the intricacies of the conjugation protocol routinely being used in the laboratory. The data that emerged made the protocol more specific and unravelled the influence of previously hidden factors negatively effecting the yield of transconjugant, such as prolonged cryopreservation and excessive amount of sExpress donor cells. The need for a meaningful definition of conjugation efficiency was identified leading to the proposal to calculate it relative to an internal empty vector control conjugated alongside each conjugation. Ideally, such a vector should be compatible with as many organisms and antibiotics as possible in order to enable comparisons in between species and experiments.

It was established that the decrease in transconjugants yield obtained using the WT *C. autoethanogenum* strain was not due to genetic drift, but previously unknown mutations in other strains were nonetheless identified. By inoculating donor cells directly from the transformation plate on the day of the mating instead of from a cryostock on the day before, by lowering the amount of sExpress donor cells plated, and by limiting desiccation, it proved possible to recover a consistently high transconjugant yield and arguably improved it further. In the absence of genetic basis to this improvement, it may be hypothesized that the conjugation potential of *C. autoethanogenum* and *E. coli* sExpress is subject to epigenetic changes. This could be verified by DNA methylation sequencing of the recipient strains used under the different treatments explored over the course of this thesis, and lead to a better understanding of the genes and mechanisms involved in bacterial conjugation.

The final two chapters tested and characterized a state-of-the-art targeted genome editing tool capable of changing a single base in the genome of *C. autoethanogenum*. This involved the initial development of a custom bioinformatics tool which enabled the screening of a list of protospacers for putative Target-AID targets. However, the subsequent characterisations of Target-AID on the *pyrE* locus of *C. autoethanogenum* led to conflicting results: the preliminary characterisation concluded to a mutagenesis efficiency of 70%, while the final characterisation apparently failed to mutate any colony prior to FOA selection. This incoherence raises concerns about the validity of the experimental setup employed. The quantitative characterization of Target-AID should be attempted again, preferably through the intermediary of next-generation sequencing in order to avoid some of the complications encountered here. It did not prove possible to establish whether Target-AID performed better or worse than conventional HDR-mediated mutagenesis, or whether inducing at 2.5 mM of theophylline would improve the outcome of Target-AID mutagenesis. In the process, concerns were raised over the toxicity of the theophylline, the molecule used to express Target-AID, but also showed that two alternative expression systems ($P_{facoid}$/LacI and $P_{thl}$/trTarget-AID) performed worse than the theophylline-inducible riboswitch: the former seemed unstable, while the latter had an altered editing profile.

Lastly, three genes from *C. autoethanogenum* were knocked out using Target-AID in combination with different multiplexing tactics to express an array of sgRNAs. In this context, Target-AID worked reliably, although it often resulted in mixed colonies and incorrect mutants. This help formalizing the importance of protospacer design: the ideal target should not have any non-target cytosine in the editing window. This rapidly led to a lack of suitable

protospacers, which was solved by assembling Target-AID variants which could target non-canonical PAMs (namely, NG and NAA PAMs). A collaborative bioinformatics analysis modelled that this initiative now enables the targeting of over 91% of the genes of *C. autoethanogenum*, an increase of 40% from the coverage of the canonical NGG PAM alone. Other alterations (such as using shorter sgRNAs and adding a UGI with a LVA degradation tag to Target-AID) were attempted to improve the mutagenesis efficiency of Target-AID, but without success. Eventually, only two genes could be successfully targeted simultaneously. Nonetheless, the data suggests that three genes should be targetable in one go with the right protospacers. The efficacy of Target-AID could still be improved in the future, for example by combining it to UGI without a LVA-tag or by incorporating the R221K and N394K mutations from Cas9-iSpyMac. Similarly, the possibility of off-target mutagenesis events should be assessed by whole genome sequencing and could be reduced by using a Cas9 engineered for high fidelity if necessary.

An array of separate expression cassettes was the most effective multiplexing method to express sgRNAs, although successful mutagenesis events were also obtained by separating a polycystronic array of sgRNAs with either tRNAs from *C. pasteurianum* or DR from the original *S. pyogenes* CRISPR system. To the best of my knowledge, it is the first time tRNAs are being used in this manner in prokaryotes. My design for the DR array might benefit from using 30 bp protospacers instead of 20 bp protospacers combined with a restriction site and four random nucleotides, and more tRNAs could be screened in order to find the most effective ones. Notably, tRNAs which integrate RNA-polymerase binding sites might be particularly effective at expressing and processing a sgRNA array.

Target-AID was thus proven for the first time to be an effective, albeit delicate, tool for (multiplex) genome editing in *C. autoethanogenum*. It has since been used successfully in several related species to knock out genes on which conventional HDR-mutagenesis strategies were ineffective: in *C. sporogenes* by Dr. Syeda Rivzy [401], *Clostridium beijerinckii* by Lauren Boak and *C. difficile* by Joanna Steczynska (unpublished).

Nevertheless, base editing did show some weaknesses: on top of its propensity to produce mixed colonies and its reluctance to mutate all the cytosines within its editing window, it is still limited by the choice of its protospacers. Indeed, among the eleven Target-AID-compatible protospacers tested in this thesis, only nine could actually be targeted and only seven of these resulted in the desired mutation at least once. The fusion with Cas9 recognizing alternative PAMs alleviated this concern, but did not solve it completely: indeed,

9% of *C. autoethanogenum* CDS' still have no suitable protospacers at all, with probably many more CDS harbouring only defective protospacers.

The development of prime editing in *Clostridium* – or the use of a Cas9 protein fusion and pegRNA to rewrite a short sequence of the genome via reverse-transcription – is a potential solution to this problem. Not only does it drastically increase the number of available protospacers, since it can rewrite entire codons instead of relying on existing ones which already closely resemble a STOP codon, it can also induce any kind of short mutation with great flexibility – allowing, for example, to undertake *in vivo* protein engineering. Exploiting the alternative NHEJ pathway or introducing a heterologous NHEJ pathway in *Clostridium* could also be a successful alternative strategy of producing knock-outs independently from the HR system.

With Bookmarks, additional insights into bacterial conjugation, and multiplexed Target-AID-NG/iSpymac the work described in this thesis have made a modest contribution to the synthetic biology toolbox of *Clostridium* and many other organisms. It is hoped that it will accelerate research and metabolic engineering strategies, enabling bold and innovative solutions to the current sustainability crises.

In the meantime, the reader, as well as policymakers, are urged to also consider non-technological solutions to the sustainability crises. Indeed, the surest and most cost-effective way of reducing $CO_2$ emissions is not to capture them with gas fermentation or any other Carbon Capture, Utilisation and Storage (CCUS) technology in the distant future, but to stop emitting them today. This can be achieved through many social, political, economic or financial reforms; sometimes even facilitated by already tried and tested – but often under-exploited – technologies. If $CO_2$ emissions are reduced quickly enough, more technological solutions such as the ones this thesis contributed to might be given enough time to emerge and help us maintain high standards of living, this time while respecting the planetary boundaries.

# References


1. Cornwall, W. Europe's Deadly Floods Leave Scientists Stunned Available online: https://www.sciencemag.org/news/2021/07/europe-s-deadly-floods-leave-scientists-stunned (accessed on 31 August 2021).

2. Burton, K. As China and UK Fall out, Millions Are Displaced in Devastating China Floods Available online: http://ftp.geographical.co.uk/nature/climate/item/3761-millions-displaced-in-devastating-china-floods (accessed on 31 August 2021).

3. Chan, V. Death Toll Continues to Rise after Floods in Central China That Displaced over 1.2 Million Peop Available online: https://www.scmp.com/video/china/3142054/death-toll-continues-rise-after-floods-central-china-displaced-over-12-million (accessed on 31 August 2021).

4. Philip, S.Y.; Kew, S.F.; Oldenborgh, G.J. Van; Yang, W.; Vecchi, G.A.; Anslow, F.S.; Li, S.; Seneviratne, S.I.; Luu, L.N.; Arrighi, J.; et al. Rapid Attribution Analysis of the Extraordinary Heatwave on the Pacific Coast of the US and Canada June 2021 . Available online: https://www.worldweatherattribution.org/wp-content/uploads/NW-US-extreme-heat-2021-scientific-report-WWA.pdf.

5. Dixon, R. Siberia's Wildfires Are Bigger than All the World's Other Blazes Combined Available online: https://www.washingtonpost.com/world/2021/08/11/siberia-fires-russia-climate/ (accessed on 31 August 2021).

6. Russia's 2021 Wildfires Now Largest in Its Recorded History Available online: https://www.themoscowtimes.com/2021/08/17/russias-2021-wildfires-now-largest-in-its-recorded-history-a74814 (accessed on 31 August 2021).

7. Rohde, R. June 2021 Temperature Update Available online: https://berkeleyearth.org/june-2021-temperature-update/ (accessed on 31 August 2021).

8. IPCC Summary for Policymakers. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, S. Berger, C., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J.B.R., Maycock, T.K., Waterfield, T., Yelekçi, O., Yu, R., Zhou, B., Eds.; Cambridge University Press: Cambridge, 2021; pp.

1–41.

9. World Meterological Organisation *State of the Global Climate 2020*; 2021; ISBN 4146702018922.

10. Hausfather, Z. State of the Climate: 2021 Sees Widespread Climate Extremes despite a Cool Start Available online: https://www.carbonbrief.org/state-of-the-climate-2021-sees-widespread-climate-extremes-despite-a-cool-start (accessed on 31 August 2021).

11. Xu, C.; Kohler, T.A.; Lenton, T.M.; Svenning, J.C.; Scheffer, M. Future of the Human Climate Niche. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, doi:10.1073/pnas.1910114117.

12. IPCC Summary for Policymakers. In *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change,*; Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P.R., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J.B.R., Chen, Y., Zhou, X., Gomis, M.I., Lonnoy, E., Maycock, T., Tignor, M., Waterfield, T., Eds.; 2018; Vol. 2, pp. 17–20.

13. Hansen, J.; Ruedy, R.; Sato, M.; Lo, K. Global Surface Temperature Change. *Rev. Geophys.* **2010**, *48*, RG4004, doi:10.1029/2010RG000345.

14. Schweiger, A.; Lindsay, R.; Zhang, J.; Steele, M.; Stern, H.; Kwok, R. Uncertainty in Modeled Arctic Sea Ice Volume. *J. Geophys. Res. Ocean.* **2011**, *116*, 1–21, doi:10.1029/2011JC007084.

15. Zhang, J.L.; Rothrock, D.A. PIOMAS Arctic Sea Ice Volume Reanalysis Available online: http://psc.apl.uw.edu/research/projects/arctic-sea-ice-volume-anomaly/ (accessed on 31 August 2021).

16. Ehhalt, D.; Prather, M.; Dentener, F.; Derwent, R.; Dlugokencky, E.; Holland, E.; Isaksen, I.; Katima, J.; Kirchhoff, V.; Matson, P.; et al. *Atmospheric Chemistry and Greenhouse Gases*; Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Dai, X., Maskell, K., Johnson, C.A., Eds.; Cambridge University Press: Cambridge, 2001;

17. Lashof, D.A.; Ahuja, D.R. Relative Global Warming Potentials of Greenhouse Gas Emissions. *Nature* **1990**, *344*, 529–531.

18. Ramaswamy, V.; Dentener, F.; Derwent, R.; Dlugokencky, E.; Holland, E.; Isaksen, I.; Katima, J.; Kirchhoff, V.; Matson, P.; Midgley, P.; et al. *Radiative Forcing of Climate Change*; Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Dai, X., Maskell, K., Johnson, C.A., Eds.; Cambridge University Press: Cambridge, 2001;

19. Zillman, J. A History of Climate Activities. *World Meteorol. Organ. Bull.* **2009**, *58*, 141.

20. Peters, G.P.; Andrew, R.M.; Boden, T.; Canadell, J.G.; Ciais, P.; Le Quéré, C.; Marland, G.; Raupach, M.R.; Wilson, C. The Challenge to Keep Global Warming below 2 °C. *Nat. Clim. Chang.* **2013**, *3*, 4–6, doi:10.1038/nclimate1783.

21. Stockwell, C.; Geiges, A.; Ramalope, D.; Gidden, M.; Hare, B.; Fekete, H.; Gonzales-Zuñiga, S.; Hans, F.; Höhne, N. *Climate Summit Momentum: Paris Commitments Improved Warming Estimate to 2.4°C*; 2021;

22. Rockström, J.; Steffen, W.; Noone, K.; Persson, Å.; Chapin, F.S.; Lambin, E.F.; Lenton, T.M.; Scheffer, M.; Folke, C.; Schellnhuber, H.J.; et al. A Safe Operating Space for Humanity. *Nature* **2009**, *461*, 472–475, doi:10.1038/461472a.

23. Steffen, W.; Richardson, K.; Rockström, J.; Cornell, S.E.; Fetzer, I.; Bennett, E.M.; Biggs, R.; Carpenter, S.R.; De Vries, W.; De Wit, C.A.; et al. Planetary Boundaries: Guiding Human Development on a Changing Planet. *Science (80-. ).* **2015**, *347*, doi:10.1126/science.1259855.

24. Montoya, J.M.; Donohue, I.; Pimm, S.L. Why a Planetary Boundary, If It Is Not Planetary, and the Boundary Is Undefined? A Reply to Rockström et Al. *Trends Ecol. Evol.* **2018**, *33*, 234, doi:10.1016/j.tree.2018.01.008.

25. Montoya, J.M.; Donohue, I.; Pimm, S.L. Planetary Boundaries for Biodiversity: Implausible Science, Pernicious Policies. *Trends Ecol. Evol.* **2018**, *33*, 71–73, doi:10.1016/j.tree.2017.10.004.

26. Rockström, J.; Richardson, K.; Steffen, W.; Mace, G. Planetary Boundaries: Separating Fact from Fiction. A Response to Montoya et Al. *Trends Ecol. Evol.* **2018**, *33*, 233–234, doi:10.1016/j.tree.2018.01.010.

27. Downing, A.S.; Chang, M.; Kuiper, J.J.; Campenni, M.; Haÿhä, T.; Cornell, S.E.; Svedin, U.; Mooij, W.M. Learning from Generations of Sustainability Concepts. *Environ. Res. Lett.* **2020**, *15*, doi:10.1088/1748-9326/ab7766.

28. De Vries, W.; Kros, J.; Kroeze, C.; Seitzinger, S.P. Assessing Planetary and Regional Nitrogen Boundaries Related to Food Security and Adverse Environmental Impacts. *Curr. Opin. Environ. Sustain.* **2013**, *5*, 392–402, doi:10.1016/j.cosust.2013.07.004.

29. Mace, G.M.; Reyers, B.; Alkemade, R.; Biggs, R.; Chapin, F.S.; Cornell, S.E.; Díaz, S.; Jennings, S.; Leadley, P.; Mumby, P.J.; et al. Approaches to Defining a Planetary Boundary for Biodiversity. *Glob. Environ. Chang.* **2014**, *28*, 289–297, doi:10.1016/j.gloenvcha.2014.07.009.

30. Li, M.; Wiedmann, T.; Fang, K.; Hadjikakou, M. The Role of Planetary Boundaries in Assessing Absolute Environmental Sustainability across Scales. *Environ. Int.* **2021**, *152*, doi:10.1016/j.envint.2021.106475.

31. Isbell, F.; Craven, D.; Connolly, J.; Loreau, M.; Schmid, B.; Beierkuhnlein, C.; Bezemer, T.M.; Bonin, C.; Bruelheide, H.; De Luca, E.; et al. Biodiversity Increases the Resistance of Ecosystem Productivity to Climate Extremes. *Nature* **2015**, *526*, 574–577, doi:10.1038/nature15374.

32. Sponsel, L.E. Human Impact on Biodiversity, Overview. *Encycl. Biodivers. Second Ed.* **2013**, *4*, 137–152, doi:10.1016/B978-0-12-384719-5.00250-1.

33. Gitay, H.; Suárez, A.; Watson, R.; Anisimov, O.; Chapin, F.S.; Cruz, R.V.; Finlayson, M.; Hohenstein, W.; Insarov, G.; Kundzewicz, B.; et al. *Climate Change and Biodiversity*; Gitay, H., Suárez, A., Watson, R., Dokken, D., Eds.; IPCC: Geneva, 2002; Vol. 5;.

34. Thushari, G.G.N.; Senevirathna, J.D.M. Plastic Pollution in the Marine Environment. *Heliyon* **2020**, *6*, e04709, doi:10.1016/j.heliyon.2020.e04709.

35. Carpenter, S.R.; Bennett, E.M. Reconsideration of the Planetary Boundary for Phosphorus. *Environ. Res. Lett.* **2011**, *6*, doi:10.1088/1748-9326/6/1/014009.

36. Kopittke, P.M.; Menzies, N.W.; Dalal, R.C.; McKenna, B.A.; Husted, S.; Wang, P.; Lombi, E. The Role of Soil in Defining Planetary Boundaries and the Safe Operating Space for Humanity. *Environ. Int.* **2021**, *146*, doi:10.1016/j.envint.2020.106245.

37. Fowler, D.; Coyle, M.; Skiba, U.; Sutton, M.A.; Cape, J.N.; Reis, S.; Sheppard, L.J.;

Jenkins, A.; Grizzetti, B.; Galloway, J.N.; et al. The Global Nitrogen Cycle in the Twentyfirst Century. *Philos. Trans. R. Soc. B Biol. Sci.* **2013**, *368*, doi:10.1098/rstb.2013.0164.

38. Engström, G.; Gars, J.; Krishnamurthy, C.; Spiro, D.; Calel, R.; Lindahl, T.; Narayanan, B. Carbon Pricing and Planetary Boundaries. *Nat. Commun.* **2020**, *11*, 1–11, doi:10.1038/s41467-020-18342-7.

39. Dasgupta, P. *The Economics of Biodiversity: The Dasgupta Review Abridged Version*; 2021; ISBN 9781911680307.

40. Sala, S.; Crenna, E.; Secchi, M.; Sanyé-Mengual, E. Environmental Sustainability of European Production and Consumption Assessed against Planetary Boundaries. *J. Environ. Manage.* **2020**, *269*, doi:10.1016/j.jenvman.2020.110686.

41. Nash, K.L.; Cvitanovic, C.; Fulton, E.A.; Halpern, B.S.; Milner-Gulland, E.J.; Watson, R.A.; Blanchard, J.L. Planetary Boundaries for a Blue Planet. *Nat. Ecol. Evol.* **2017**, *1*, 1625–1634, doi:10.1038/s41559-017-0319-z.

42. Algunaibet, I.M.; Pozo, C.; Galán-Martín, Á.; Huijbregts, M.A.J.; Dowell, N. Mac; Guillén-Gosálbez, G. Powering Sustainable Development within Planetary Boundaries. *Energy Environ. Sci.* **2019**, *12*, 1890–1900, doi:10.1039/c8ee03423k.

43. Akrofi, M.M. An Analysis of Energy Diversification and Transition Trends in Africa. *Int. J. Energy Water Resour.* **2021**, *5*, 1–12, doi:10.1007/s42108-020-00101-5.

44. Azzuni, A.; Breyer, C. Global Energy Security Index and Its Application on National Level. *Energies* **2020**, *13*, doi:10.3390/en13102502.

45. Bluszcz, A. European Economies in Terms of Energy Dependence. *Qual. Quant.* **2017**, *51*, 1531–1548, doi:10.1007/s11135-016-0350-1.

46. Karatayev, M.; Hall, S. Establishing and Comparing Energy Security Trends in Resource-Rich Exporting Nations (Russia and the Caspian Sea Region). *Resour. Policy* **2020**, *68*, 101746, doi:10.1016/j.resourpol.2020.101746.

47. Agencia Internacional de la Energía Key World Energy Statistics 2020. *Int. Energy Agency* **2020**, *33*, 4649.

48. International Energy Agency *The Future of Petrochemicals – Analysis*; 2018;

49. Colgan, J.D. Fueling the Fire: Pathways from Oil to War. *Int. Secur.* **2013**, *38*, 147–180, doi:10.1162/ISEC_a_00135.

50. San-Akca, B.; Sever, S.D.; Yilmaz, S. Does Natural Gas Fuel Civil War? Rethinking Energy Security, International Relations, and Fossil-Fuel Conflict. *Energy Res. Soc. Sci.* **2020**, *70*, 101690, doi:10.1016/j.erss.2020.101690.

51. Kettell, S. Oil Crisis Available online: https://www.britannica.com/topic/oil-crisis (accessed on 1 September 2021).

52. Macalister, T. Background: What Caused the 1970s Oil Price Shock? Available online: https://www.theguardian.com/environment/2011/mar/03/1970s-oil-price-shock (accessed on 1 September 2021).

53. Correljé, A.; van der Linde, C. Energy Supply Security and Geopolitics: A European Perspective. *Energy Policy* **2006**, *34*, 532–543, doi:10.1016/j.enpol.2005.11.008.

54. Ritchie, H.; Roser, M. Energy Available online: https://ourworldindata.org/energy (accessed on 2 September 2021).

55. Buck, M.; Redl, C.; Hein, F.; Jones, D. *The European Power Sector in 2020: Up-to-Date Analysis on the Electricity Transition*; 2021;

56. BP p.l.c *Statistical Review of World Energy 2021*; London, 2021; Vol. 70;.

57. Dukes *Digest of UK Energy Statistics (Chapter 6)*; 2020;

58. Organisation for Economic Co-operation and Development *Industrial Biotechnology and Climate Change*; 2011;

59. Luque, R.; Herrero-Davila, L.; Campelo, J.M.; Clark, J.H.; Hidalgo, J.M.; Luna, D.; Marinas, J.M.; Romero, A.A. Biofuels: A Technological Perspective. *Energy Environ. Sci.* **2008**, doi:10.1039/b807094f.

60. Burk, M.J.; Van Dien, S. Biotechnology for Chemical Production: Challenges and Opportunities. *Trends Biotechnol.* **2016**, *34*, 187–190, doi:10.1016/j.tibtech.2015.10.007.

61. De Jong, E.; Stichnothe, H.; Bell, G.; Jorgensen, H. *Bio-Based Chemicals: A 2020 Update*; IEA Bioenergy, 2020; ISBN 9781910154694.

62. Lee, U.; Kwon, H.; Wu, M.; Wang, M. Retrospective Analysis of the U.S. Corn Ethanol

Industry for 2005–2019: Implications for Greenhouse Gas Emission Reductions. *Biofuels, Bioprod. Biorefining* **2021**, 1318–1331, doi:10.1002/bbb.2225.

63. Flugge, M.; Lewandrowski, J.; Rosenfeld, J.; Boland, C.; Hendrickson, T.; Jaglo, K.; Kolansky, S.; Moffroid, K.; Riley-Gilbert, M.; Pape, D. *A Life-Cycle Analysis of the Greenhouse Gas Emissions of Corn-Based Ethanol*; 2017;

64. Oxfam *Another Inconvenient Truth: How Biofuel Policies Are Deepening Poverty and Accelerating Climate Change*; 2008; Vol. 114;.

65. Fargione, J.; Hill, J.; Tilman, D.; Polasky, S.; Hawthorne, P. Land Clearing and the Biofuel Carbon Debt. *Science (80-. ).* **2008**, *319*, 1235–1238, doi:10.1126/science.1152747.

66. Jeswani, H.K.; Chilvers, A.; Azapagic, A. Environmental Sustainability of Biofuels: A Review. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2020**, *476*, doi:10.1098/rspa.2020.0351.

67. Searchinger, T.; Heimlich, R.; Houghton, R.A.; Dong, F.; Elobeid, A.; Fabiosa, J.; Tokgoz, S.; Hayes, D.; Yu, T.-H. Use of U.S. Croplands for Biofuels Increases Greenhouse Gases Through Emissions from Land-Use Change. *Science (80-. ).* **2008**, *319*, 1238–1240, doi:10.1126/science.1151861.

68. An, H.; Qiu, F.; Rude, J. Volatility Spillovers between Food and Fuel Markets: Do Administrative Regulations Affect the Transmission? *Econ. Model.* **2021**, *102*, 105552, doi:10.1016/j.econmod.2021.105552.

69. Vural Gursel, I.; Moretti, C.; Hamelin, L.; Jakobsen, L.G.; Steingrimsdottir, M.M.; Junginger, M.; Høibye, L.; Shen, L. Comparative Cradle-to-Grave Life Cycle Assessment of Bio-Based and Petrochemical PET Bottles. *Sci. Total Environ.* **2021**, *793*, 148642, doi:10.1016/j.scitotenv.2021.148642.

70. Yang, J.; Xu, M.; Zhang, X.; Hu, Q.; Sommerfeld, M.; Chen, Y. Life-Cycle Analysis on Biodiesel Production from Microalgae: Water Footprint and Nutrients Balance. *Bioresour. Technol.* **2011**, doi:10.1016/j.biortech.2010.07.017.

71. Policy Analysis Life-Cycle Assessment of Biodiesel Production from Microalgae., doi:10.1021/es900705j.

72. Kalmykova, Y.; Sadagopan, M.; Rosado, L. Circular Economy - From Review of Theories and Practices to Development of Implementation Tools. *Resour. Conserv.*

*Recycl.* **2018**, *135*, 190–201, doi:10.1016/j.resconrec.2017.10.034.

73. Phillips, J.; Huhnke, R.; Atiyeh, H. Syngas Fermentation: A Microbial Conversion Process of Gaseous Substrates to Various Products. *Fermentation* **2017**, *3*, 28, doi:10.3390/fermentation3020028.

74. Teixeira, L. V.; Moutinho, L.F.; Romão-Dumaresq, A.S. Gas Fermentation of C1 Feedstocks: Commercialization Status and Future Prospects. *Biofuels, Bioprod. Biorefining* **2018**, *12*, 1103–1117, doi:10.1002/bbb.1912.

75. Köpke, M.; Simpson, S.D. Pollution to Products: Recycling of 'above Ground' Carbon by Gas Fermentation. *Curr. Opin. Biotechnol.* **2020**, *65*, 180–189, doi:10.1016/j.copbio.2020.02.017.

76. Sun, X.; Atiyeh, H.K.; Huhnke, R.L.; Tanner, R.S. Syngas Fermentation Process Development for Production of Biofuels and Chemicals: A Review. *Bioresour. Technol. Reports* **2019**, *7*, 100279, doi:10.1016/j.biteb.2019.100279.

77. Griffin, D.W.; Schultz, M.A. Fuel and Chemical Products from Biomass Syngas: A Comparison of Gas Fermentation to Thermochemical Conversion Routes. *Environ. Prog. Sustain. Energy* **2012**, *31*, 219–224, doi:10.1002/ep.11613.

78. Mahinpey, N.; Gomez, A. Review of Gasification Fundamentals and New Findings: Reactors, Feedstock, and Kinetic Studies. *Chem. Eng. Sci.* **2016**, *148*, 14–31, doi:10.1016/j.ces.2016.03.037.

79. Molitor, B.; Richter, H.; Martin, M.E.; Jensen, R.O.; Juminaga, A.; Mihalcea, C.; Angenent, L.T. Carbon Recovery by Fermentation of CO-Rich off Gases – Turning Steel Mills into Biorefineries. *Bioresour. Technol.* **2016**, *215*, 386–396, doi:10.1016/j.biortech.2016.03.094.

80. Simon, M.; Frey, R.; Zangemeister-Wittke, U.; Plückthun, A. Orthogonal Assembly of a Designed Ankyrin Repeat Protein-Cytotoxin Conjugate with a Clickable Serum Albumin Module for Half-Life Extension. *Bioconjug. Chem.* **2013**, doi:10.1021/bc4004102.

81. Yasin, M.; Cha, M.; Chang, I.S.; Atiyeh, H.K.; Munasinghe, P.; Khanal, S.K. *Syngas Fermentation into Biofuels and Biochemicals*; 2nd ed.; Elsevier Inc., 2019; ISBN 9780128168561.

82. Min, F.; Kopke, M.; Dennis, S. Gas Fermentation for Commercial Biofuels Production. In *Liquid, Gaseous and Solid Biofuels - Conversion Techniques*; 2013; pp. 125–173 ISBN 978-953-51-1050-7.

83. *Industrialization of Biology A Roadmap to Accelerate the Advanced Manufacturing of Chemicals*; 2015; Vol. 4;.

84. Nagaraju, S.; Davies, N.K.; Walker, D.J.F.; Köpke, M.; Simpson, S.D. Genome Editing of Clostridium Autoethanogenum Using CRISPR/Cas9. *Biotechnol. Biofuels* **2016**, *9*, doi:10.1186/s13068-016-0638-3.

85. Gyulev, I.S.; Willson, B.J.; Hennessy, R.C.; Krabben, P.; Jenkinson, E.R.; Thomas, G.H. Part by Part: Synthetic Biology Parts Used in Solventogenic Clostridia. *ACS Synth. Biol.* **2018**, *7*, 311–327, doi:10.1021/acssynbio.7b00327.

86. Smith, P.; Nkem, J.; Calvin, K.; Campbell, D.; Cherubini, F.; Grassi, G.; Korotkov, V.; Hoang, A.L.; Lwasa, S.; McElwee, P.; et al. 9: Interlinkages Between Desertification, Land Degradation, Food Security and Greenhouse Gas Fluxes: Synergies, Trade-Offs and Integrated Response Options. In *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*; Shukla, P.R., Skea, J., Calvo Buendia, E., Masson-Delmotte, V., Portner, H.-O., Roberts, D.C., Zhai, P., Slade, R., Connors, S., van Diemen, R., Ferrat, M., Haughey, E., Luz, S., Neogi, S., Pathak, M., Petzold, J., Portugal Pereira, J., Vyas, P., Huntley, E., Kissick, K., Belkacemi, M., Malley, J., Eds.; 2019; pp. 580–583.

87. Rogelj, J.; Shindell, D.; Jiang, K.; Fifita, S.; Forster, P.; Ginzburg, V.; Handa, C.; Kheshgi, H.; Kobayashi, S.; Kriegler, E.; et al. Mitigation Pathways Compatible with 1.5°C in the Context of Sustainable Development. In *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change,*; Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P.R., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J.B.R., Chen, X. Zhou, Y., Gomis, M.I., Lonnoy, E., Maycock, T., Tignor, M., Waterfield, T., Eds.; 2018; p. 96.

88. Fuhrman, J.; McJeon, H.; Patel, P.; Doney, S.C.; Shobe, W.M.; Clarens, A.F. Food–Energy–Water Implications of Negative Emissions Technologies in a +1.5 °C Future. *Nat. Clim. Chang.* **2020**, *10*, 920–927, doi:10.1038/s41558-020-0876-z.

89. Lenzi, D.; Lamb, W.F.; Hilaire, J.; Kowarsch, M.; Minx, J.C. Weigh the Ethics of Plans to Mop up Carbon Dioxide. *Nature* **2018**, *561*, 303–305.

90. Berg, I.A. Ecological Aspects of the Distribution of Different Autotrophic CO2 Fixation Pathways. *Appl. Environ. Microbiol.* **2011**, *77*, 1925–1936, doi:10.1128/AEM.02473-10.

91. Sánchez-Andrea, I.; Guedes, I.A.; Hornung, B.; Boeren, S.; Lawson, C.E.; Sousa, D.Z.; Bar-Even, A.; Claassens, N.J.; Stams, A.J.M. The Reductive Glycine Pathway Allows Autotrophic Growth of Desulfovibrio Desulfuricans. *Nat. Commun.* **2020**, *11*, 1–12, doi:10.1038/s41467-020-18906-7.

92. Bengelsdorf, F.R.; Beck, M.H.; Erz, C.; Hoffmeister, S.; Karl, M.M.; Riegler, P.; Wirth, S.; Poehlein, A.; Weuster-Botz, D.; Dürre, P. Bacterial Anaerobic Synthesis Gas (Syngas) and CO2 + H2 Fermentation. *Adv. Appl. Microbiol.* **2018**, *103*, 143–221, doi:10.1016/bs.aambs.2018.01.002.

93. Valgepea, K.; Loi, K.Q.; Behrendorff, J.B.; Lemgruber, R. de S.P.; Plan, M.; Hodson, M.P.; Köpke, M.; Nielsen, L.K.; Marcellin, E. Arginine Deiminase Pathway Provides ATP and Boosts Growth of the Gas-Fermenting Acetogen Clostridium Autoethanogenum. *Metab. Eng.* **2017**, *41*, 202–211, doi:10.1016/j.ymben.2017.04.007.

94. Köpke, M.; Mihalcea, C.; Liew, F.M.; Tizard, J.H.; Ali, M.S.; Conolly, J.J.; Al-Sinawi, B.; Simpson, S.D. 2,3-Butanediol Production by Acetogenic Bacteria, an Alternative Route to Chemical Synthesis, Using Industrial Waste Gas. *Appl. Environ. Microbiol.* **2011**, *77*, 5467–5475, doi:10.1128/AEM.00355-11.

95. Abrini, J.; Naveau, H.; Nyns, E.-J. Clostridium Autoethanogenum, Sp. Nov., an Anaerobic Bacterium That Produces Ethanol from Carbon Monoxide. *Arch. Microbiol.* **1994**, *161*, 345–351, doi:10.1007/BF00303591.

96. Ragsdale, S.W.; Pierce, E. Acetogenesis and the Wood-Ljungdahl Pathway of CO2 Fixation. *Biochim. Biophys. Acta - Proteins Proteomics* **2008**, *1784*, 1873–1898, doi:10.1016/j.bbapap.2008.08.012.

97. Wang, S.; Huang, H.; Kahnt, H.H.; Mueller, A.P.; Köpke, M.; Thauer, R.K. NADP-Specific Electron-Bifurcating [FeFe]-Hydrogenase in a Functional Complex with Formate Dehydrogenase in Clostridium Autoethanogenum Grown on CO. *J. Bacteriol.* **2013**, *195*, 4373–4386, doi:10.1128/JB.00678-13.

98. Xu, H.; Liang, C.; Yuan, Z.; Xu, J.; Hua, Q.; Guo, Y. A Study of CO/Syngas Bioconversion by Clostridium Autoethanogenum with a Flexible Gas-Cultivation System. *Enzyme Microb. Technol.* **2017**, *101*, 24–29, doi:10.1016/j.enzmictec.2017.03.002.

99. Bertsch, J.; Müller, V. Bioenergetic Constraints for Conversion of Syngas to Biofuels in Acetogenic Bacteria. *Biotechnol. Biofuels* **2015**, doi:10.1186/s13068-015-0393-x.

100. Peters, J.W.; Miller, A.F.; Jones, A.K.; King, P.W.; Adams, M.W.W. Electron Bifurcation. *Curr. Opin. Chem. Biol.* **2016**, *31*, 146–152, doi:10.1016/j.cbpa.2016.03.007.

101. Buckel, W.; Thauer, R.K. Flavin-Based Electron Bifurcation, Ferredoxin, Flavodoxin, and Anaerobic Respiration With Protons (Ech) or NAD+ (Rnf) as Electron Acceptors: A Historical Review. *Front. Microbiol.* **2018**, *9*, doi:10.3389/fmicb.2018.00401.

102. Katsyv, A.; Müller, V. Overcoming Energetic Barriers in Acetogenic C1 Conversion. *Front. Bioeng. Biotechnol.* **2020**, *8*, 1–23, doi:10.3389/fbioe.2020.621166.

103. Buckel, W.; Thauer, R.K. Energy Conservation via Electron Bifurcating Ferredoxin Reduction and Proton/Na+ Translocating Ferredoxin Oxidation. *Biochim. Biophys. Acta - Bioenerg.* **2013**, *1827*, 94–113, doi:10.1016/j.bbabio.2012.07.002.

104. Schuchmann, K.; Müller, V. Energetics and Application of Heterotrophy in Acetogenic Bacteria. *Appl. Environ. Microbiol.* **2016**, *82*, 4056–4069, doi:10.1128/AEM.00882-16.

105. Marcellin, E.; Behrendorff, J.B.; Nagaraju, S.; DeTissera, S.; Segovia, S.; Palfreyman, R.W.; Daniell, J.; Licona-Cassani, C.; Quek, L.-E.; Speight, R.; et al. Low Carbon Fuels and Commodity Chemicals from Waste Gases – Systematic Approach to Understand Energy Metabolism in a Model Acetogen. *Green Chem.* **2016**, *18*, 3020–3028, doi:10.1039/C5GC02708J.

106. Abubackar, H.N.; Fernández-Naveira, Á.; Veiga, M.C.; Kennes, C. Impact of Cyclic PH

Shifts on Carbon Monoxide Fermentation to Ethanol by Clostridium Autoethanogenum. *Fuel* **2016**, *178*, 56–62, doi:10.1016/J.FUEL.2016.03.048.

107. Mock, J.; Zheng, Y.; Mueller, A.P.; Ly, S.; Tran, L.; Segovia, S.; Nagaraju, S.; Köpke, M.; Dürre, P.; Thauer, R.K. Energy Conservation Associated with Ethanol Formation from H 2 and CO 2 in Clostridium Autoethanogenum Involving Electron Bifurcation. *J. Bacteriol.* **2015**, *197*, 2965–2980, doi:10.1128/JB.00399-15.

108. Liew, F.; Henstra, A.M.; Winzer, K.; Köpke, M.; Simpson, S.D.; Minton, N.P. Insights into CO2 Fixation Pathway of Clostridium Autoethanogenum by Targeted Mutagenesis. *MBio* **2016**, doi:10.1128/mBio.00427-16.

109. Weghoff, M.C.; Bertsch, J.; Müller, V. A Novel Mode of Lactate Metabolism in Strictly Anaerobic Bacteria. *Environ. Microbiol.* **2015**, *17*, 670–677, doi:10.1111/1462-2920.12493.

110. Köpke, M.; Gerth, M.L.; Maddock, D.J.; Mueller, A.P.; Liew, F.M.; Simpson, S.D.; Patrick, W.M. Reconstruction of an Acetogenic 2,3-Butanediol Pathway Involving a Novel NADPH-Dependent Primary-Secondary Alcohol Dehydrogenase. *Appl. Environ. Microbiol.* **2014**, *80*, 3394–3303, doi:10.1128/AEM.00301-14.

111. Fu, Y.; Zhu, Y.; Zhang, K.; Yeung, M.T.; Durocher, D.; Xiao, W. Rad6-Rad18 Mediates a Eukaryotic SOS Response by Ubiquitinating the 9-1-1 Checkpoint Clamp. *Cell* **2008**, *133*, 601–611, doi:10.1016/j.cell.2008.02.050.

112. Maslowska, K.H.; Makiela-Dzbenska, K.; Fijalkowska, I.J. The SOS System: A Complex and Tightly Regulated Response to DNA Damage. *Environ. Mol. Mutagen.* **2019**, *60*, 368–384, doi:10.1002/em.22267.

113. Banach-Orlowska, M.; Fijalkowska, I.J.; Schaaper, R.M.; Jonczyk, P. DNA Polymerase II as a Fidelity Factor in Chromosomal DNA Synthesis in Escherichia Coli. *Mol. Microbiol.* **2005**, *58*, 61–70, doi:10.1111/j.1365-2958.2005.04805.x.

114. Fujii, S.; Isogawa, A.; Fuchs, R.P. Chronological Switch from Translesion Synthesis to Homology- Dependent Gap Repair in Vivo. *Toxicol. Res.* **2018**, *34*, 297–302, doi:10.5487/TR.2018.34.4.297.

115. Fujii, S.; Fuchs, R.P. A Comprehensive View of Translesion Synthesis in Escherichia Coli . *Microbiol. Mol. Biol. Rev.* **2020**, *84*, 1–29, doi:10.1128/mmbr.00002-20.

116. Nickoloff, J.A.; Hoekstra, M.F.; Hatahet, Z.; Wallace, S.S. Translesion DNA Synthesis. *DNA Damage Repair* **2003**, *5*, 229–262, doi:10.1385/0-89603-356-2:229.

117. Vaisman, A.; Woodgate, R. Translesion DNA Polymerases in Eukaryotes: What Makes Them Tick? *Crit. Rev. Biochem. Mol. Biol.* **2017**, *52*, 274–303, doi:10.1080/10409238.2017.1291576.

118. Hedglin, M.; Benkovic, S.J. Eukaryotic Translesion DNA Synthesis on the Leading and Lagging Strands: Unique Detours around the Same Obstacle. *Chem. Rev.* **2017**, *117*, 7857–7877, doi:10.1021/acs.chemrev.7b00046.

119. Li, G.M. Mechanisms and Functions of DNA Mismatch Repair. *Cell Res.* 2008, *18*, 85–98.

120. Spampinato, C.P.; Gomez, R.L.; Galles, C.; Lario, L.D. From Bacteria to Plants: A Compendium of Mismatch Repair Assays. *Mutat. Res. Mutat. Res.* **2009**, *682*, 110–128, doi:10.1016/j.mrrev.2009.07.001.

121. Fukui, K. DNA Mismatch Repair in Eukaryotes and Bacteria. *J. Nucleic Acids* 2010, *2010*, 1–16.

122. Modrich, P. Mechanisms in Eukaryotic Mismatch Repair. *J. Biol. Chem.* **2006**, *281*, 30305–30309, doi:10.1074/jbc.R600022200.

123. Williams, J.S.; Kunkel, T.A. Ribonucleotides in DNA: Origins, Repair and Consequences. *DNA Repair (Amst).* **2014**, *19*, 27–37, doi:10.1016/j.dnarep.2014.03.029.

124. Parsons, J.L.; Edmonds, M.J. The Base Excision Repair Pathway. *Encycl. Cell Biol.* **2016**, *1*, 442–450, doi:10.1016/B978-0-12-394447-4.10046-X.

125. Weller, G.R.; Kysela, B.; Roy, R.; Tonkin, L.M.; Scanlan, E.; Della, M.; Devine, S.K.; Day, J.P.; Wilkinson, A.; D'Adda di Fagagna, F.; et al. Identification of a DNA Nonhomologous End-Joining Complex in Bacteria. *Science (80-. ).* **2002**, *297*, 1686–1689, doi:10.1126/science.1074584.

126. Pitcher, R.S.; Brissett, N.C.; Doherty, A.J. Nonhomologous End-Joining in Bacteria: A Microbial Perspective. *Annu. Rev. Microbiol.* **2007**, *61*, 259–282, doi:10.1146/annurev.micro.61.080706.093354.

127. Dupuy, P.; Sauviac, L.; Bruand, C. Stress-Inducible NHEJ in Bacteria: Function in DNA Repair and Acquisition of Heterologous DNA. *Nucleic Acids Res.* **2019**, *47*, 1335–1349, doi:10.1093/nar/gky1212.

128. Ceccaldi, R.; Rondinelli, B.; D 'andrea, A.D. Repair Pathway Choices and Consequences at the Double- Strand Break Mechanisms of DNA DSB Repair. *Trends Biochem. Sci.* **2016**, *26*, 52–64, doi:10.1016/j.tcb.2015.07.009.

129. Anzalone, A. V.; Koblan, L.W.; Liu, D.R. Genome Editing with CRISPR–Cas Nucleases, Base Editors, Transposases and Prime Editors. *Nat. Biotechnol.* 2020, *38*, 824–844.

130. Huang, C.; Ding, T.; Wang, J.; Wang, X.; Guo, L.; Wang, J.; Zhu, L.; Bi, C.; Zhang, X.; Ma, X.; et al. CRISPR-Cas9-Assisted Native End-Joining Editing Offers a Simple Strategy for Efficient Genetic Engineering in Escherichia Coli. *Appl. Microbiol. Biotechnol.* **2019**, *103*, 8497–8509, doi:10.1007/s00253-019-10104-w.

131. Chayot, R.; Montagne, B.; Mazel, D.; Ricchetti, M. An End-Joining Repair Mechanism in Escherichia Coli. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 2141–2146, doi:10.1073/pnas.0906355107.

132. Finger-Bou, M.; Orsi, E.; van der Oost, J.; Staals, R.H.J. CRISPR with a Happy Ending: Non-Templated DNA Repair for Prokaryotic Genome Engineering. *Biotechnol. J.* **2020**, *1900404*, e1900404, doi:10.1002/biot.201900404.

133. Wyman, C.; Kanaar, R. DNA Double-Strand Break Repair: All's Well That Ends Well. *Annu. Rev. Genet.* **2006**, *40*, 363–383, doi:10.1146/annurev.genet.40.110405.090451.

134. Yang, K.; Guo, R.; Xu, D. Non-Homologous End Joining: Advances and Frontiers. *Acta Biochim. Biophys. Sin. (Shanghai).* **2016**, *48*, 632–640, doi:10.1093/abbs/gmw046.

135. Rass, E.; Grabarz, A.; Bertrand, P.; Lopez, B.S. Réparation Des Cassures Double-Brin de l'ADN, Un Mécanisme Peut En Cacher Un Autre: La Ligature d'extrémités Non Homologues Alternative. *Cancer/Radiotherapie* **2012**, *16*, 1–10, doi:10.1016/j.canrad.2011.05.004.

136. Bhargava, R.; Onyango, D.O.; Stark, J.M. Regulation of Single Strand Annealing. **2017**, *32*, 566–575, doi:10.1016/j.tig.2016.06.007.Regulation.

137. Aravind, L.; Koonin, E. V. Prokaryotic Homologs of the Eukaryotic DNA-End-Binding

Protein Ku, Novel Domains in the Ku Protein and Prediction of a Prokaryotic Double-Strand Break Repair System. *Genome Res.* **2001**, *11*, 1365–1374, doi:10.1101/gr.181001.

138. Bowater, R.; Doherty, A.J. Making Ends Meet: Repairing Breaks in Bacterial DNA by Non-Homologous End-Joining. *PLoS Genet.* **2006**, *2*, 93–99, doi:10.1371/journal.pgen.0020008.

139. Shuman, S.; Glickman, M.S. Bacterial DNA Repair by Non-Homologous End Joining. *Nat. Rev. Microbiol.* **2007**, *5*, 852–861, doi:10.1038/nrmicro1768.

140. Dillingham, M.S.; Kowalczykowski, S.C. RecBCD Enzyme and the Repair of Double-Stranded DNA Breaks. *Microbiol. Mol. Biol. Rev.* **2008**, *72*, 642–671, doi:10.1128/mmbr.00020-08.

141. Smith, G.R. How RecBCD Enzyme and Chi Promote DNA Break Repair and Recombination: A Molecular Biologist's View. *Microbiol. Mol. Biol. Rev.* **2012**, *76*, 217–228, doi:10.1128/mmbr.05026-11.

142. Liu, M.; Rehman, S.; Tang, X.; Gu, K.; Fan, Q.; Chen, D.; Ma, W. Methodologies for Improving HDR Efficiency. *Front. Genet.* 2019, *10*.

143. Verma, P.; Greenberg, R.A. Noncanonical Views of Homology-Directed DNA Repair. *Genes Dev.* **2016**, *30*, 1138–1154, doi:10.1101/gad.280545.116.

144. Davis, L.; Maizels, N. Homology-Directed Repair of DNA Nicks via Pathways Distinct from Canonical Double-Strand Break Repair. *Proc. Natl. Acad. Sci.* **2014**, *111*, E924–E932, doi:10.1073/pnas.1400236111.

145. Maher, R.L.; Branagan, A.M.; Morrical, S.W. Coordination of DNA Replication and Recombination Activities in the Maintenance of Genome Stability. *J. Cell. Biochem.* **2011**, *112*, 2672–2682, doi:10.1002/jcb.23211.

146. Pâques, F.; Haber, J.E. Multiple Pathways of Recombination Induced by Double-Strand Breaks in Saccharomyces Cerevisiae. *Microbiol. Mol. Biol. Rev.* **1999**, *63*, 349–404.

147. Rosenberg, S.M.; Hastings, P.J. The Split-End Model for Homologous Recombination at Double-Strand Breaks and at Chi. *Biochimie* **1991**, *73*, 385–397, doi:10.1016/0300-9084(91)90105-A.

148. Smith, G.R. RecBCD Enzyme and Pathway. In *Reference Module in Life Sciences*; Elsevier, 2017.

149. Smith, G.R. Chi Sequences. *Ref. Modul. Life Sci.* **2017**, 1–5, doi:10.1016/b978-0-12-809633-8.06196-3.

150. Petrosino, J.F.; Hastings, P.J.; Rosenberg, S.M. RecBCD Enzyme, Pathway. In *Encyclopedia of Genetics*; Elsevier, 2001; pp. 1623–1631.

151. Mawer, J.S.P.; Leach, D.R.F. Branch Migration Prevents DNA Loss during Double-Strand Break Repair. *PLoS Genet.* **2014**, *10*, doi:10.1371/journal.pgen.1004485.

152. Michel, B.; Sinha, A.K.; Leach, D.R.F. Replication Fork Breakage and Restart in Escherichia Coli . *Microbiol. Mol. Biol. Rev.* **2018**, *82*, 1–19, doi:10.1128/mmbr.00013-18.

153. Zerbib, D.; Mézard, C.; George, H.; West, S.C. Coordinated Actions of RuvABC in Holliday Junction Processing. *J. Mol. Biol.* **1998**, *281*, 621–630, doi:10.1006/jmbi.1998.1959.

154. Van Gool, A.J.; Hajibagheri, N.M.A.; Stasiak, A.; West, S.C. Assembly of the Escherichia Coli RuvABC Resolvasome Directs the Orientation of Holliday Junction Resolution. *Genes Dev.* **1999**, *13*, 1861–1870, doi:10.1101/gad.13.14.1861.

155. Azeroglu, B.; Mawer, J.S.P.; Cockram, C.A.; White, M.A.; Hasan, A.M.M.; Filatenkova, M.; Leach, D.R.F. RecG Directs DNA Synthesis during Double-Strand Break Repair. *PLoS Genet.* **2016**, *12*, 1–23, doi:10.1371/journal.pgen.1005799.

156. Tanaka, T.; Masai, H. Stabilization of a Stalled Replication Fork by Concerted Actions of Two Helicases. *J. Biol. Chem.* **2006**, *281*, 3484–3493, doi:10.1074/jbc.M510979200.

157. Kogoma, T.; Cadwell, G.W.; Barnard, K.G.; Asai, T. The DNA Replication Priming Protein, PriA, Is Required for Homologous Recombination and Double-Strand Break Repair. *J. Bacteriol.* **1996**, *178*, 1258–1264, doi:10.1128/jb.178.5.1258-1264.1996.

158. Lee, C.; Kamal, S.M.; Römling, U. High Frequency of Double Crossover Recombination Facilitates Genome Engineering in Pseudomonas Aeruginosa PA14 and Clone C Strains. *Microbiol. (United Kingdom)* **2019**, *165*, 757–760, doi:10.1099/mic.0.000812.

159. Rédei, G.P. Polar Mutation. *Encycl. Genet. Genomics, Proteomics Informatics* 2008, 1519–1519.

160. Rédei, G.P. Operon. *Encycl. Genet. Genomics, Proteomics Informatics* 2008, 1399–1399.

161. Margolin, P. Genetic Polarity in Bacteria. *Am. Nat.* **1967**, *101*, 301–312.

162. Zipser, D. Polar Mutations and Operon Function. *Nature* **1969**, *221*, 21–25, doi:10.1038/221021a0.

163. Morse, D.E.; Primakoff, P. Relief of Polarity in E. Coli by "SuA". *Nature* **1970**, *226*, 28–31, doi:10.1038/226028a0.

164. Richardson, J.P.; Grimley, C.; Lowery, C. Transcription Termination Factor Rho Activity Is Altered in Escherichia Coli with SuA Gene Mutations. *Proc. Natl. Acad. Sci. U. S. A.* **1975**, *72*, 1725–1728, doi:10.1073/pnas.72.5.1725.

165. Boudvillain, M.; Figueroa-Bossi, N.; Bossi, L. Terminator Still Moving Forward: Expanding Roles for Rho Factor. *Curr. Opin. Microbiol.* 2013, *16*, 118–124.

166. Carpousis, A.J.; Luisi, B.F.; McDowall, K.J. Chapter 3 Endonucleolytic Initiation of MRNA Decay in Escherichia Coli. *Prog. Mol. Biol. Transl. Sci.* 2009, *85*, 91–135.

167. Dreyfus, M. *Chapter 11 Killer and Protective Ribosomes*; 1st ed.; Elsevier Inc., 2009; Vol. 85;.

168. Heap, J.T.; Pennington, O.J.; Cartman, S.T.; Minton, N.P. A Modular System for Clostridium Shuttle Plasmids. *J. Microbiol. Methods* **2009**, *78*, 79–85, doi:10.1016/j.mimet.2009.05.004.

169. Ingle, P.; Groothuis, D.; Rowe, P.; Huang, H.; Cockayne, A.; Kuehne, S.A.; Jiang, W.; Gu, Y.; Humphreys, C.M.; Minton, N.P. Generation of a Fully Erythromycin-Sensitive Strain of Clostridioides Difficile Using a Novel CRISPR-Cas9 Genome Editing System. *Sci. Rep.* **2019**, *9*, doi:10.1038/s41598-019-44458-y.

170. Karberg, M.; Guo, H.; Zhong, J.; Coon, R.; Perutka, J.; Lambowitz, A.M. Group II Introns as Controllable Gene Targeting Vectors for Genetic Manipulation of Bacteria. *Nat. Biotechnol.* **2001**, *19*, 1162–1167, doi:10.1038/nbt1201-1162.

171. Heap, J.T.; Kuehne, S.A.; Ehsaan, M.; Cartman, S.T.; Cooksley, C.M.; Scott, J.C.;

Minton, N.P. The ClosTron: Mutagenesis in Clostridium Refined and Streamlined. *J. Microbiol. Methods* **2010**, *80*, 49–55, doi:10.1016/j.mimet.2009.10.018.

172. Heap, J.T.; Pennington, O.J.; Cartman, S.T.; Carter, G.P.; Minton, N.P. The ClosTron: A Universal Gene Knock-out System for the Genus Clostridium. *J. Microbiol. Methods* **2007**, *70*, 452–464, doi:10.1016/j.mimet.2007.05.021.

173. Zhong, J.; Karberg, M.; Lambowitz, A.M. Targeted and Random Bacterial Gene Disruption Using a Group II Intron (Targetron) Vector Containing a Retrotransposition-Activated Selectable Marker. *Nucleic Acids Res.* **2003**, *31*, 1656–1664.

174. Heap, J.T.; Ehsaan, M.; Cooksley, C.M.; Ng, Y.-K.; Cartman, S.T.; Winzer, K.; Minton, N.P. Integration of DNA into Bacterial Chromosomes from Plasmids without a Counter-Selection Marker. *Nucleic Acids Res.* **2012**, *40*, e59, doi:10.1093/nar/gkr1321.

175. Ng, Y.K.; Ehsaan, M.; Philip, S.; Collery, M.M.; Janoir, C.; Collignon, A.; Cartman, S.T.; Minton, N.P. Expanding the Repertoire of Gene Tools for Precise Manipulation of the Clostridium Difficile Genome: Allelic Exchange Using PyrE Alleles. *PLoS One* **2013**, *8*, e54051, doi:10.1371/journal.pone.0056051.

176. Minton, N.P.; Ehsaan, M.; Humphreys, C.M.; Little, G.T.; Baker, J.; Henstra, A.M.; Liew, F.; Kelly, M.L.; Sheng, L.; Schwarz, K.; et al. A Roadmap for Gene System Development in Clostridium. *Anaerobe* **2016**, *41*, 104–112, doi:10.1016/j.anaerobe.2016.05.011.

177. Storici, F.; Durham, C.L.; Gordenin, D.A.; Resnick, M.A. Chromosomal Site-Specific Double-Strand Breaks Are Efficiently Targeted for Repair by Oligonucleotides in Yeast. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 14994–14999, doi:10.1073/pnas.2036296100.

178. Dicarlo, J.E.; Norville, J.E.; Mali, P.; Rios, X.; Aach, J.; Church, G.M. Genome Engineering in Saccharomyces Cerevisiae Using CRISPR-Cas Systems. *Nucleic Acids Res.* **2013**, doi:10.1093/nar/gkt135.

179. Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J.A.; Charpentier, E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (80-. ).* **2012**, *337*, 816–821, doi:10.1126/science.1225829.

180. Symington, L.S.; Rothstein, R.; Lisby, M. Mechanisms and Regulation of Mitotic Recombination in Saccharomyces Cerevisiae. *Genetics* **2014**, *198*, 795–835, doi:10.1534/genetics.114.166140.

181. Jakočinas, T.; Bonde, I.; Herrgård, M.; Harrison, S.J.; Kristensen, M.; Pedersen, L.E.; Jensen, M.K.; Keasling, J.D. Multiplex Metabolic Pathway Engineering Using CRISPR/Cas9 in Saccharomyces Cerevisiae. *Metab. Eng.* **2015**, *28*, 213–222, doi:10.1016/j.ymben.2015.01.008.

182. Jiang, W.; Bikard, D.; Cox, D.; Zhang, F.; Marraffini, L.A. RNA-Guided Editing of Bacterial Genomes Using CRISPR-Cas Systems. *Nat. Biotechnol.* **2013**, *31*, 233–239, doi:10.1038/nbt.2508.

183. Altenbuchner, J. Editing of the Bacillus Subtilis Genome by the CRISPR-Cas9 System. *Appl. Environ. Microbiol.* **2016**, *82*, 5421–5427, doi:10.1128/AEM.01453-16.

184. Wang, Y.; Zhang, Z.-T.; Seo, S.-O.; Lynn, P.; Lu, T.; Jin, Y.-S.; Blaschek, H.P. Bacterial Genome Editing with CRISPR-Cas9: Deletion, Integration, Single Nucleotide Modification, and Desirable "Clean" Mutant Selection in Clostridium Beijerinckii as an Example. *ACS Synth. Biol.* **2016**, *5*, 721–732, doi:10.1021/acssynbio.6b00060.

185. Xu, T.; Li, Y.; Shi, Z.; Hemme, C.L.; Li, Y.; Zhu, Y.; Van Nostrand, J.D.; He, Z.; Zhou, J. Efficient Genome Editing in Clostridium Cellulolyticum via CRISPR-Cas9 Nickase. *Appl. Environ. Microbiol.* **2015**, *81*, 4423–4431, doi:10.1128/AEM.00873-15.

186. Wasels, F.; Jean-Marie, J.; Collas, F.; L?pez-Contreras, A.M.; Ferreira, N.L. A Two-Plasmid Inducible CRISPR/Cas9 Genome Editing Tool for Clostridium Acetobutylicum. *J. Microbiol. Methods* **2017**, *140*, 5–11, doi:10.1016/j.mimet.2017.06.010.

187. Barrangou, R.; Marraffini, L.A. CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. *Mol. Cell* 2014, *54*, 234–244.

188. Ishino, Y.; Shinagawa, H.; Makino, K.; Amemura, M.; Nakata, A. Nucleotide Sequence of the Iap Gene, Responsible for Alkaline Phosphatase Isozyme Conversion in Escherichia Coli, and Identification of the Gene Product. *J. Bacteriol.* **1987**, *169*, 5429–5433, doi:10.1128/jb.169.12.5429-5433.1987.

189. Nakata, A.; Amemura, M.; Makino, K. Unusual Nucleotide Arrangement with

Repeated Sequences in the Escherichia Coli K-12 Chromosome. *J. Bacteriol.* **1989**, *171*, 3553–3556.

190. Jansen, R.; Embden, J.D.A. van; Gaastra, W.; Schouls, L.M. Identification of Genes That Are Associated with DNA Repeats in Prokaryotes. *Mol. Microbiol.* **2002**, *43*, 1565–1575, doi:10.1046/j.1365-2958.2002.02839.x.

191. Mojica, F.J.M.; Díez-Villaseñor, C.; Soria, E.; Juez, G. Biological Significance of a Family of Regularly Spaced Repeats in the Genomes of Archaea, Bacteria and Mitochondria. *Mol. Microbiol.* 2000, *36*, 244–246.

192. Pourcel, C.; Salvignol, G.; Vergnaud, G. CRISPR Elements in Yersinia Pestis Acquire New Repeats by Preferential Uptake of Bacteriophage DNA, and Provide Additional Tools for Evolutionary Studies. *Microbiology* **2005**, *151*, 653–663, doi:10.1099/mic.0.27437-0.

193. Bolotin, A.; Quinquis, B.; Sorokin, A.; Dusko Ehrlich, S. Clustered Regularly Interspaced Short Palindrome Repeats (CRISPRs) Have Spacers of Extrachromosomal Origin. *Microbiology* **2005**, *151*, 2551–2561, doi:10.1099/mic.0.28048-0.

194. Mojica, F.J.M.; Díez-Villaseñor, C.; García-Martínez, J.; Soria, E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *J. Mol. Evol.* **2005**, *60*, 174–182, doi:10.1007/s00239-004-0046-3.

195. Makarova, K.S.; Grishin, N. V; Shabalina, S.A.; Wolf, Y.I.; Koonin, E. V A Putative RNA-Interference-Based Immune System in Prokaryotes: Computational Analysis of the Predicted Enzymatic Machinery, Functional Analogies with Eukaryotic RNAi, and Hypothetical Mechanisms of Action. *Biol. Direct* 2006, *1*, 7.

196. Wright, A. V.; Nuñez, J.K.; Doudna, J.A. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* 2016, *164*, 29–44.

197. Deltcheva, E.; Chylinski, K.; Sharma, C.M.; Gonzales, K.; Chao, Y.; Pirzada, Z.A.; Eckert, M.R.; Vogel, J.; Charpentier, E. CRISPR RNA Maturation by Trans-Encoded Small RNA and Host Factor RNase III. *Nature* **2011**, *471*, 602–607, doi:10.1038/nature09886.

198. Kleinstiver, B.P.; Prew, M.S.; Tsai, S.Q.; Topkar, V. V.; Nguyen, N.T.; Zheng, Z.; Gonzales, A.P.W.; Li, Z.; Peterson, R.T.; Yeh, J.-R.J.; et al. Engineered CRISPR-Cas9

Nucleases with Altered PAM Specificities. *Nature* **2015**, *523*, 481–485, doi:10.1038/nature14592.

199. Slaymaker, I.M.; Gao, L.; Zetsche, B.; Scott, D.A.; Yan, W.X.; Zhang, F. Rationally Engineered Cas9 Nucleases with Improved Specificity. *Science (80-. ).* **2016**, *351*, doi:10.1126/science.aad5227.

200. Liu, Z.; Shan, H.; Chen, S.; Chen, M.; Song, Y.; Lai, L.; Li, Z. Efficient Base Editing with Expanded Targeting Scope Using an Engineered Spy-Mac Cas9 Variant. *Cell Discov.* 2019, *5*, 58.

201. Vakulskas, C.A.; Dever, D.P.; Rettig, G.R.; Turk, R.; Jacobi, A.M.; Collingwood, M.A.; Bode, N.M.; McNeill, M.S.; Yan, S.; Camarena, J.; et al. A High-Fidelity Cas9 Mutant Delivered as a Ribonucleoprotein Complex Enables Efficient Gene Editing in Human Hematopoietic Stem and Progenitor Cells. *Nat. Med.* **2018**, *24*, 1216–1224, doi:10.1038/s41591-018-0137-0.

202. Lee, J.K.; Jeong, E.; Lee, J.; Jung, M.; Shin, E.; Kim, Y. hoon; Lee, K.; Jung, I.; Kim, D.; Kim, S.; et al. Directed Evolution of CRISPR-Cas9 to Increase Its Specificity. *Nat. Commun.* **2018**, *9*, 1–10, doi:10.1038/s41467-018-05477-x.

203. Huang, T.P.; Zhao, K.T.; Miller, S.M.; Gaudelli, N.M.; Oakes, B.L.; Fellmann, C.; Savage, D.F.; Liu, D.R. Circularly Permuted and PAM-Modified Cas9 Variants Broaden the Targeting Scope of Base Editors. *Nat. Biotechnol.* **2019**, *37*, 626–631, doi:10.1038/s41587-019-0134-y.

204. Kleinstiver, B.P.; Sousa, A.A.; Walton, R.T.; Tak, Y.E.; Hsu, J.Y.; Clement, K.; Welch, M.M.; Horng, J.E.; Malagon-Lopez, J.; Scarfò, I.; et al. Engineered CRISPR–Cas12a Variants with Increased Activities and Improved Targeting Ranges for Gene, Epigenetic and Base Editing. *Nat. Biotechnol.* **2019**, *37*, 276–282, doi:10.1038/s41587-018-0011-0.

205. Pickar-Oliver, A.; Gersbach, C.A. The next Generation of CRISPR–Cas Technologies and Applications. *Nat. Rev. Mol. Cell Biol.* **2019**, 1, doi:10.1038/s41580-019-0131-5.

206. Yan, M.Y.; Yan, H.Q.; Ren, G.X.; Zhao, J.P.; Guo, X.P.; Sun, Y.C. CRISPR-Cas12a-Assisted Recombineering in Bacteria. *Appl. Environ. Microbiol.* **2017**, *83*, doi:10.1128/AEM.00947-17.

207. Ao, X.; Yao, Y.; Li, T.; Yang, T.T.; Dong, X.; Zheng, Z.T.; Chen, G.Q.; Wu, Q.; Guo, Y. A Multiplex Genome Editing Method for Escherichia Coli Based on CRISPR-Cas12a. *Front. Microbiol.* **2018**, *9*, 1–13, doi:10.3389/fmicb.2018.02307.

208. Zhang, J.; Zong, W.; Hong, W.; Zhang, Z.T.; Wang, Y. Exploiting Endogenous CRISPR-Cas System for Multiplex Genome Editing in Clostridium Tyrobutyricum and Engineer the Strain for High-Level Butanol Production. *Metab. Eng.* **2018**, *47*, 49–59, doi:10.1016/j.ymben.2018.03.007.

209. Pyne, M.E.; Bruder, M.R.; Moo-Young, M.; Chung, D.A.; Chou, C.P. Harnessing Heterologous and Endogenous CRISPR-Cas Machineries for Efficient Markerless Genome Editing in Clostridium. *Sci. Rep.* **2016**, *6*, 25666, doi:10.1038/srep25666.

210. Farboud, B.; Meyer, B.J. Dramatic Enhancement of Genome Editing by CRISPR/Cas9 through Improved Guide RNA Design. *Genetics* **2015**, *199*, 959–971, doi:10.1534/genetics.115.175166.

211. Tsai, S.Q.; Zheng, Z.; Nguyen, N.T.; Liebers, M.; Topkar, V. V; Thapar, V.; Wyvekens, N.; Khayter, C.; Iafrate, A.J.; Le, L.P.; et al. GUIDE-Seq Enables Genome-Wide Profiling of off-Target Cleavage by CRISPR-Cas Nucleases. *Nat. Biotechnol.* **2015**, *33*, 187–197, doi:10.1038/nbt.3117.

212. Doench, J.G.; Hartenian, E.; Graham, D.B.; Tothova, Z.; Hegde, M.; Smith, I.; Sullender, M.; Ebert, B.L.; Xavier, R.J.; Root, D.E. Rational Design of Highly Active SgRNAs for CRISPR-Cas9-Mediated Gene Inactivation. *Nat. Biotechnol.* **2014**, *32*, 1262–1267, doi:10.1038/nbt.3026.

213. Doench, J.G.; Fusi, N.; Sullender, M.; Hegde, M.; Vaimberg, E.W.; Donovan, K.F.; Smith, I.; Tothova, Z.; Wilen, C.; Orchard, R.; et al. Optimized SgRNA Design to Maximize Activity and Minimize Off-Target Effects of CRISPR-Cas9. *Nat. Biotechnol.* **2016**, *34*, 184–191, doi:10.1038/nbt.3437.

214. Moreno-Mateos, M.A.; Vejnar, C.E.; Beaudoin, J.D.; Fernandez, J.P.; Mis, E.K.; Khokha, M.K.; Giraldez, A.J. CRISPRscan: Designing Highly Efficient SgRNAs for CRISPR-Cas9 Targeting in Vivo. *Nat. Methods* **2015**, *12*, 982–988, doi:10.1038/nmeth.3543.

215. Chari, R.; Mali, P.; Moosburner, M.; Church, G.M. Unraveling CRISPR-Cas9 Genome Engineering Parameters via a Library-on-Library Approach. *Nat. Methods* **2015**, *12*,

823–826, doi:10.1038/nmeth.3473.

216. Wilson, L.O.W.; O'Brien, A.R.; Bauer, D.C. The Current State and Future of CRISPR-Cas9 GRNA Design Tools. *Front. Pharmacol.* **2018**, *9*, 1–6, doi:10.3389/fphar.2018.00749.

217. Zhang, G.; Dai, Z.; Dai, X. C-RNNCrispr: Prediction of CRISPR/Cas9 SgRNA Activity Using Convolutional and Recurrent Neural Networks. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 344–354, doi:10.1016/j.csbj.2020.01.013.

218. Wong, N.; Liu, W.; Wang, X. WU-CRISPR: Characteristics of Functional Guide RNAs for the CRISPR/Cas9 System. *Genome Biol.* **2015**, *16*, 1–8, doi:10.1186/s13059-015-0784-0.

219. Haeussler, M.; Schönig, K.; Eckert, H.; Eschstruth, A.; Mianné, J.; Renaud, J.B.; Schneider-Maunoury, S.; Shkumatava, A.; Teboul, L.; Kent, J.; et al. Evaluation of Off-Target and on-Target Scoring Algorithms and Integration into the Guide RNA Selection Tool CRISPOR. *Genome Biol.* **2016**, *17*, 1–12, doi:10.1186/s13059-016-1012-2.

220. Guo, J.; Wang, T.; Guan, C.; Liu, B.; Luo, C.; Xie, Z.; Zhang, C.; Xing, X.H. Improved SgRNA Design in Bacteria via Genome-Wide Activity Profiling. *Nucleic Acids Res.* **2018**, *46*, 7052–7069, doi:10.1093/nar/gky572.

221. Wang, L.; Zhang, J. Prediction of SgRNA On-Target Activity in Bacteria by Deep Learning. *BMC Bioinformatics* **2019**, *20*, 1–14, doi:10.1186/s12859-019-3151-4.

222. Hsu, P.D.; Scott, D.A.; Weinstein, J.A.; Ran, F.A.; Konermann, S.; Agarwala, V.; Li, Y.; Fine, E.J.; Wu, X.; Shalem, O.; et al. DNA Targeting Specificity of RNA-Guided Cas9 Nucleases. *Nat. Biotechnol.* **2013**, *31*, 827–832, doi:10.1038/nbt.2647.

223. Cong, L.; Ran, F.A.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P.D.; Wu, X.; Jiang, W.; Marraffini, L.A.; et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-. ).* **2013**, *339*, 819–823, doi:10.1126/science.1231143.

224. Cho, S.W.; Kim, S.; Kim, Y.; Kweon, J.; Kim, H.S.; Bae, S.; Kim, J.S. Analysis of Off-Target Effects of CRISPR/Cas-Derived RNA-Guided Endonucleases and Nickases. *Genome Res.* **2014**, *24*, 132–141, doi:10.1101/gr.162339.113.

225. Wiedenheft, B.; Sternberg, S.H.; Doudna, J.A. RNA-Guided Genetic Silencing Systems

in Bacteria and Archaea. *Nature* **2012**, *482*, 331–338, doi:10.1038/nature10886.

226. Semenova, E.; Jore, M.M.; Datsenko, K.A.; Semenova, A.; Westra, E.R.; Wanner, B.; van der Oost, J.; Brouns, S.J.J.; Severinov, K. Interference by Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) RNA Is Governed by a Seed Sequence. *Proc. Natl. Acad. Sci.* **2011**, *108*, 10098–10103, doi:10.1073/pnas.1104144108.

227. Stemmer, M.; Thumberger, T.; Del Sol Keyer, M.; Wittbrodt, J.; Mateo, J.L. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PLoS One* **2015**, *10*, 1–11, doi:10.1371/journal.pone.0124633.

228. Pattanayak, V.; Lin, S.; Guilinger, J.P.; Ma, E.; Doudna, J.A.; Liu, D.R. High-Throughput Profiling of off-Target DNA Cleavage Reveals RNA-Programmed Cas9 Nuclease Specificity. *Nat. Biotechnol.* **2013**, *31*, 839–843, doi:10.1038/nbt.2673.

229. Lin, Y.; Cradick, T.J.; Brown, M.T.; Deshmukh, H.; Ranjan, P.; Sarode, N.; Wile, B.M.; Vertino, P.M.; Stewart, F.J.; Bao, G. CRISPR/Cas9 Systems Have off-Target Activity with Insertions or Deletions between Target DNA and Guide RNA Sequences. *Nucleic Acids Res.* **2014**, *42*, 7473–7485, doi:10.1093/nar/gku402.

230. Abadi, S.; Yan, W.X.; Amar, D.; Mayrose, I. A Machine Learning Approach for Predicting CRISPR-Cas9 Cleavage Efficiencies and Patterns Underlying Its Mechanism of Action. *PLoS Comput. Biol.* **2017**, *13*, 1–24, doi:10.1371/journal.pcbi.1005807.

231. Listgarten, J.; Weinstein, M.; Kleinstiver, B.P.; Sousa, A.A.; Joung, J.K.; Crawford, J.; Gao, K.; Hoang, L.; Elibol, M.; Doench, J.G.; et al. Prediction of Off-Target Activities for the End-to-End Design of CRISPR Guide RNAs. *Nat. Biomed. Eng.* **2018**, *2*, 38–47, doi:10.1038/s41551-017-0178-6.

232. Lin, J.; Wong, K.C. Off-Target Predictions in CRISPR-Cas9 Gene Editing Using Deep Learning. *Bioinformatics* **2018**, *34*, i656–i663, doi:10.1093/bioinformatics/bty554.

233. Koo, T.; Lee, J.; Kim, J.S. Measuring and Reducing Off-Target Activities of Programmable Nucleases Including CRISPR-Cas9. *Mol. Cells* 2015, *38*, 475–481.

234. Fu, Y.; Sander, J.D.; Reyon, D.; Cascio, V.M.; Joung, J.K. Improving CRISPR-Cas Nuclease Specificity Using Truncated Guide RNAs. *Nat. Biotechnol.* **2014**, *32*, 279–284, doi:10.1038/nbt.2808.

235. Ran, F.A.; Hsu, P.D.; Lin, C.Y.; Gootenberg, J.S.; Konermann, S.; Trevino, A.E.; Scott, D.A.; Inoue, A.; Matoba, S.; Zhang, Y.; et al. Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **2013**, *154*, 1380–1389, doi:10.1016/j.cell.2013.08.021.

236. Woods, C.; Humphreys, C.M.; Rodrigues, R.M.; Ingle, P.; Rowe, P.; Henstra, A.M.; Köpke, M.; Simpson, S.D.; Winzer, K.; Minton, N.P. A Novel Conjugal Donor Strain for Improved DNA Transfer into Clostridium Spp. *Anaerobe* **2019**, *59*, 184–191, doi:10.1016/j.anaerobe.2019.06.020.

237. Geissmann, Q. OpenCFU, a New Free and Open-Source Software to Count Cell Colonies and Other Circular Objects. *PLoS One* **2013**, *8*, 1–10, doi:10.1371/journal.pone.0054072.

238. Chen, C.Y.; Nace, G.W.; Irwin, P.L. A 6x6 Drop Plate Method for Simultaneous Colony Counting and MPN Enumeration of Campylobacter Jejuni, Listeria Monocytogenes, and Escherichia Coli. *J. Microbiol. Methods* **2003**, *55*, 475–479, doi:10.1016/S0167-7012(03)00194-5.

239. Zuker, M. Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Res.* **2003**, *31*, 3406–3415.

240. Humphreys, C.M.; Mclean, S.; Schatschneider, S.; Millat, T.; Henstra, A.M.; Annan, F.J.; Breitkopf, R.; Pander, B.; Piatek, P.; Rowe, P.; et al. Whole Genome Sequence and Manual Annotation of Clostridium Autoethanogenum, an Industrially Relevant Bacterium. *BMC Genomics* **2015**, *16*, doi:10.1186/s12864-015-2287-5.

241. NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2017**, *45*, D12–D17, doi:10.1093/nar/gkw1071.

242. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277, doi:10.1016/S0168-9525(00)02024-2.

243. Chin, J.X.; Chung, B.K.-S.; Lee, D.-Y. Codon Optimization OnLine (COOL): A Web-Based Multi-Objective Optimization Platform for Synthetic Gene Design. *Bioinformatics* **2014**, *30*, 2210–2212, doi:10.1093/bioinformatics/btu192.

244. Willson, B.J.; Kovács, K.; Wilding-Steele, T.; Markus, R.; Winzer, K.; Minton, N.P. Production of a Functional Cell Wall-Anchored Minicellulosome by Recombinant Clostridium Acetobutylicum ATCC 824. *Biotechnol. Biofuels* **2016**, *9*, 109, doi:10.1186/s13068-016-0526-x.

245. Cañadas, I.C.; Groothuis, D.; Zygouropoulou, M.; Rodrigues, R.; Minton, N.P. RiboCas: A Universal CRISPR-Based Editing Tool for Clostridium. *ACS Synth. Biol.* **2019**, *8*, 1379–1390, doi:10.1021/acssynbio.9b00075.

246. Pander, B. Nutrient Limitation in Clostridium Autoethanogenum and Characterisation of Its Carbonic Anhydrase Bart Pander , MSc, University of Nottingham, 2017.

247. Ullmann, A.; Jacob, F.; Monod, J. Characterization by in Vitro Complementation of a Peptide Corresponding to an Operator-Proximal Segment of the β-Galactosidase Structural Gene of Escherichia Coli. *J. Mol. Biol.* **1967**, *24*, 339–343, doi:10.1016/0022-2836(67)90341-5.

248. Tannous, J.; Barda, O.; Luciano-Rosario, D.; Prusky, D.B.; Sionov, E.; Keller, N.P. New Insight Into Pathogenicity and Secondary Metabolism of the Plant Pathogen Penicillium Expansum Through Deletion of the Epigenetic Reader SntB. *Front. Microbiol.* **2020**, *11*, doi:10.3389/fmicb.2020.00610.

249. Guérout-Fleury, A.M.; Frandsen, N.; Stragier, P. Plasmids for Ectopic Integration in Bacillus Subtilis. *Gene* **1996**, *180*, 57–61, doi:10.1016/S0378-1119(96)00404-0.

250. Carvalho, N.D.S.P.; Arentshorst, M.; Jin Kwon, M.; Meyer, V.; Ram, A.F.J. Expanding the Ku70 Toolbox for Filamentous Fungi: Establishment of Complementation Vectors and Recipient Strains for Advanced Gene Analyses. *Appl. Microbiol. Biotechnol.* **2010**, *87*, 1463–1473, doi:10.1007/s00253-010-2588-1.

251. Rose, M.D. Isolation of Genes by Complementation in Yeast. *Methods Enzymol.* **1987**, *152*, 481–504, doi:10.1016/0076-6879(87)52056-0.

252. Ehsaan, M.; Kuit, W.; Zhang, Y.; Cartman, S.T.; Heap, J.T.; Winzer, K.; Minton, N.P. Mutant Generation by Allelic Exchange and Genome Resequencing of the Biobutanol Organism Clostridium Acetobutylicum ATCC 824. *Biotechnol. Biofuels* **2016**, *9*, doi:10.1186/s13068-015-0410-0.

253. Higgins, D.; Dworkin, J. Recent Progress in Bacillus Subtilis Sporulation. *FEMS Microbiol. Rev.* 2012, *36*, 131–148.

254. Oh, J.H.; Van Pijkeren, J.P. CRISPR-Cas9-Assisted Recombineering in Lactobacillus Reuteri. *Nucleic Acids Res.* **2014**, *42*, doi:10.1093/nar/gku623.

255. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410, doi:10.1016/S0022-2836(05)80360-2.

256. Finnigan, G.C.; Thorner, J. MCAL: A New Approach for Versatile Multiplex Action of Cas9 Using One SgRNA and Loci Flanked by a Programmed Target Sequence. *G3* **2016**, *6*, 2147–2156, doi:10.1534/g3.116.029801.

257. Jakimo, N.; Chatterjee, P.; Nip, L.; Jacobson, J. A Cas9 with Complete PAM Recognition for Adenine Dinucleotides. *bioRxiv* **2018**, *9*, 429654, doi:10.1101/429654.

258. Tan, J.; Zhang, F.; Karcher, D.; Bock, R. Expanding the Genome-Targeting Scope and the Site Selectivity of High-Precision Base Editors. *Nat. Commun.* **2020**, *11*, doi:10.1038/s41467-020-14465-z.

259. Jiang, W.; Feng, S.; Huang, S.; Yu, W.; Li, G.; Yang, G.; Liu, Y.; Zhang, Y.; Zhang, L.; Hou, Y.; et al. BE-PLUS: A New Base Editing Tool with Broadened Editing Window and Enhanced Fidelity. *Cell Res.* **2018**, *28*, 855–861, doi:10.1038/s41422-018-0052-4.

260. Nishimasu, H.; Shi, X.; Ishiguro, S.; Gao, L.; Hirano, S.; Okazaki, S.; Noda, T.; Abudayyeh, O.O.; Gootenberg, J.S.; Mori, H.; et al. Engineered CRISPR-Cas9 Nuclease with Expanded Targeting Space. *Science (80-. ).* **2018**, *361*, 1259–1262, doi:10.1126/science.aas9129.

261. Zhang, Y.; Ge, X.; Yang, F.; Zhang, L.; Zheng, J.; Tan, X.; Jin, Z.B.; Qu, J.; Gu, F. Comparison of Non-Canonical PAMs for CRISPR/Cas9-Mediated DNA Cleavage in Human Cells. *Sci. Rep.* **2014**, *4*, 1–5, doi:10.1038/srep05405.

262. Kocak, D.D.; Josephs, E.A.; Bhandarkar, V.; Adkar, S.S.; Kwon, J.B.; Gersbach, C.A. Increasing the Specificity of CRISPR Systems with Engineered RNA Secondary Structures. *Nat. Biotechnol.* **2019**, *37*, 657–666, doi:10.1038/s41587-019-0095-1.

263. Edraki, A.; Mir, A.; Ibraheim, R.; Gainetdinov, I.; Yoon, Y.; Song, C.Q.; Cao, Y.; Gallant,

J.; Xue, W.; Rivera-Pérez, J.A.; et al. A Compact, High-Accuracy Cas9 with a Dinucleotide PAM for In Vivo Genome Editing. *Mol. Cell* **2019**, doi:10.1016/j.molcel.2018.12.003.

264. Blakely, G.W.; Murray, N.E. DNA Restriction and Modification. *Encycl. Microbiol.* **2009**, 538–549, doi:10.1016/B978-012373944-5.00033-X.

265. Wilson, G.G. Organization of Restriction-Modification Systems. *Nucleic Acids Res.* **1991**, *19*, 2539–2566.

266. Kolek, J.; Sedlar, K.; Provaznik, I.; Patakova, P. Dam and Dcm Methylations Prevent Gene Transfer into Clostridium Pasteurianum NRRL B-598: Development of Methods for Electrotransformation, Conjugation, and Sonoporation. *Biotechnol. Biofuels* **2016**, *9*, doi:10.1186/s13068-016-0436-y.

267. Soucy, S.M.; Huang, J.; Gogarten, J.P. Horizontal Gene Transfer: Building the Web of Life. *Nat. Rev. Genet.* **2015**, *16*, 472–482, doi:10.1038/nrg3962.

268. McGee, D.J.; Coker, C.; Harro, J.M.; Mobley, H.L. Bacterial Genetic Exchange. *Encycl. Life Sci.* **2001**, doi:10.1038/npg.els.0001416.

269. Asif, A.; Mohsin, H.; Tanvir, R.; Rehman, Y. Revisiting the Mechanisms Involved in Calcium Chloride Induced Bacterial Transformation. *Front. Microbiol.* **2017**, *8*, 6–10, doi:10.3389/fmicb.2017.02169.

270. Chang, A.Y.; Chau, V.W.; Landas, J.A.; Yvonne Preparation of Calcium Competent Escherichia Coli and Heat-Shock Transformation. *J. Exp. Microbiol. Immunol.* **2017**, *1*, 22–25.

271. Jennert, K.C.B.; Young, D.I.; Young, M.; Tardif, C. Gene Transfer to Clostridium Cellulolyticum ATCC 35319. *Microbiology* **2000**, *146*, 3071–3080, doi:10.1099/00221287-146-12-3071.

272. Zhang, J.; Hong, W.; Guo, L.; Wang, Y.; Wang, Y. Enhancing Plasmid Transformation Efficiency and Enabling CRISPR-Cas9/Cpf1-Based Genome Editing in Clostridium Tyrobutyricum. *Biotechnol. Bioeng.* **2020**, *117*, 2911–2917, doi:10.1002/bit.27435.

273. Williams, D.R.; Young, D.I.; Young, M. Conjugative Plasmid Transfer from Escherichia Coli to Clostridium Acetobutylicum. *J. Gen. Microbiol.* **1990**, *136*, 819–826, doi:10.1099/00221287-136-5-819.

274. Hedges, R.W.; Jacob, A.E. Transposition of Ampicillin Resistance from RP4 to Other Replicons. *MGG Mol. Gen. Genet.* **1974**, *132*, 31–40, doi:10.1007/BF00268228.

275. Levin, B.R.; Stewart, F.M.; Rice, V.A. The Kinetics of Conjugative Plasmid Transmission: Fit of a Simple Mass Action Model. *Plasmid* **1979**, *2*, 247–260, doi:10.1016/0147-619X(79)90043-X.

276. MACDONALD, J.; SMETS, B.; RITTMANN, B. The Effects of Energy Availability on the Conjugative-Transfer Kinetics of Plasmid RP4. *Water Res.* **1992**, *26*, 461–468, doi:10.1016/0043-1354(92)90046-7.

277. Philipsen, K.R.; Christiansen, L.E.; Hasman, H.; Madsen, H. Modelling Conjugation with Stochastic Differential Equations. *J. Theor. Biol.* **2010**, *263*, 134–142, doi:10.1016/j.jtbi.2009.11.011.

278. Kirk, J.A.; Fagan, R.P. Heat Shock Increases Conjugation Efficiency in Clostridium Difficile. *Anaerobe* **2016**, *42*, 1–5, doi:10.1016/j.anaerobe.2016.06.009.

279. Annan, F.J.; Al-Sinawi, B.; Humphreys, C.M.; Norman, R.; Winzer, K.; Köpke, M.; Simpson, S.D.; Minton, N.P.; Henstra, A.M. Engineering of Vitamin Prototrophy in Clostridium Ljungdahlii and Clostridium Autoethanogenum. *Appl. Microbiol. Biotechnol.* **2019**, *103*, 4633–4648, doi:10.1007/s00253-019-09763-6.

280. Patching, S.G.; Baldwin, S.A.; Baldwin, A.D.; Young, J.D.; Gallagher, M.P.; Henderson, P.J.F.; Herbert, R.B. The Nucleoside Transport Proteins, NupC and NupG, from Escherichia Coli: Specific Structural Motifs Necessary for the Binding of Ligands. *Org. Biomol. Chem.* **2005**, *3*, 462–470, doi:10.1039/b414739a.

281. Musick, W.D.L.; Nyhan, W.L. Structural Features of the Phosphoribosyl-Transferases and Their Relationship to the Human Deficiency Disorders of Purine and Pyrimidine Metabolis. *Crit. Rev. Biochem. Mol. Biol.* **1981**, *11*, 1–34, doi:10.3109/10409238109108698.

282. Kim, M.K.; Song, H.E.; Kim, Y.S.; Rho, S.H.; Im, Y.J.; Lee, J.H.; Kang, G.B.; Eom, S.H. Crystallization and Preliminary X-Ray Crystallographic Analysis of Orotate Phosphoribosyltransferase from Helicobacter Pylori. *Mol. Cells* **2003**, *15*, 361–363.

283. Wu, H.; Li, Y.; Ma, Q.; Li, Q.; Jia, Z.; Yang, B.; Xu, Q.; Fan, X.; Zhang, C.; Chen, N.; et al. Metabolic Engineering of Escherichia Coli for High-Yield Uridine Production. *Metab.*

*Eng.* **2018**, *49*, 248–256, doi:10.1016/j.ymben.2018.09.001.

284. Kravanja, M.; Engelmann, R.; Dossonnet, V.; Blüggel, M.; Meyer, H.E.; Frank, R.; Galinier, A.; Deutscher, J.; Schnell, N.; Hengstenberg, W. The HprK Gene of Enterococcus Faecalis Encodes a Novel Bifunctional Enzyme: The HPr Kinase/Phosphatase. *Mol. Microbiol.* **1999**, *31*, 59–66, doi:10.1046/j.1365-2958.1999.01146.x.

285. Rowe, P. Development of Molecular Tools for Optimisation of C1 Gas Fermentation in Acetogens, University of Nottingham, 2017.

286. Heuer, H.; Abdo, Z.; Smalla, K. Patchy Distribution of Flexible Genetic Elements in Bacterial Populations Mediates Robustness to Environmental Uncertainty. *FEMS Microbiol. Ecol.* **2008**, *65*, 361–371, doi:10.1111/j.1574-6941.2008.00539.x.

287. Koraimann, G.; Wagner, M.A. Social Behavior and Decision Making in Bacterial Conjugation. *Front. Cell. Infect. Microbiol.* **2014**, *4*, 1–7, doi:10.3389/fcimb.2014.00054.

288. Rösch, T.C.; Graumann, P.L. Induction of Plasmid Conjugation in Bacillus Subtilis Is Bistable and Driven by a Direct Interaction of a Rap/Phr Quorum-Sensing System with a Master Repressor. *J. Biol. Chem.* **2015**, *290*, 20221–20232, doi:10.1074/jbc.M115.664110.

289. Marinus, M.G.; Løbner-Olesen, A. DNA Methylation. *EcoSal Plus* **2013**, *1*, doi:10.1128/ecosalplus.ESP-0003-2013.

290. Wyrzykowski, J.; Volkert, M.R. The Escherichia Coli Methyl-Directed Mismatch Repair System Repairs Base Pairs Containing Oxidative Lesions. *J. Bacteriol.* **2003**, *185*, 1701–1704, doi:10.1128/JB.185.5.1701-1704.2003.

291. Robbins-Manke, J.L.; Zdraveski, Z.Z.; Marinus, M.; Essigmann, J.M. Analysis of Global Gene Expression and Double-Strand-Break Formation in DNA Adenine Methyltransferase- and Mismatch Repair-Deficient Escherichia Coli. *J. Bacteriol.* **2005**, *187*, 7027–7037, doi:10.1128/JB.187.20.7027-7037.2005.

292. Cherry, J.L. Methylation-Induced Hypermutation in Natural Populations of Bacteria. *J. Bacteriol.* **2018**, *200*, 1–12, doi:10.1128/JB.00371-18.

293. Hernandez-Valdes, J.A.; van Gestel, J.; Kuipers, O.P. A Riboswitch Gives Rise to

Multi-Generational Phenotypic Heterogeneity in an Auxotrophic Bacterium. *Nat. Commun.* **2020**, *11*, 1–13, doi:10.1038/s41467-020-15017-1.

294. Müller, S.; Harms, H.; Bley, T. Origin and Analysis of Microbial Population Heterogeneity in Bioprocesses. *Curr. Opin. Biotechnol.* **2010**, *21*, 100–113, doi:10.1016/j.copbio.2010.01.002.

295. Blow, M.J.; Clark, T.A.; Daum, C.G.; Deutschbauer, A.M.; Fomenkov, A.; Fries, R.; Froula, J.; Kang, D.D.; Malmstrom, R.R.; Morgan, R.D.; et al. The Epigenomic Landscape of Prokaryotes. *PLOS Genet.* **2016**, *12*, e1005854, doi:10.1371/journal.pgen.1005854.

296. Vandenbussche, I.; Sass, A.; Van Nieuwerburgh, F.; Pinto-Carbó, M.; Mannweiler, O.; Eberl, L.; Coenye, T. Detection of Cytosine Methylation in Burkholderia Cenocepacia by Single-Molecule Real-Time Sequencing and Whole-Genome Bisulfite Sequencing. *Microbiol. (United Kingdom)* **2021**, *167*, doi:10.1099/mic.0.001027.

297. Morrison, J.; Koeman, J.M.; Johnson, B.K.; Foy, K.K.; Beddows, I.; Zhou, W.; Chesla, D.W.; Rossell, L.L.; Siegwald, E.J.; Adams, M.; et al. Evaluation of Whole-Genome DNA Methylation Sequencing Library Preparation Protocols. *Epigenetics and Chromatin* **2021**, *14*, 1–15, doi:10.1186/s13072-021-00401-y.

298. Nishida, K.; Arazoe, T.; Yachie, N.; Banno, S.; Kakimoto, M.; Tabata, M.; Mochizuki, M.; Miyabe, A.; Araki, M.; Hara, K.Y.; et al. Targeted Nucleotide Editing Using Hybrid Prokaryotic and Vertebrate Adaptive Immune Systems. *Science (80-. ).* **2016**, *353*, aaf8729–aaf8729, doi:10.1126/science.aaf8729.

299. Komor, A.C.; Kim, Y.B.; Packer, M.S.; Zuris, J.A.; Liu, D.R. Programmable Editing of a Target Base in Genomic DNA without Double-Stranded DNA Cleavage. *Nature* **2016**, *533*, 420–424, doi:10.1038/nature17946.

300. Wang, L.; Xue, W.; Yan, L.; Li, X.; Wei, J.; Chen, M.; Wu, J.; Yang, B.; Yang, L.; Chen, J. Enhanced Base Editing by Co-Expression of Free Uracil DNA Glycosylase Inhibitor. *Cell Res.* **2017**, *27*, 1289–1292, doi:10.1038/cr.2017.111.

301. Storb, U.; Stavnezer, J. Immunoglobulin Genes: Generating Diversity with AID and UNG. *Curr. Biol.* **2002**, *12*, 725–727, doi:10.1016/S0960-9822(02)01247-2.

302. Takahashi, I.; Marmur, J. Replacement of Thymidylic Acid by Deoxyuridylic Acid in

the Deoxyribonucleic Acid of a Transducing Phage for Bacillus Subtilis. *Nature* **1963**, *197*, 794–795, doi:10.1038/197794a0.

303. Zhigang, W.; Smith, D.G.; Mosbaugh, D.W. Overproduction and Characterization of the Uracil-DNA Glycosylase Inhibitor of Bacteriophage PBS2. *Gene* **1991**, *99*, 31–37, doi:10.1016/0378-1119(91)90030-F.

304. Banno, S.; Nishida, K.; Arazoe, T.; Mitsunobu, H.; Kondo, A. Deaminase-Mediated Multiplex Genome Editing in Escherichia Coli. *Nat. Microbiol.* **2018**, *3*, 423–429, doi:10.1038/s41564-017-0102-6.

305. Molla, K.A.; Yang, Y. CRISPR/Cas-Mediated Base Editing: Technical Considerations and Practical Applications. *Trends Biotechnol.* **2019**, *37*, 1121–1142, doi:10.1016/J.TIBTECH.2019.03.008.

306. Eid, A.; Alshareef, S.; Mahfouz, M.M. CRISPR Base Editors: Genome Editing without Double-Stranded Breaks. *Biochem. J.* 2018, *475*, 1955–1964.

307. Marx, V. Base Editing a CRISPR Way. *Nat. Methods* **2018**, *15*, 767–770, doi:10.1038/s41592-018-0146-4.

308. Gaudelli, N.M.; Komor, A.C.; Rees, H.A.; Packer, M.S.; Badran, A.H.; Bryson, D.I.; Liu, D.R. Programmable Base Editing of A•T to G•C in Genomic DNA without DNA Cleavage. *Nature* **2017**, *551*, 464–471, doi:10.1038/nature24644.

309. Chatterjee, P.; Lee, J.; Nip, L.; Koseki, S.R.T.; Tysinger, E.; Sontheimer, E.J.; Jacobson, J.M.; Jakimo, N. A Cas9 with PAM Recognition for Adenine Dinucleotides. *Nat. Commun.* **2020**, *11*, 2474, doi:10.1038/s41467-020-16117-8.

310. Anders, C.; Bargsten, K.; Jinek, M. Structural Plasticity of PAM Recognition by Engineered Variants of the RNA-Guided Endonuclease Cas9. *Mol. Cell* **2016**, *61*, 895–902, doi:10.1016/j.molcel.2016.02.020.

311. Ma, D.; Xu, Z.; Zhang, Z.; Chen, X.; Zeng, X.; Zhang, Y.; Deng, T.; Ren, M.; Sun, Z.; Jiang, R.; et al. Engineer Chimeric Cas9 to Expand PAM Recognition Based on Evolutionary Information. *Nat. Commun.* **2019**, *10*, 560, doi:10.1038/s41467-019-08395-8.

312. Hu, J.H.; Miller, S.M.; Geurts, M.H.; Tang, W.; Chen, L.; Sun, N.; Zeina, C.M.; Gao, X.; Rees, H.A.; Lin, Z.; et al. Evolved Cas9 Variants with Broad PAM Compatibility and

High DNA Specificity. *Nature* **2018**, *556*, 57–63, doi:10.1038/nature26155.

313. Thuronyi, B.W.; Koblan, L.W.; Levy, J.M.; Yeh, W.-H.; Zheng, C.; Newby, G.A.; Wilson, C.; Bhaumik, M.; Shubina-Oleinik, O.; Holt, J.R.; et al. Continuous Evolution of Base Editors with Expanded Target Compatibility and Improved Activity. *Nat. Biotechnol.* **2019**, *37*, 1070–1079, doi:10.1038/s41587-019-0193-0.

314. Walton, R.T.; Christie, K.A.; Whittaker, M.N.; Kleinstiver, B.P. Unconstrained Genome Targeting with Near-PAMless Engineered CRISPR-Cas9 Variants. *Science (80-. ).* **2020**, *368*, 290–296, doi:10.1126/science.aba8853.

315. Li, X.; Wang, Y.; Liu, Y.; Yang, B.; Wang, X.; Wei, J.; Lu, Z.; Zhang, Y.; Wu, J.; Huang, X.; et al. Base Editing with a Cpf1-Cytidine Deaminase Fusion. *Nat. Biotechnol.* **2018**, *36*, 324–327, doi:10.1038/nbt.4102.

316. Veillet, F.; Perrot, L.; Guyon-Debast, A.; Kermarrec, M.P.; Chauvin, L.; Chauvin, J.E.; Gallois, J.L.; Mazier, M.; Nogué, F. Expanding the CRISPR Toolbox in P. Patens Using SpCas9-NG Variant and Application for Gene and Base Editing in Solanaceae Crops. *Int. J. Mol. Sci.* **2020**, *21*, doi:10.3390/ijms21031024.

317. Kim, H.K.; Lee, S.; Kim, Y.; Park, J.; Min, S.; Choi, J.W.; Huang, T.P.; Yoon, S.; Liu, D.R.; Kim, H.H. High-Throughput Analysis of the Activities of XCas9, SpCas9-NG and SpCas9 at Matched and Mismatched Target Sequences in Human Cells. *Nat. Biomed. Eng.* **2020**, *4*, 111–124, doi:10.1038/s41551-019-0505-1.

318. Li, Q.; Seys, F.M.; Minton, N.P.; Yang, J.; Jiang, Y.; Jiang, W.; Yang, S. CRISPR–Cas9 D10A Nickase-Assisted Base Editing in the Solvent Producer Clostridium Beijerinckii. *Biotechnol. Bioeng.* **2019**, *116*, 1475–1483, doi:10.1002/bit.26949.

319. Garrity; Cavanagh CRISPR Mechanism , Tufts University, 2014.

320. Zhang, Y.; Grosse-Honebrink, A.; Minton, N.P.; Choi, J.; Papoutsakis, E.; Antoniewicz, M. A Universal Mariner Transposon System for Forward Genetic Studies in the Genus Clostridium. *PLoS One* **2015**, *10*, e0122411, doi:10.1371/journal.pone.0122411.

321. Perret, S.; Casalot, L.; Fierobe, H.P.; Tardif, C.; Sabathe, F.; Belaich, J.P.; Belaich, A. Production of Heterologous and Chimeric Scaffoldins by Clostridium Acetobutylicum ATCC 824. *J. Bacteriol.* **2004**, *186*, 253–257, doi:10.1128/JB.186.1.253-257.2004.

322. Dong, H.; Tao, W.; Zhang, Y.; Li, Y. Development of an Anhydrotetracycline-Inducible Gene Expression System for Solvent-Producing Clostridium Acetobutylicum: A Useful Tool for Strain Engineering. *Metab. Eng.* **2012**, *14*, 59–67, doi:10.1016/j.ymben.2011.10.004.

323. Girbal, L.; Mortier-Barrière, I.; Raynaud, F.; Rouanet, C.; Croux, C.; Soucaille, P. Development of a Sensitive Gene Expression Reporter System and an Inducible Promoter-Repressor System for Clostridium Acetobutylicum. *Appl. Environ. Microbiol.* **2003**, *69*, 4985–4988, doi:10.1128/AEM.69.8.4985-4988.2003.

324. Hartman, A.H.; Liu, H.; Melville, S.B. Construction and Characterization of a Lactose-Inducible Promoter System for Controlled Gene Expression in Clostridium Perfringens. *Appl. Environ. Microbiol.* **2011**, *77*, 471–478, doi:10.1128/AEM.01536-10.

325. Banerjee, A.; Leang, C.; Ueki, T.; Nevin, K.P.; Lovley, D.R. Lactose-Inducible System for Metabolic Engineering of Clostridium Ljungdahlii. *Appl. Environ. Microbiol.* **2014**, *80*, 2410–2416, doi:10.1128/AEM.03666-13.

326. Desai, S.K.; Gallivan, J.P. Genetic Screens and Selections for Small Molecules Based on a Synthetic Riboswitch That Activates Protein Translation. *J. Am. Chem. Soc.* **2004**, *126*, 13247–13254, doi:10.1021/ja048634j.

327. Thattai, M.; Van Oudenaarden, A. Intrinsic Noise in Gene Regulatory Networks. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 8614–8619, doi:10.1073/pnas.151588598.

328. Kierzek, A.M.; Zaim, J.; Zielenkiewicz, P. The Effect of Transcription and Translation Initiation Frequencies on the Stochastic Fluctuations in Prokaryotic Gene Expression. *J. Biol. Chem.* **2001**, *276*, 8165–8172, doi:10.1074/jbc.M006264200.

329. Swain, P.S. Efficient Attenuation of Stochasticity in Gene Expression through Post-Transcriptional Control. *J. Mol. Biol.* **2004**, *344*, 965–976, doi:10.1016/j.jmb.2004.09.073.

330. Kamiura, R.; Toya, Y.; Matsuda, F.; Shimizu, H. Theophylline-Inducible Riboswitch Accurately Regulates Protein Expression at Low Level in Escherichia Coli. *Biotechnol. Lett.* **2019**, *41*, 743–751, doi:10.1007/s10529-019-02672-8.

331. Cui, W.; Han, L.; Cheng, J.; Liu, Z.; Zhou, L.; Guo, J.; Zhou, Z. Engineering an Inducible

Gene Expression System for Bacillus Subtilis from a Strong Constitutive Promoter and a Theophylline-Activated Synthetic Riboswitch. *Microb. Cell Fact.* **2016**, *15*, 199, doi:10.1186/s12934-016-0599-z.

332. Hwang, G.H.; Park, J.; Lim, K.; Kim, S.; Yu, J.; Kim, S.T.; Eils, R.; Kim, J.S.; Bae, S. Web-Based Design and Analysis Tools for CRISPR Base Editing. *bioRxiv* **2018**, 1–7, doi:10.1101/373944.

333. Roberts, R.J.; Vincze, T.; Posfai, J.; Macelis, D. REBASE: Restriction Enzymes and Methyltransferases. *Nucleic Acids Res.* **2003**, *31*, 418–420.

334. Roberts, R.J.; Vincze, T.; Posfai, J.; Macelis, D. REBASE--a Database for DNA Restriction and Modification: Enzymes, Genes and Genomes. *Nucleic Acids Res.* **2015**, *43*, D298-9, doi:10.1093/nar/gku1046.

335. Müller-Hill, B.; Crapo, L.; Gilbert, W. Mutants That Make More Lac Repressor. *Proc. Natl. Acad. Sci. U. S. A.* **1968**, *59*, 1259–1264.

336. Anzalone, A. V.; Randolph, P.B.; Davis, J.R.; Sousa, A.A.; Koblan, L.W.; Levy, J.M.; Chen, P.J.; Wilson, C.; Newby, G.A.; Raguram, A.; et al. Search-and-Replace Genome Editing without Double-Strand Breaks or Donor DNA. *Nature* **2019**, *576*, 149–157, doi:10.1038/s41586-019-1711-4.

337. Arroyo-Olarte, R.D.; Bravo Rodríguez, R.; Morales-Ríos, E. Genome Editing in Bacteria: Crispr-Cas and Beyond. *Microorganisms* **2021**, *9*, doi:10.3390/microorganisms9040844.

338. Opgenorth, P.; Costello, Z.; Okada, T.; Goyal, G.; Chen, Y.; Gin, J.; Benites, V.; De Raad, M.; Northen, T.R.; Deng, K.; et al. Lessons from Two Design-Build-Test-Learn Cycles of Dodecanol Production in Escherichia Coli Aided by Machine Learning. *ACS Synth. Biol.* **2019**, *8*, 1337–1351, doi:10.1021/acssynbio.9b00020.

339. Adiego-Pérez, B.; Randazzo, P.; Daran, J.M.; Verwaal, R.; Roubos, J.A.; Daran-Lapujade, P.; Van Der Oost, J. Multiplex Genome Editing of Microorganisms Using CRISPR-Cas. *FEMS Microbiol. Lett.* **2019**, *366*, 1–19, doi:10.1093/femsle/fnz086.

340. Jakočiunas, T.; Rajkumar, A.S.; Zhang, J.; Arsovska, D.; Rodriguez, A.; Jendresen, C.B.; Skjødt, M.L.; Nielsen, A.T.; Borodina, I.; Jensen, M.K.; et al. CasEMBLR: Cas9-Facilitated Multiloci Genomic Integration of in Vivo Assembled DNA Parts in

Saccharomyces Cerevisiae. *ACS Synth. Biol.* **2015**, *4*, 1126–1134, doi:10.1021/acssynbio.5b00007.

341. Kurata, M.; Wolf, N.K.; Lahr, W.S.; Weg, M.T.; Kluesner, M.G.; Lee, S.; Hui, K.; Shiraiwa, M.; Webber, B.R.; Moriarity, B.S. Highly Multiplexed Genome Engineering Using CRISPR/Cas9 GRNA Arrays. *PLoS One* **2018**, *13*, e0198714, doi:10.1371/journal.pone.0198714.

342. Liu, H.; Hou, G.; Wang, P.; Guo, G.; Wang, Y.; Yang, N.; Rehman, M.N.U.; Li, C.; Li, Q.; Zheng, J.; et al. A Double-Locus Scarless Genome Editing System in Escherichia Coli. *Biotechnol. Lett.* **2020**, *42*, 1457–1465, doi:10.1007/s10529-020-02856-7.

343. Ding, T.; Huang, C.; Liang, Z.; Ma, X.; Wang, N.; Huo, Y.X. Reversed Paired-GRNA Plasmid Cloning Strategy for Efficient Genome Editing in Escherichia Coli. *Microb. Cell Fact.* **2020**, *19*, 1–13, doi:10.1186/s12934-020-01321-4.

344. Sakuma, T.; Nishikawa, A.; Kume, S.; Chayama, K.; Yamamoto, T. Multiplex Genome Engineering in Human Cells Using All-in-One CRISPR/Cas9 Vector System. *Sci. Rep.* **2015**, *4*, 5400, doi:10.1038/srep05400.

345. Walter, J.M.; Chandran, S.S.; Horwitz, A.A. CRISPR-Cas-Assisted Multiplexing (CAM): Simple Same-Day Multi-Locus Engineering in Yeast. *J. Cell. Physiol.* **2016**, *231*, 2563–2569, doi:10.1002/jcp.25375.

346. Ferreira, R.; Skrekas, C.; Nielsen, J.; David, F. Multiplexed CRISPR/Cas9 Genome Editing and Gene Regulation Using Csy4 in Saccharomyces Cerevisiae. *ACS Synth. Biol.* **2018**, *7*, 10–15, doi:10.1021/acssynbio.7b00259.

347. Haurwitz, R.E.; Jinek, M.; Wiedenheft, B.; Zhou, K.; Doudna, J.A. Sequence- and Structure-Specific RNA Processing by a CRISPR Endonuclease. *Science (80-. ).* **2010**, doi:10.1126/science.1192272.

348. Zhang, Y.; Wang, J.; Wang, Z.; Zhang, Y.; Shi, S.; Nielsen, J.; Liu, Z. A GRNA-TRNA Array for CRISPR-Cas9 Based Rapid Multiplexed Genome Editing in Saccharomyces Cerevisiae. *Nat. Commun.* **2019**, *10*, 1053, doi:10.1038/s41467-019-09005-3.

349. Xie, K.; Minkenberg, B.; Yang, Y. Boosting CRISPR/Cas9 Multiplex Editing Capability with the Endogenous TRNA-Processing System. *Proc. Natl. Acad. Sci.* **2015**, *112*, 3570–3575, doi:10.1073/pnas.1420294112.

350. Dong, F.; Xie, K.; Chen, Y.; Yang, Y.; Mao, Y. Polycistronic TRNA and CRISPR Guide-RNA Enables Highly Efficient Multiplexed Genome Engineering in Human Cells. *Biochem. Biophys. Res. Commun.* **2017**, *482*, 889–895, doi:10.1016/j.bbrc.2016.11.129.

351. Shepherd, J.; Ibba, M. Bacterial Transfer RNAs. *FEMS Microbiol. Rev.* 2015, *39*, 280–300.

352. Marvin, M.C.; Engelke, D.R. RNase P: Increased Versatility through Protein Complexity? *RNA Biol.* **2009**, *6*, 40–42.

353. Hartmann, R.K.; Gößringer, M.; Späth, B.; Fischer, S.; Marchfelder, A. Chapter 8 The Making of TRNAs and More – RNase P and TRNase Z. In *Progress in Molecular Biology and Translational Science*; 2009; Vol. 85, pp. 319–368.

354. White, R.J. Transcription by RNA Polymerase III: More Complex than We Thought. *Nat. Rev. Genet.* **2011**, *12*, 459–463, doi:10.1038/nrg3001.

355. Turowski, T.W.; Tollervey, D. Transcription by RNA Polymerase III: Insights into Mechanism and Regulation. *Biochem. Soc. Trans.* **2016**, *44*, 1367–1375, doi:10.1042/BST20160062.

356. Schwartz, C.M.; Hussain, M.S.; Blenner, M.; Wheeldon, I. Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia Lipolytica. *ACS Synth. Biol.* **2016**, *5*, 356–359, doi:10.1021/acssynbio.5b00162.

357. Sekiya, T.; Contreras, R.; Takeya, T.; Khorana, H.G. Total Synthesis of a Tyrosine Suppressor Transfer RNA Gene. *J. Biol. Chem.* **1979**, *254*, 5802–5816.

358. Mörl, M.; Marchfelder, A. The Final Cut. The Importance of TRNA 3'-Processing. *EMBO Rep.* 2001, *2*, 17–20.

359. Mackie, G.A. RNase E: At the Interface of Bacterial RNA Processing and Decay. *Nat. Rev. Microbiol.* **2013**, *11*, 45–57, doi:10.1038/nrmicro2930.

360. Li, Z.; Deutscher, M.P. RNase E Plays an Essential Role in the Maturation of Escherichia Coli TRNA Precursors. *Rna* **2002**, *8*, 97–109, doi:10.1017/S1355838202014929.

361. Ow, M.C.; Kushner, S.R. Initiation of TRNA Maturation by RNase E Is Essential for Cell Viability in E. Coli. *Genes Dev.* **2002**, *16*, 1102–1115, doi:10.1101/gad.983502.

362. Apirion, D.; Miczak, A. RNA Processing in Prokaryotic Cells. *BioEssays* 1993, *15*, 113–120.

363. Gegenheimer, P.; Apirion, D. Processing of Procaryotic Ribonucleic Acid. *Microbiol. Rev.* **1981**, *45*, 502–541.

364. Altman, S. Biosynthesis of Transfer RNA in Escherichia Coli. *Cell* 1975, *4*, 21–29.

365. Li, Z.; Gong, X.; Joshi, V.H.; Li, M. Co-Evolution of TRNA 3' Trailer Sequences with 3' Processing Enzymes in Bacteria. *RNA* **2005**, *11*, 567–577, doi:10.1261/rna.7287505.

366. Minagawa, A.; Takaku, H.; Takagi, M.; Nashimoto, M. A Novel Endonucleolytic Mechanism to Generate the CCA 3' Termini of TRNA Molecules in Thermotoga Maritima. *J. Biol. Chem.* **2004**, *279*, 15688–15697, doi:10.1074/jbc.M313951200.

367. Madeira, F.; Park, Y.M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A.R.N.; Potter, S.C.; Finn, R.D.; et al. The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019. *Nucleic Acids Res.* **2019**, *47*, W636–W641, doi:10.1093/nar/gkz268.

368. Ryan, O.W.; Skerker, J.M.; Maurer, M.J.; Li, X.; Tsai, J.C.; Poddar, S.; Lee, M.E.; DeLoache, W.; Dueber, J.E.; Arkin, A.P.; et al. Selection of Chromosomal DNA Libraries Using a Multiplex CRISPR System. *Elife* **2014**, *3*, 25–36, doi:10.7554/eLife.03703.

369. Ryan, O.W.; Cate, J.H.D. Multiplex Engineering of Industrial Yeast Genomes Using CRISPRm. *Methods Enzymol.* **2014**, *546*, 473–489, doi:10.1016/B978-0-12-801185-0.00023-4.

370. Gao, Y.; Zhao, Y. Self-Processing of Ribozyme-Flanked RNAs into Guide RNAs in Vitro and in Vivo for CRISPR-Mediated Genome Editing. *J. Integr. Plant Biol.* **2014**, *56*, 343–349, doi:10.1111/jipb.12152.

371. de Vries, A.R.G.; de Groot, P.A.; van den Broek, M.; Daran, J.M.G. CRISPR-Cas9 Mediated Gene Deletions in Lager Yeast Saccharomyces Pastorianus. *Microb. Cell Fact.* **2017**, *16*, 1–18, doi:10.1186/s12934-017-0835-1.

372. Juergens, H.; Varela, J.A.; de Vries, A.R.G.; Perli, T.; Gast, V.J.M.; Gyurchev, N.Y.; Rajkumar, A.S.; Mans, R.; Pronk, J.T.; Morrissey, J.P.; et al. Genome Editing in Kluyveromyces and Ogataea Yeasts Using a Broad-Host-Range Cas9/GRNA Co-Expression Plasmid. *FEMS Yeast Res.* **2018**, *18*, 1–16, doi:10.1093/femsyr/foy012.

373. Karvelis, T.; Gasiunas, G.; Miksys, A.; Barrangou, R.; Horvath, P.; Siksnys, V. CrRNA and TracrRNA Guide Cas9-Mediated DNA Interference in Streptococcus Thermophilus. *RNA Biol.* **2013**, *10*, 841–851, doi:10.4161/rna.24203.

374. Charpentier, E.; Richter, H.; van der Oost, J.; White, M.F. Biogenesis Pathways of RNA Guides in Archaeal and Bacterial CRISPR-Cas Adaptive Immunity. *FEMS Microbiol. Rev.* 2015, *39*, 428–441.

375. Court, D.L.; Gan, J.; Liang, Y.-H.; Shaw, G.X.; Tropea, J.E.; Costantino, N.; Waugh, D.S.; Ji, X. RNase III: Genetics and Function; Structure and Mechanism. *Annu. Rev. Genet.* **2013**, *47*, 405–431, doi:10.1146/annurev-genet-110711-155618.

376. Conrad, C.; Rauhut, R. Ribonuclease III: New Sense from Nuisance. *Int. J. Biochem. Cell Biol.* **2002**, *34*, 116–129, doi:10.1016/S1357-2725(01)00112-1.

377. Cress, B.F.; Toparlak, Ö.D.; Guleria, S.; Lebovich, M.; Stieglitz, J.T.; Englaender, J.A.; Jones, J.A.; Linhardt, R.J.; Koffas, M.A.G. CRISPathBrick: Modular Combinatorial Assembly of Type II-A CRISPR Arrays for DCas9-Mediated Multiplex Transcriptional Repression in *E. Coli*. *ACS Synth. Biol.* **2015**, *4*, 987–1000, doi:10.1021/acssynbio.5b00012.

378. Jin, S.; Bae, J.; Song, Y.; Pearcy, N.; Shin, J.; Kang, S.; Minton, N.P.; Soucaille, P.; Cho, B.-K. Synthetic Biology on Acetogenic Bacteria for Highly Efficient Conversion of C1 Gases to Biochemicals. *Int. J. Mol. Sci.* **2020**, *21*, 7639, doi:10.3390/ijms21207639.

379. Zetsche, B.; Heidenreich, M.; Mohanraju, P.; Fedorova, I.; Kneppers, J.; DeGennaro, E.M.; Winblad, N.; Choudhury, S.R.; Abudayyeh, O.O.; Gootenberg, J.S.; et al. Multiplex Gene Editing by CRISPR-Cpf1 Using a Single CrRNA Array. *Nat. Biotechnol.* **2017**, *35*, 31–34, doi:10.1038/nbt.3737.

380. Yamano, T.; Nishimasu, H.; Zetsche, B.; Hirano, H.; Slaymaker, I.M.; Li, Y.; Fedorova, I.; Nakane, T.; Makarova, K.S.; Koonin, E. V; et al. Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **2016**, *165*, 949–962, doi:10.1016/j.cell.2016.04.003.

381. Zetsche, B.; Gootenberg, J.S.; Abudayyeh, O.O.; Slaymaker, I.M.; Makarova, K.S.; Essletzbichler, P.; Volz, S.E.; Joung, J.; Van Der Oost, J.; Regev, A.; et al. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **2015**, *163*, 759–771, doi:10.1016/j.cell.2015.09.038.

382. Moreno-Mateos, M.A.; Fernandez, J.P.; Rouet, R.; Vejnar, C.E.; Lane, M.A.; Mis, E.; Khokha, M.K.; Doudna, J.A.; Giraldez, A.J. CRISPR-Cpf1 Mediates Efficient Homology-Directed Repair and Temperature-Controlled Genome Editing. *Nat. Commun.* **2017**, *8*, 2024, doi:10.1038/s41467-017-01836-2.

383. Paul, B.; Montoya, G. CRISPR-Cas12a: Functional Overview and Applications. *Biomed. J.* **2020**, *43*, 8–17, doi:10.1016/j.bj.2019.10.005.

384. Meliawati, M.; Schilling, C.; Schmid, J. Recent Advances of Cas12a Applications in Bacteria. *Appl. Microbiol. Biotechnol.* **2021**, *105*, 2981–2990, doi:10.1007/s00253-021-11243-9.

385. Zhao, R.; Liu, Y.; Zhang, H.; Chai, C.; Wang, J.; Jiang, W.; Gu, Y. CRISPR-Cas12a-Mediated Gene Deletion and Regulation in Clostridium Ljungdahlii and Its Application in Carbon Flux Redirection in Synthesis Gas Fermentation. *ACS Synth. Biol.* **2019**, *8*, 2270–2279, doi:10.1021/acssynbio.9b00033.

386. Grissa, I.; Vergnaud, G.; Pourcel, C. The CRISPRdb Database and Tools to Display CRISPRs and to Generate Dictionaries of Spacers and Repeats. *BMC Bioinformatics* **2007**, *8*, 172, doi:10.1186/1471-2105-8-172.

387. Grissa, I.; Vergnaud, G.; Pourcel, C. CRISPRFinder: A Web Tool to Identify Clustered Regularly Interspaced Short Palindromic Repeats. *Nucleic Acids Res.* **2007**, *35*, 52–57, doi:10.1093/nar/gkm360.

388. Xu, Z.; Kuang, Y.; Ren, B.; Yan, D.; Yan, F.; Spetz, C.; Sun, W.; Wang, G.; Zhou, X.; Zhou, H. SpRY Greatly Expands the Genome Editing Scope in Rice with Highly Flexible PAM Recognition. *Genome Biol.* **2021**, *22*, 1–15, doi:10.1186/s13059-020-02231-9.

389. Woods, C.; Humphreys, C.M.; Tomi-Andrino, C.; Henstra, A.M.; Köpke, M.; Simpson, S.D.; Winzer, K.; Minton, N.P. Application of Transposon-Insertion Sequencing to Determine Gene Essentiality in the Acetogen Clostridium Autoethanogenum. *bioRxiv* **2021**, 1–32, doi:https://doi.org/10.1101/2021.05.19.444907.

390. Norman, R.O.J. (University of N. Genome-Scale Metabolic Modelling of Clostridium Autoethanogenum, University of Nottingham, 2019.

391. Bao, Z.; Xiao, H.; Liang, J.; Zhang, L.; Xiong, X.; Sun, N.; Si, T.; Zhao, H. Homology-Integrated CRISPR-Cas (HI-CRISPR) System for One-Step Multigene Disruption in Saccharomyces Cerevisiae. *ACS Synth. Biol.* **2015**, *4*, 585–594, doi:10.1021/sb500255k.

392. Chan, P.P.; Lowe, T.M. GtRNAdb: A Database of Transfer RNA Genes Detected in Genomic Sequence. *Nucleic Acids Res.* **2009**, *37*, D93–D97, doi:10.1093/nar/gkn787.

393. Chan, P.P.; Lowe, T.M. GtRNAdb 2.0: An Expanded Database of Transfer RNA Genes Identified in Complete and Draft Genomes. *Nucleic Acids Res.* **2016**, *44*, D184–D189, doi:10.1093/nar/gkv1309.

394. Gruber, A.R.; Lorenz, R.; Bernhart, S.H.; Neubock, R.; Hofacker, I.L. The Vienna RNA Websuite. *Nucleic Acids Res.* **2008**, *36*, W70–W74, doi:10.1093/nar/gkn188.

395. Jiang, F.; Doudna, J.A. CRISPR-Cas9 Structures and Mechanisms. **2017**, doi:10.1146/annurev-biophys.

396. Kim, Y.B.; Komor, A.C.; Levy, J.M.; Packer, M.S.; Zhao, K.T.; Liu, D.R. Increasing the Genome-Targeting Scope and Precision of Base Editing with Engineered Cas9-Cytidine Deaminase Fusions. *Nat. Biotechnol.* **2017**, *35*, 371–376, doi:10.1038/nbt.3803.

397. Liu, Z.; Chen, M.; Shan, H.; Chen, S.; Xu, Y.; Song, Y.; Zhang, Q.; Yuan, H.; Ouyang, H.; Li, Z.; et al. Expanded Targeting Scope and Enhanced Base Editing Efficiency in Rabbit Using Optimized XCas9(3.7). *Cell. Mol. Life Sci.* **2019**, *76*, 4155–4164, doi:10.1007/s00018-019-03110-8.

398. Ren, Q.; Sretenovic, S.; Liu, S.; Tang, X.; Huang, L.; He, Y.; Liu, L.; Guo, Y.; Zhong, Z.; Liu, G.; et al. PAM-Less Plant Genome Editing Using a CRISPR–SpRY Toolbox. *Nat. Plants* **2021**, *7*, 25–33, doi:10.1038/s41477-020-00827-4.

399. Larson, S.B.; Jones, J.A.; McPherson, A. The Structure of an Iron-Containing Alcohol Dehydrogenase from a Hyperthermophilic Archaeon in Two Chemical States. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2019**, *75*, 217–226, doi:10.1107/S2053230X19001201.

400. Guex, N.; Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: An Environment for Comparative Protein Modeling. *Electrophoresis* **1997**, *18*, 2714–2723, doi:10.1002/elps.1150181505.

401. Rizvi, S.-R.-Z. Studying Novel Sporulation Genes in Clostridium Sporogenes, University of Nottingham, 2020.

# Chapter VIII: Supplementary material

## VIII.1. Algorithm descriptions

### VIII.1.1. Target-AID sgRNA design

The Target-AID sgRNA design Excel table was constructed so as to be as user-friendly as possible. All steps were made explicit, so that the end-user could approach the algorithm critically. In addition to the sgRNA design table from Benchling (of which an example was given in Table 12), it only requires the length of the gene being analysed and the direction of the coding strand (1 if it is in the forward direction, -1 if it is in the reverse-complementary direction) (cf. Table S. 1).

**Table S. 1: Additional input required for the Target-AID sgRNA design Excel algorithm, using *C. autoethanogenum pyrE* as an example.** [@CDS] and [@Length] define each variable in the algorithm.

| | | |
|---|---|---|
| ***pyrE*** | Coding strand `[@CDS]` | -1 |
| | Gene length `[@Length]` | 573 |

gRNAs compatible with Target-AID are isolated by the algorithm in five simple steps:

1. Extract the position of the first cytosine in the protospacer sequence, and return the string “No C” if there is no cytosine in the protospacer.

```
[@C1]=IFERROR(SEARCH("c",[@Sequence],1),"No C")
```

2. Find out if the position of the cytosine is within the predefined editing window (in positions -20 to -16 from the PAM, or 1 to 5 from the start of the protospacer), and calculate the position of the base relative to the start of the gene. If the cytosine falls out of the editing window, return `"C]16-20["`, and if there was no C in the protospacer to start with, return an empty space (`""`).

```
[@C1[16-20]]=
IF(ISNUMBER([@C1]),
   IF([@C1]<=5,
      IF([@CDS]<0,
         IF([@Strand]<0,
            ([@Length]-([@Position]+17-[@C1])+1,
            ([@Length]-([@Position]-18+[@C1])+1),
         IF([@Strand]>0,
            [@Position]-18+[@C1],
            [@Position]+17-[@C1])),
      "C]16-20["),
   "")
```

3. Calculate the position (or rank) of the base within its codon (in between 1 and 3), using the size of the gene and its orientation given in Table S. 1. If there was no cytosine within the editing window, return an empty space.

```
[@Rank]=
IF(ISNUMBER([@[C1'[16-20']]]),
   IF(INT(([@[C1'[16-20']]]+0)/3
      =([@[C1'[16-20']]]+0)/3,
      3,
      IF(INT(([@[C1'[16-20']]]+1)/3)
         =([@[C1'[16-20']]]+1)/3,
         2,
         IF(INT(([@[C1'[16-20']]]+2)/3)
            =([@[C1'[16-20']]]+2)/3,
            1))),
   " ")
```

4. Extract the sequence of the base triplet which composes the codon. If the protospacer is on the non-coding strand, this triplet is actually the reverse-complement of the codon sequence. To reflect both possibilities, the variable is simply designated as “Triplet”. If there was no rank value from the previous step, return the code “#N/A”. In some cases, no triplet can be defined because it would start or end outside of the sgRNA sequence. Then, the default error “#Value!” is returned.

```
[@Triplet]=
IF(ISNUMBER([@[Rank]),
   IF([@CDS]=[@Strand],
      MID([@Sequence],[@C1]+1-[@[Rank]],3),
      MID([@Sequence],[@C1]-3+[@[Rank]],3)),
   NA())
```

5. Finally, compare this base triplet with a reference table containing all the codons which, if their cytosine is mutated into a thymine, result in a nonsense mutation (cf. Table S. 2). If such a codon is identified, return “Yes”. If it is not identified, return “No”. If the codon defined in the previous step was incomplete (i.e. if its value was the error “#Value!”), the algorithm returns the instruction “To check”, directing the operator to verify that protospacer manually. At last, if no base triplet was identified in the previous stage, it returns another blank value (“ “).

```
[@Nonsense?]=
IF(ISTEXT(L14),
   IF([@Strand]=[@CDS],
      IFERROR(VLOOKUP([@["Triplet"]],
                      Coding,3,FALSE),
              "No"),
      IFERROR(VLOOKUP([@["Triplet"]],
                      Non-coding,3,FALSE),
              "No")),
   IF(ISERR([@["Triplet"]]), "To check"," "))
```

**Table S. 2: Reference table of Target-AID-inducible nonsense mutations.** The vlookup() function compares a triplet with a cytosine ([@Codon]) with the WT column of the appropriate strand (Coding or Non-coding). If the triplet matches a triplet which can be mutated by Target-AID to produce a STOP codon (TAA, TAG or TGA), then it returns the value in the "Nonsense" column (which is always "Yes"). The "Mutated" column is not used in the algorithm but it illustrates which STOP codon can be expected from each mutated triplet.

| Strand | WT | Mutated | Nonsense |
|---|---|---|---|
| Coding | CAA | TAA | Yes |
| | CAG | TAG | Yes |
| | CGA | TGA | Yes |
| Non-coding | CCA | TAA | Yes |
| | CTA | TAA | Yes |
| | TCA | TAA | Yes |
| | CCA | TAG | Yes |
| | CCA | TGA | Yes |

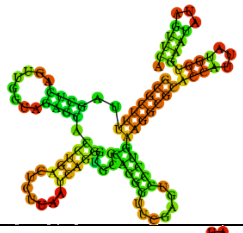
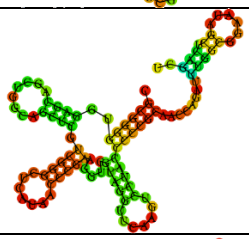
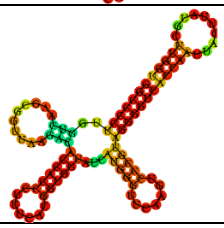
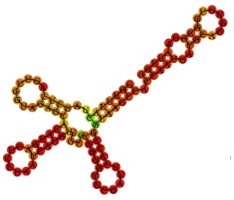
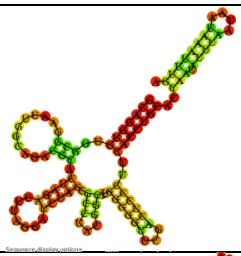
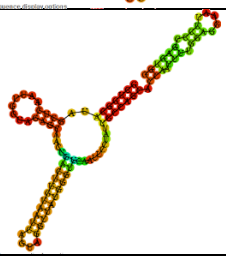
The whole algorithm is then repeated two additional times in order to test more cytosines in the same protospacers.

```
[@C2]=
IF(ISNUMBER([@C1]),
   IFERROR(SEARCH("c",[@Sequence],[@C1]+1),""),
   " ")

[@C3]=
IF(ISNUMBER([@C2]),
   IFERROR(SEARCH("c",[@Sequence],[@C2]+1),""),
   " ")
```

### VIII.1.2. tRNA decision matrix

**Table S. 3: Summary and decision matrix of the pre-tRNAs considered for use in the TA_MultitRNA construct.** The selection criteria and the bioinformatics tools that have been used are detailed in **Table 18**. tRNA-Thr-TGT-1-1 and tRNA-fMet-CAT-1-1 ended up being selected.

| Name | Structure | Stem-loop | Poly-U | sgRNA OK | Codon usage | Homology |
|---|---|---|---|---|---|---|
| **tRNA-Lys-CTT-1-1** |  | 19 bp AU-rich short | No | No | 25/1000 (14th) | None found |
| **tRNA-fMet-CAT-1-1** |  | 22bp AU-rich short | No | Yes | 26/1000 (13th) | 96% |
| **tRNA-Glu-TTC-1-1** |  | 16 bp AT rich loop; strong GC-rich stem | No | Not attempted | 53/1000 (3$^{rd}$) | 89% |
| **tRNA-Lys-TTT-1-1** |  | 24 bp strong GC-rich long | No | Not attempted | 67/1000 (1$^{st}$) | 99% |
| **tRNA-Leu-TAG-2-1** |  | 20 bp AU-rich long | Maybe? | Not attempted | 9/1000 (35$^{th}$) | 94% |
| **tRNA-Thr-TGT-1-1** |  | 23 bp AU-rich long | No | Yes | 20/1000 (19th) | 89% |

### VIII.1.3. Base editing coverage algorithm

The code written by Claudio Tomi Andrino is collecting and organizing the sequences of all the protospacers which could produce a STOP codon through Target-AID base editing, then simply summing the number of CDS's containing at least one protospacer and dividing that by the total number of CDS's (3964).

In order to find all the protospacers which can lead to a STOP codon, the sense and antisense strands are processed separately. The algorithm proceeds as described in the two following subsections.

#### *VIII.1.3.1) Sense strand*

First, all the PAMs are detected and sorted according to their category (`NGG`, `NGD`,`NG`, `NGA`, `NAA`), as well as their presence on the sense or antisense strand. The protospacers on each strand are then processed separately. The distance of each PAM from to the start of the gene is calculated, both in terms of length (in bp) and as a proportion of the whole gene. This is can later be used to discard all the PAMs which located at a lower distance than the arbitrary 75% threshold.

The 21 bases upstream of the PAM are extracted. The sequence of the protospacer is directly adjusted to only be composed of full, in-frame codons. It is then divided into the bases upstream of the editing window, inside the editing window, and after the editing window. All the `C`'s inside the editing window are changed into `T`'s, and the mutant protospacer is assembled by combining the mutant editing window with the upstream and downstream sequences once again. At last, the sequence of the resulting mutant protospacer is screened with the appropriate reading frame, and flagged if a STOP codon is found.

#### *VIII.1.3.2) Antisense strand*

To find protospacers on the antisense strand, the reverse-complementary sequence of the CDS is first being computed in order to proceed to the *in silico* mutagenesis. The algorithm then proceeds much in the same way as with the sense strand. When the mutant protospacer is finally obtained, the sequence is once again replaced by its reverse-complementary in order to restore the sense sequence, and the search for the STOP codon proceeds as previously described.

## VIII.2. Key sequences

### VIII.2.1. Target-AID

#### *VIII.2.1.1) nCas9(D10A)-RSKO*

ATGGATAAGAAATACTCAATAGGCTTAGcTATCGGCACAAATAGCGTCGGATGGGCGGTGATCACTGATGAATAT
AAGGTTCCGTCTAAAAAGTTCAAGGTTCTGGGAAATACAGACCGCCACAGTATCAAAAAAAATCTTATAGGGGCT
CTTTTATTTGACAGTGGAGAGACAGCGGAAGCGACTCGTCTCAAACGGACAGCTCGTAGAAGGTATACACGTCGG
AAGAATCGTATTTGTTATCTACAGGAGATTTTTTCAAATGAGATGGCGAAAGTAGATGATAGTTTCTTTCATCGA
CTTGAAGAGTCTTTTTTGGTGGAAGAAGACAAGAAGCATGAACGTCATCCTATTTTTGGAAATATAGTAGATGAA
GTTGCTTATCATGAGAAATATCCAACTATCTATCATCTGCGAAAAAAATTGGTAGATTCTACTGATAAAGCGGAT
TTGCGCTTAATCTATTTGGCCTTAGCGCAcATGATTAAGTTTCGTGGTCATTTTTTGATTGAGGGAGATTTAAAT
CCTGATAATAGTGATGTGGACAAACTATTTATCCAGTTGGTACAAACCTACAATCAATTATTTGAAGAAAACCCT
ATaAACGCAAGTGGAGTAGATGCTAAAGCGATTCTTTCTGCACGATTGAGTAAATCAAGACGATTAGAAAATCTC
ATTGCTCAGCTCCCCGGTGAGAAGAAAAATGGCTTATTTGGGAATCTCATTGCTTTGTCATTGGGTTTGACCCCT
AATTTTAAATCAAATTTTGATTTGGCAGAAGATGCTAAATTACAGCTTTCAAAAGATACTTACGATGATGATTTA
GATAATTTATTGGCGCAAATTGGAGATCAATATGCTGATTTGTTTTTGGCAGCTAAGAATTTATCAGATGCTATT
TTACTTTCAGATATCCTAAGAGTAAATACTGAAATAACTAAGGCTCCCCTATCAGCTTCAATGATTAAACGCTAC
GATGAACATCATCAAGACTTGACTCTTTTAAAAGCTTTAGTTCGACAACAACTTCCAGAAAAGTATAAAGAAATC
TTTTTTGATCAATCAAAAAACGGATATGCAGGTTATATTGATGGGGGAGCTAGCCAAGAAGAATTTTATAAATTT
ATCAAACCAATTTTAGAAAAAATGGATGGTACTGAGGAATTATTGGTGAAACTAAATCGTGAAGATTTGCTGCGC
AAGCAACGGACCTTTGACAACGGCTCTATTCCCCATCAAATTCACTTGGGTGAGCTGCATGCTATTTTGAGAAGA
CAAGAAGACTTTTATCCATTTTTAAAAGACAATCGTGAGAAGATTGAAAAAATCTTGACTTTTCGAATTCCTTAT
TATGTTGGTCCATTGGCGCGTGGCAATAGTCGTTTTGCATGGATGACTCGGAAGTCTGAAGAAACAATTACCCCA
TGGAATTTTGAAGAAGTTGTCGATAAAGGTGCTTCAGCTCAATCATTTATTGAACGCATGACAAACTTTGATAAA
AATCTTCCAAATGAAAAAGTACTACCAAAACATAGTTTGCTTTATGAGTATTTTACGGTTTATAACGAATTGACA
AAGGTCAAATATGTTACTGAAGGAATGCGAAAACCAGCATTTCTTTCAGGTGAACAGAAGAAAGCCATTGTTGAT
TTACTCTTCAAAACAAATCGAAAAGTAACCGTTAAGCAATTAAAAGAAGATTATTTCAAAAAAATAGAATGTTTT
GATAGTGTTGAAATTTCAGGAGTTGAAGATAGATTTAATGCTTCATTAGGTACCTACCATGATTTGCTAAAAATT
ATTAAAGATAAAGATTTTTTGGATAATGAAGAAAATGAAGATATCTTAGAGGATATTGTTTTAACATTGACCTTA
TTTGAAGATAGGGAGATGATTGAGGAAAGACTTAAAACtTATGCTCACCTCTTTGATGATAAGGTGATGAAACAG
CTTAAACGTCGCCGTTATACTGGTTGGGGACGTTTGTCTCGAAAATTGATTAATGGTATTAGGGATAAGCAATCT
GGCAAAACAATATTAGATTTTTTGAAATCAGATGGTTTTGCCAATCGCAATTTTATGCAGCTGATCCATGATGAT
AGTTTGACATTTAAAGAAGACATTCAAAAAGCACAAGTGTCTGGACAAGGCGATAGTTTACATGAACATATTGCA
AATTTAGCTGGTAGCCCTGCTATTAAAAAAGGTATTTTACAGACTGTAAAAGTTGTTGATGAATTGGTCAAAGTA
ATGGGGCGGCATAAGCCAGAAAATATCGTTATTGAAATGGCACGTGAAAATCAGACAACTCAAAAGGGCCAGAAA
AATTCGCGAGAGCGTATGAAACGAATCGAAGAAGGTATCAAAGAATTAGGAAGTCAGATTCTTAAAGAGCATCCT
GTTGAAAATACTCAATTGCAAAATGAAAAGCTCTATCTCTATTATCTCCAAAATGGAAGAGACATGTATGTGGAC
CAAGAATTAGATATTAATCGTTTAAGTGATTATGATGTCGATCACATTGTTCCACAAAGTTTCCTTAAAGACGAT
TCAATAGACAATAAGGTCTTAACGCGTTCTGATAAAAATCGTGGTAAATCGGATAACGTTCCAAGTGAAGAAGTA
GTCAAAAAGATGAAAAACTATTGGAGACAACTTCTAAACGCCAAaTTAATCACTCAACGTAAGTTTGATAATTTA
ACGAAAGCTGAACGTGGAGGTTTGAGTGAACTTGATAAAGCTGGTTTTATCAAACGCCAATTGGTTGAAACTCGC
CAAATCACTAAGCATGTGGCACAAATTTTGGATAGTCGCATGAATACTAAATACGATGAAAATGATAAACTTATT
CGAGAGGTTAAAGTGATTACCTTAAAATCTAAATTAGTTTCTGACTTCCGAAAAGATTTCCAATTCTATAAAGTA
CGTGAGATaAACAATTACCATCATGCCCATGATGCGTATCTAAATGCCGTCGTTGGAACTGCTTTGATTAAGAAA
TATCCAAAACTTGAATCGGAGTTTGTCTATGGTGATTATAAAGTTTATGATGTTCGTAAAATGATTGCTAAGTCT
GAGCAAGAAATAGGCAAAGCAACCGCAAAATATTTCTTTTACTCTAATATCATGAACTTCTTCAAAACAGAAATT
ACACTTGCAAATGGAGAGATTCGCAAACGCCCTCTAATCGAAACTAATGGGGAAACTGGAGAAATTGTCTGGGAT
AAAGGGCGAGATTTTGCCACAGTGCGCAAAGTATTGTCCATGCCCCAAGTCAATATTGTCAAGAAAACAGAAGTA
CAGACAGGCGGATTCTCCAAGGAGTCAATTTTACCAAAAAGAAATTCGGACAAGCTTATTGCTCGTAAAAAAGAC
TGGGATCCAAAAAAATATGGTGGTTTTGATAGTCCAACGGTAGCTTATTCAGTCCTAGTGGTTGCTAAGGTGGAA
AAAGGGAAATCGAAGAAGTTAAAATCCGTTAAAGAGTTACTAGGGATCACAATTATGGAAAGAAGTTCCTTTGAA
AAAAATCCGATTGACTTTTTAGAAGCTAAAGGATATAAGGAAGTTAAAAAAGACTTAATCATTAAACTACCTAAA
TATAGTCTTTTTGAGTTAGAAAACGGTCGTAAACGGATGCTGGCTAGTGCCGGAGAATTACAAAAAGGAAATGAG
CTGGCTCTGCCAAGCAAATATGTGAATTTTTTATATTTAGCTAGTCATTATGAAAAGTTGAAGGGTAGTCCAGAA
GATAACGAACAAAAACAATTGTTTGTGGAGCAGCATAAGCATTATTTAGATGAGATTATTGAGCAAATCAGTGAA
TTTTCTAAGCGTGTTATTTTAGCAGATGCCAATTTAGATAAAGTTCTTAGTGCATATAACAAACATAGAGACAAA
CCAATACGTGAACAAGCAGAAAATATTATTCATTTATTTACGTTGACGAATCTTGGAGCTCCCGCTGCTTTTAAA
TATTTTGATACAACAATTGATCGTAAACGATATACGTCTACAAAAGAAGTTTTAGATGCCACTCTTATCCATCAA
TCCATCACTGGTCTTTATGAAACACGCATTGATTTGAGTCAGCTAGGAGGTGAC

Purpose of SNPs compared to WT: D10A (inactivate RuvC); Disruption of Cau10061II; disruption of NdeI.

### *VIII.2.1.2) Codon-optimized PmCDA1*

agtagagcagatcctaagaagaaaagaaaagttggtggtggtggaagtggtggtggtggaagtgcagaatatgtt
agagcattatttgattttaatggaaatgatgaagaagatttacctttaagaaaggagatatacttagaataaga
gataaaccagaagaacaatggtggaatgcagaagattctgaaggaaaaagaggaatgatacctgtaccttatgta
gaaaaatattcaggagattataaagatcatgatggtgactacaaagatcatgatatagattataaagatgatgat
gataaaagtagaATGACTGATGCTGAATATGTTAGAATACATGAAAAATTAGATATATATACATTTAAAAAACAA
TTTTTCAATAATAAAAAATCTGTATCACATAGATGTTATGTATTATTTGAATTAAAAAGAAGAGGAGAAAGAAGA
GCATGTTTTTGGGGATATGCTGTAAATAAACCTCAATCAGGAACTGAAAGAGGAATACATGCAGAAATATTTAGT
ATAAGAAAAGTAGAAGAATATTTAAGAGATAATCCTGGACAATTTACAATAAACTGGTATAGTTCTTGGAGTCCA
TGTGCAGATTGTGCTGAAAAAATATTAGAATGGTATAATCAAGAACTTAGAGGAAATGGACATACTCTTAAAATA
TGGGCATGTAAACTTTATTATGAAAAGAATGCTAGAAATCAAATAGGATTATGGAATTTAAGAGATAATGGAGTA
GGATTAAATGTAATGGTATCTGAACATTATCAATGTTGTAGAAAAATATTTATACAATCAAGTCATAATCAATTA
AATGAAAATAGATGGCTTGAAAAAACACTTAAAAGAGCTGAAAAGAGAAGAAGTGAATTATCTATAATGATTCAG
GTAAAAATACTTCATACAACAAAATCACCAGCAGTTTCAAGAGGTTCAGGATAA

The linker is in lowercase, the CDS of PmCDA1 is in uppercase.

Linker features: SV40 NLS; GSG linker; SH3 domain; 3xFLAG

### VIII.2.2. Multiplexing

#### *VIII.2.2.1) S. pyogenes CRISPR array*

**Table S. 4: Sequence of the Direct Repeat (DR) and tracrRNA from the CRISPR system of *S. pyogenes*** [391]**; as well as the Hammerhead (HH) and Human Δ-virus (HDV) ribozymes** [371]**.** Lowercase nucleotides represent the part of HH which anneals to the sequence of the tracrRNA.

| Part | Sequence |
|---|---|
| **DR** | GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC |
| **tracrRNA** | GGAACCATTCAAAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGA GTCGGTGCTTTTTTT |
| **HH** | aactggttccCTGATGAGTCCGTGAGGACGAAACGAGTAAGCTCGTCGGAACCATTCAAAACAGC |
| **HDV** | gctagccatggtcccagcctcctcgctggcggctagtgggcaacatgcttcggcatggcgaatgggac |

#### *VIII.2.2.2) tRNA array*

**Table S. 5: List of the two pre-tRNAs used in the sgRNA polycistron of vFS49_TA_mtRNA.** Nomenclature from GtRNAdb**.** Capitalized letters represent the sequence of the final tRNA obtained after post-transcriptional processing.

| Name | Sequence |
|---|---|
| **tRNA-Thr-TGT-1-1** | GCTGGCATAGCTCAATTGGTAGAGCAACTGACTTGTAATCAGTAGGTTATGGGTTCAATTC CTATTGCCAGCACCAattttggaggaattcccgagt |
| **tRNA-fMet-CAT-1-1** | CGCGGGGTGGAGCAGCTGGCAGCTCGTCGGGCTCATAACCCGAAGGTCGTAGGTTCAAGTC CTACCCCCGCAACCAgatttggcggaatagctcagct |

#### *VIII.2.2.1) UGI-LVA*

ATGACAAATTTATCAGATATTATAGAAAAAGAAACAGGAAAACAGTTAGTTATACAAGAATCAATACTTATGTTA
CCAGAAGAAGTTGAAGAAGTAATAGGAAATAAACCTGAAAGTGATATATTAGTACATACTGCTTATGATGAAAGT
ACAGATGAAAATGTAATGTTACTTACTTCTGATGCACCTGAATATAAACCTTGGGCTTTAGTAATACAGGATAGT
AATGGAGAAAATAAAATAAAAATGTTAGGTCTTGTTGCATAA

Feature: GLVA degradation tag.

### VIII.2.3. List of primers and oligos

**Table S. 6: Sequence, annealing template and association of each primer used in this project.** Uppercase letters are used to represent the annealing region of each primer; conversely, when present, lowercase letter represent the overhand of each primer. The purpose of a primer is either to be used for sequencing (Seq) or to assemble a new vector (vFSxx).

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS003** | AAGAGAAACAgTAACTATATATATTCAATTTATG | 62 | vFS02 | oFS004 | vFS45 |
| **oFS004** | GTATGTCGGTACATTTGAAATATTG | 62 | vFS02 | oFS003 | vFS45 |
| **oFS017** | gcttaatctatttggccttagcgcaCATGATTAAGTTTCGTGGTC | 58 | vFS04 | oFS018 | vFS08 |
| **oFS018** | tccacttgcgtttATAGGGTTTTCTTCAAATAATTG | 55.5 | vFS04 | oFS017 | vFS08 |
| **oFS019** | aagaaaaccctatAAACGCAAGTGGAGTAGATG | 62.1 | vFS04 | oFS020 | vFS08 |
| **oFS020** | gaggtgagcataaGTTTTAAGTCTTTCCTCAATCATC | 58.8 | vFS04 | oFS019 | vFS08 |
| **oFS021** | aaagacttaaaacTTATGCTCACCTCTTTGATG | 59.1 | vFS04 | oFS022 | vFS08 |
| **oFS022** | ttacgttgagtgaTTAATTTGGCGTTTAGAAGTTG | 58.4 | vFS04 | oFS021 | vFS08 |
| **oFS023** | aacgccaaattaaTCACTCAACGTAAGTTTG | 56 | vFS04 | oFS024 | vFS08 |
| **oFS024** | acgcatcatgggcatgatggtaattgttTATCTCACGTACTTTATAGAATTG | 56.4 | vFS04 | oFS023 | vFS08 |
| **oFS026** | TGCGCTAAGGCCAAATAGATTAAGC | 67 | vFS04 | oFS027 | vFS08 |
| **oFS027** | AACAATTACCATCATGCCCATG | 63 | vFS04 | oFS026 | vFS08 |
| **oFS028** | tttatctacaatttttttatcaggaaacagctatgaccgcggccgcTTATCCTGAACCTCTTGAAAC | 58.9 | vFS07 | oFS029 | vFS31, vFS72 |
| **oFS029** | gctaggaggtgacAGTAGAGCAGATCCTAAG | 56.2 | vFS07 | oFS028 | vFS31 |
| **oFS030** | ttcttaggatctgctctactGTCACCTCCTAGCTGACTC | 59 | vFS04 | oFS031 | vFS31 |
| **oFS031** | gattaaaatttttaaggaggtgtatttcatATGGATAAGAAATACTCAATAGGC | 59 | vFS04 | oFS030 | vFS31 |
| **oFS034** | cgatggtgtcTCTAGATTTATATTTAGTCCCTTGCCTTGCCTAC | 68 | gFS02 | oFS035 | vFS13 |
| **oFS035** | tttcaatttggcgcgcctaaaagtaaaGCGATCGCATAAAAATAAGAAGCCTGC | 69 | gFS02 | oFS034 | vFS13 |
| **oFS039** | atcttaaggaggagttttcgtcgacAGGGTTGTGGGTTGTACGGAgttttagagctagaaatagcaagtt | 84.9 | vFS17 | oFS079 | vFS21 |
| **oFS040** | atcttaaggaggagttttcgtcgacATTTCTGATATTACTGTCACgttttagagctagaaatagcaagtt | 81.2 | vFS17 | oFS080 | vFS22 |
| **oFS041** | atcttaaggaggagttttcgtcgacACCGATACCGTTTACGAAATgttttagagctagaaatagcaagtt | 83.2 | vFS17 | oFS081 | vFS23 |

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS042** | atcttaaggaggagttttcgtcgacTGAAGATCAGGCTATCACTGgttttagagctagaaatagcaagtt | 83.1 | vFS17 | oFS082 | vFS24 |
| **oFS043** | atcttaaggaggagttttcgtcgacTCCGGAGCTCCGATAAAAAAgttttagagctagaaatagcaagtt | 83.5 | vFS17 | oFS083 | vFS25 |
| **oFS044** | atcttaaggaggagttttcgtcgacTATTGATTCTCTTCAAGTAGgttttagagctagaaatagcaagtt | 80.2 | vFS17 | oFS084 | vFS26 |
| **oFS045** | atcttaaggaggagttttcgtcgacCCATTGTACTATCATGCTAGgttttagagctagaaatagcaagtt | 81.9 | vFS17 | oFS085 | vFS27 |
| **oFS046** | atcttaaggaggagttttcgtcgacATGCAGTCGGCTGTAGAAAGgttttagagctagaaatagcaagtt | 84.4 | vFS17 | oFS086 | vFS28 |
| **oFS047** | atcttaaggaggagttttcgtcgacCGACTGCATTTTATTATGTAgttttagagctagaaatagcaagtt | 81 | vFS17 | oFS087 | vFS29 |
| **oFS053** | GGCCGCGGTCATAGCTG | 67 | vFS05 | oFS054 | vFS30 |
| **oFS054** | ctgttgtttgtcggtgaacgctctcactagtGAGATAGTATATGATGCATATTCTTTAAATATAGATAAAGTTATAGAAGCAATAGAAGATTTAGGATTTACTG | 67 | vFS05 | oFS053 | vFS30 |
| **oFS055** | caatttttttatcaggaaacagctatgaccgcggccgCGACACCATCGAATGGTGC | 67 | vFS06 | oFS056 | vFS30 |
| **oFS056** | tatctcactagtgagagcgttcaccgacaaacaacagataaaacgaaaggcccagtctttcgactgagcctttcgttttatttgatgcctggTCACTGCCCGCTTTCCAG | 68 | vFS06 | oFS055 | vFS30 |
| **oFS057** | GAAACTTAATCATATGCGCTAAGG | 61 | Cas9 | | Seq. |
| **oFS058** | ATGGATAAGAAATACTCAATAGGCTTAG | 61 | Cas9 | | Seq. |
| **oFS059** | GCTTTGTCATTGGGTTTGAC | 62 | Cas9 | | Seq. |
| **oFS060** | GTCGATAAAGGTGCTTCAGC | 63 | Cas9 | | Seq. |
| **oFS061** | GAACATATTGCAAATTTAGCTGG | 58 | Cas9 | | Seq. |
| **oFS062** | CTGACTTCCGAAAAGATTTCC | 61 | Cas9 | | Seq. |
| **oFS063** | GAGTTAGAAAACGGTCGTAAACG | 63 | Cas9 | | Seq. |
| **oFS064** | CGGCGGTGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATaAACTATCCGCTGGATGACCAGGATG | N/A | vFS06 | oFS076 | vFS20 |
| **oFS066** | atacataaatgcgatcgcGTACCGCTGCTATCTGCC | 66 | C. auto | oFS067 | vFS17 |
| **oFS067** | taatatatatggcgcgccCCTGTAATCGGAGCATCTGG | 65 | pyrD | oFS066 | vFS17 |
| **oFS068** | GCAAAATACATTCGTTGATG | 57 | pCB102 | | Seq. |
| **oFS069** | GTCAAGTATGAAATCATAAATAAAG | 55 | pCB102 | | Seq. |
| **oFS070** | GATAAATAGTTAACTTCAGGTTTGTC | 59 | catP | | Seq. |
| **oFS071** | CTGTGGATAACCGTATTACC | 59 | ColE1 | oFS074 | Seq. |
| **oFS072** | CAAGAAGAGCGACTTCGC | 64 | tra | | Seq. |

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS073** | CTAGATTTATATTTAGTCCCTTGCCTTGC | 65 | sgRNA | | Seq. |
| **oFS074** | CTGTTATGCCTTTTGACTATC | 58 | pCB102 | oFS071 | Seq. |
| **oFS075** | GTCAAAATACTCTTTTCTGTTCC | 59 | CatP | oFS071 | Seq. |
| **oFS076** | GCAATGGCATCCTGGTCATCCAGCGGATAGTTTATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCAC CGCCGCTTTAC | N/A | vFS06 | oFS064 | vFS20 |
| **oFS077** | CATTGAAAGAAGTAGGAGCAC | 61 | *pyrE* HA | | Seq. |
| **oFS078** | CTTTTCTACGGGGTCTGAC | 59 | ColE1 | oFS070 | Seq. |
| **oFS079** | aacttgctatttctagctctaaaacTACATAATAAAATGCAGTCGGTCGACgaaaactcctccttaagat | N/A | vFS17 | oFS039 | vFS21 |
| **oFS080** | aacttgctatttctagctctaaaacCTTTCTACAGCCGACTGCATGTCGACgaaaactcctccttaagat | N/A | vFS17 | oFS040 | vFS22 |
| **oFS081** | aacttgctatttctagctctaaaacCTAGCATGATAGTACAATGGGTCGACgaaaactcctccttaagat | N/A | vFS17 | oFS041 | vFS23 |
| **oFS082** | aacttgctatttctagctctaaaacCTACTTGAAGAGAATCAATAGTCGACgaaaactcctccttaagat | N/A | vFS17 | oFS042 | vFS24 |
| **oFS083** | aacttgctatttctagctctaaaacTTTTTTATCGGAGCTCCGGAGTCGACgaaaactcctccttaagat | N/A | vFS17 | oFS043 | vFS25 |
| **oFS084** | aacttgctatttctagctctaaaacCAGTGATAGCCTGATCTTCAGTCGACgaaaactcctccttaagat | N/A | vFS17 | oFS044 | vFS26 |
| **oFS085** | aacttgctatttctagctctaaaacATTTCGTAAACGGTATCGGTGTCGACgaaaactcctccttaagat | N/A | vFS17 | oFS045 | vFS27 |
| **oFS086** | aacttgctatttctagctctaaaacGTGACAGTAATATCAGAAATGTCGACgaaaactcctccttaagat | N/A | vFS17 | oFS046 | vFS28 |
| **oFS087** | aacttgctatttctagctctaaaacTCCGTACAACCCACAACCCTGTCGACgaaaactcctccttaagat | N/A | vFS17 | oFS047 | vFS29 |
| **oFS088** | GGTCATAGCTGTTTCCTG | 60 | vFS09 | | Seq. |
| **oFS091** | gcaATTGTTCCACAAAGTTTCCTTAAAGACG | 65 | vFS19 | oFS092 | vFS39 |
| **oFS092** | ATCGACATCATAATCACTTAAACGATTAATATC | 62 | vFS19 | oFS091 | vFS39 |
| **oFS093** | CTTGTTGTTACCTCCTTAGCAG | 63 | pMTL-IC111-E | oFS094 | vFS36 |
| **oFS094** | GTGTAGTAGCCTGTGAAATAAGTAAG | 62 | pMTL-IC111-E | oFS093 | vFS36 |
| **oFS095** | gcagcaccctgctaaggaggtaacaacaagATGGATAAGAAATACTCAATAGGCTTAG | 61 | vFS19 | oFS096 | vFS36 |
| **oFS096** | ccttacttatttcacaggctactacacTCTAGATTTATATTTAGTCCCTTGCC | 61 | vFS19 | oFS095 | vFS36 |
| **oFS105** | gagcttatgcaattcaagtaggtactgcaaac | 59 | *pyrE* | oFS106 | Seq. |
| **oFS106** | catcaaagctatactattttccgtatttacatttggg | 57 | *pyrE* | oFS105 | Seq. |

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS109** | caattgttcaaaaaaataatggcggcgcgccCCTGTAATCGGAGCATCTGG | 65 | *pyrE* | oFS110 | vFS42, vFS40 |
| **oFS110** | tttttcgacttatttATCCATAACTGTCCTCCTAAATTATTCCTC | 65 | *pyrE* | oFS109 | vFS40 |
| **oFS111** | aggacagttatggatAAATAAGTCGAAAAAATCAATGCACGATGC | 66 | *pyrE* | oFS112 | vFS40 |
| **oFS112** | catttgcaggcttcttatttttatgcgatcgcGTACCGCTGCTATCTGCC | 66 | *pyrE* | oFS111 | vFS42, vFS40 |
| **oFS113** | tatcaggaaacagctatgaccgcggccgctcaGTCACCTCCTAGCTGACTCAAATC | 67 | vFS33 | oFS114 | vFS40 |
| **oFS114** | GCGATCGCATAAAAATAAGAAGCCTG | 66 | vFS33 | oFS113 | vFS40 |
| **oFS115** | ACATATATAAATCTTAAGGAGGAGTTTTCGTCGACaattcaggaattaggtggagGTTTTAGAGCTAGAAATAGCAAGTTAAAAT | N/A | oFS116 | oFS116 | vFS35 |
| **oFS116** | ATTTTAACTTGCTATTTCTAGCTCTAAAACctccacctaattcctgaattGTCGACGAAAACTCCTCCTTAAGATTTATATATGT | N/A | oFS115 | oFS115 | vFS35 |
| **oFS117** | CTTACTTATTTCACAGGCTACTACACTCTAGAcaaggcgtcgacctatagGCGATCGCTTTACTTTTAGGCGCGCCGCC | N/A | oFS118 | oFS118 | vFS37 |
| **oFS118** | GGCGGCGCGCCTAAAAGTAAAGCGATCGCctataggtcgacgccttgTCTAGAGTGTAGTAGCCTGTGAAATAAGTAAG | N/A | oFS117 | oFS117 | vFS37 |
| **oFS119** | ATCCATAACTGTCCTCCTAAATTATTCCTC | 65 | *pyrE* | oFS109 | vFS42 |
| **oFS120** | AAATAAGTCGAAAAAATCAATGCACGATGC | 66 | *pyrE* | oFS112 | vFS42 |
| **oFS121** | tttccacatatataaatcttaaggaggagttttcgtcgacctccagtcaggtgttgtgcagttttagagctagaaatagcAAGTTAAAATAAGGCTAGTCCGTTATC | 62 | vFS36 | oFS122 | vFS49, vFS48, vFS56 |
| **oFS122** | ggaattcctccaaaattggtgctggcaataggaattgaacccataacctactgattacaagtcagttgctctaccaattgagctatgccagcAAAAAAAGCACCGACTCG | 60 | vFS36 | oFS121 | vFS48 |
| **oFS123** | caattcctattgccagcaccaattttggaggaattcccgagtagcccaatgtctagctgggagttttagagctagaaatagcAAGTTAAAATAAGGCTAGTCCGTTATC | 62 | vFS36 | oFS124 | vFS48 |
| **oFS124** | gctattccgccaaatctggttgcgggggtaggacttgaacctacgaccttcgggttatgagcccgacgagctgccagctgctccaccccgcgAAAAAAAGCACCGACTCG | 60 | vFS36 | oFS123 | vFS48 |
| **oFS125** | caagtcctacccccgcaaccagatttggcggaatagctcagctaaacaagcaattgttccgttgttttagagctagaaatagcAAGTTAAAATAAGGCTAGTCCGTTATC | 62 | vFS36 | oFS126 | vFS48 |

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS126** | caaaaaaataatggcggcgcgcctaaaagtaaagcgatcgcataaaaataagaagcctgcaaatgcaggcttct tatttttatAAAAAAAGCACCGACTCG | 60 | vFS36 | oFS125 | vFS48, vFS56 |
| **oFS129** | CCCAAAACtgaaCTCGAGagcccaatgtctagctgggatGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAA CactaACTAGTaaacaag | 81 | oFS130 | oFS130 | vFS51 |
| **oFS130** | cttgtttACTAGTtagtGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAACatcccagctagacattgggct CTCGAGttcaGTTTTGGG | 81 | oFS129 | oFS129 | vFS51 |
| **oFS131** | gcaagacgtcctccagtcaggtgttgtgcagttttagagctatgctgttttgaatggtcCCAAAACTGAACTCG AGAG | 60 | oFS129 | oFS158 | vFS51 |
| **oFS137** | gtataaaaataagaagcctgcaaatgcaggcttcttatttttatAAAAAAAGCACCGACTC | 62 | vFS36 | oFS121 | vFS49 |
| **oFS138** | gaagcctgcatttgcaggcttcttatttttatactagtGACTTTGTTAAAAAAGTTTAATAAATATAATTTGAA TAAATGGTATAAATAGACAGATATTTAG | 65 | vFS47 | oFS139 | vFS49 |
| **oFS139** | ctcccagctagacattgggctgacgtccaaaaatctatatttttttCTAACTGATTTAATTATAACCATAAATTT ACTAGC | 64 | vFS47 | oFS138 | vFS49 |
| **oFS140** | gaaaaaatatagatttttggacgtcagcccaatgtctagctgggagttttagagctagaaatagcAAGTTAAAA TAAGGCTAGTCCGTTATC | 62 | vFS36 | oFS141 | vFS49 |
| **oFS141** | caaacaacagataaaacgaaaggcccagtctttcgactgagcctttcgttttatttgatgcctggAAAAAAAGC ACCGACTCG | 60 | vFS36 | oFS140 | vFS49 |
| **oFS142** | ggctcagtcgaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctcagatctaattaaatttt ttaatACGGTATAGGGGTATTCTTTAGC | 63 | Ppffdx | oFS143 | vFS49 |
| **oFS143** | caacggaacaattgcttgtttctcgagTACCATATTTATATTATCATATTTTTGCTAATTTTTAAAGTATTTAA TATCTC | 62 | Ppffdx | oFS142 | vFS49 |
| **oFS144** | gataatataaatatggtactcgagaaacaagcaattgttccgttgttttagagctagaaatagcAAGTTAAAAT AAGGCTAGTCCGTTATC | 62 | vFS36 | oFS145 | vFS49 |
| **oFS145** | caaaaaaataatggcggcgcgcctaaaagtaaagcgatcgcaaataaaaaacgccccttcgtttacacgaaggg acgattgattatAAAAAAAGCACCGACTC | 60 | vFS36 | oFS144 | vFS49, vFS94 |
| **oFS154** | tttccacatatataaatcttaaggagg | 59 | oFS121-oFS139 | oFS155 | vFS50 |
| **oFS155** | ctcccagctagacattg | 59 | oFS121-oFS139 | oFS154 | vFS50 |
| **oFS158** | ctaaaacaacggaacaattGCTTGTTTACTAGTTAGTGTTTTGGGAC | 65 | oFS130 | oFS131 | Seq. |

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS159** | gatttccacatatataaatcttaaggaggagttttcgtcgacgttttagagctatgctgttttgaatggtcccaaaacgcaagacgtcCTCCAGTCAGGTGTTGTGCAG | 68 | oFS131-oFS158 | oFS160 | vFS51 |
| **oFS160** | agtcagatctataaaaataagaagcctgcaaatgcaggcttcttatttttatgttttgggaccattcaaaacagcatagctCTAAAACAACGGAACAATTGCTTGTTTAC | 65 | oFS131-oFS158 | oFS159 | vFS51 |
| **oFS161** | catttgcaggcttcttatttttatagatctGACTTTGTTAAAAAAGTTTAATAAATATAATTTGAATAAATGGTATAAATAGACAGATATTTAG | 65 | vFS47 | oFS162 | vFS51 |
| **oFS162** | CAAAAATCTATATTTTTTCTAACTGATTTAATTATAACCATAAATTTACTAGC | 64 | vFS47 | oFS161 | vFS51 |
| **oFS163** | aaatatagatttttgggaaccattcaaaacagcATAGCAAGTTAAAATAAGGCTAGTCCG | 64 | vFS36 | oFS164 | vFS51 |
| **oFS164** | gagagcgttcaccgacaaacaacagataaaacgaaaggcccagtctttcgactgagcctttcgttttatttgatgcctggAAAAAAAGCACCGACTCGGTG | 66 | vFS36 | oFS163 | vFS51 |
| **oFS165** | tggttataattaaatcagttagaaaaaatataGATTTTTGGGAACCATTCAAAACAGC | 65 | oFS163-oFS164 | oFS166 | vFS51 |
| **oFS166** | aaaaaaataatggcggcgcgcctaaaagtaaagcgatcgcGAGAGCGTTCACCGACAAAC | 66 | oFS163-oFS164 | oFS165 | vFS51, vFS60 |
| **oFS167** | ATCTTAAGGAGGAGTTTTCGTCGACctccagtcaggtgttgtgcaGTTTTAGAGCTAGAAATAGCAAGTT | N/A | vFS36 | oFS168 | vFS52 |
| **oFS168** | AACTTGCTATTTCTAGCTCTAAAACtgcacaacacctgactggagGTCGACGAAAACTCCTCCTTAAGAT | N/A | vFS36 | oFS167 | vFS52 |
| **oFS169** | ATCTTAAGGAGGAGTTTTCGTCGACagcccaatgtctagctgggaGTTTTAGAGCTAGAAATAGCAAGTT | N/A | vFS36 | oFS170 | vFS53 |
| **oFS170** | AACTTGCTATTTCTAGCTCTAAAACtcccagctagacattgggctGTCGACGAAAACTCCTCCTTAAGAT | N/A | vFS36 | oFS169 | vFS53 |
| **oFS171** | ATCTTAAGGAGGAGTTTTCGTCGACaaacaagcaattgttccgttGTTTTAGAGCTAGAAATAGCAAGTT | N/A | vFS36 | oFS171 | vFS54 |
| **oFS172** | AACTTGCTATTTCTAGCTCTAAAACaacggaacaattgcttgtttGTCGACGAAAACTCCTCCTTAAGAT | N/A | vFS36 | oFS170 | vFS54 |
| **oFS173** | CATACTGGCACAACATTTGC | 62 | CLAU532 | oFS174 | Seq. |
| **oFS174** | TCACATTTTTTTAATGCGACAG | 59 | CLAU532 | oFS173 | Seq. |
| **oFS175** | CCCCAATAGTTAAATTCTAAAAAAGTAATG | 60 | CLAU534 | oFS176 | Seq. |
| **oFS176** | GAGCATACTGCTTTAGTTTG | 58 | CLAU534 | oFS175 | Seq. |
| **oFS177** | GTTTATATTAGGGTTATATTGAGGTC | 58 | CLAU1794 | oFS178 | Seq. |
| **oFS178** | GTTGATTGGTACGAATTTTGC | 60 | CLAU1794 | oFS177 | Seq. |
| **oFS193** | ATCTTAAGGAGGAGTTTTCGTCGACatcacaatgtttagcaggtaGTTTTAGAGCTAGAAATAGCAAGTT | N/A | vFS54 | oFS194 | vFS58 |
| **oFS194** | AACTTGCTATTTCTAGCTCTAAAACtacctgctaaacattgtgatGTCGACGAAAACTCCTCCTTAAGAT | N/A | vFS54 | oFS193 | vFS58 |

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS195** | TAGCAAAAATATGATAATATAAATATGGTACTCGAGatcacaatgtttagcaggtaGTTTTAGAGCTAGAAATAGCAAGTT | N/A | vFS50 | oFS196 | vFS57 |
| **oFS196** | AACTTGCTATTTCTAGCTCTAAAACtacctgctaaacattgtgatCTCGAGTACCATATTTATATTATCATATTTTTGCTA | N/A | vFS50 | oFS195 | vFS57 |
| **oFS197** | ctaaaacTCCCAGCTAGACATTGGG | 64 | vFS48 | oFS121 | vFS56 |
| **oFS198** | cccgagtagcccaatgtctagctgggagttttagagctagaaatagcAAGTTAAAATAAGGCTAGTCCGTTATC | 62 | vFS48 | oFS199 | vFS56 |
| **oFS199** | ctaaaactacctgctaaacattgtgatACTCGGGAATTCCTCC | 60 | vFS48 | oFS198 | vFS56 |
| **oFS200** | cccgagtatcacaatgtttagcaggtagttttagagctagaaatagcAAGTTAAAATAAGGCTAGTCCGTTATC | 62 | vFS36 | oFS126 | vFS56 |
| **oFS201** | gataaatttataaaattcttcttggcttGCTCCCCCATCAATATAACC | 62 | vFS51 | oFS202 | vFS60 |
| **oFS202** | ctcgtttcgtcctcacggactcatcaggttttagagcatatttgtcgacgaaaactCCTCCTTAAGATTTATATATGTGGAAATC | 60 | vFS51 | oFS201 | vFS60 |
| **oFS203** | ctaaaacctgatgagtccgtgaggacgaaacgagtaagctcgtcgttttagagctatgctgttttgaatggtccccaaaacgcaagacgtcCTCCAGTCAGGTGTTGTGC | 66 | vFS51 | oFS204 | vFS60 |
| **oFS204** | gaatggttccgacgagcttactcgtttcgtcctcacggactcatcagggaaccagttCAAAAATCTATATTTTTTCTAACTGATTTAATTATAACCATAAATTTACTAGC | 64 | vFS51 | oFS203 | vFS60 |
| **oFS205** | gaggacgaaacgagtaagctcgtcggaaccattcaaaacagcATAGCAAGTTAAAATAAGGCTAGTCCG | 64 | vFS36 | oFS206 | vFS60 |
| **oFS206** | gatgcctggttattgtcccattcgccatgccgaagcatgttgcccactagccgccagcgaggaggctgggaccatggctagcAAAAAAAGCACCGACTCGG | 63 | vFS36 | oFS205 | vFS60 |
| **oFS207** | catggcgaatgggacaataaCCAGGCATCAAATAAAACGAAAGG | 64 | vFS51 | oFS166 | vFS60 |
| **oFS209** | GGAGATCTTTAGAAGTCCAAG | 59 | CLAU532 | | Seq. |
| **oFS210** | TGTTACTCATGTATTTGATGGTG | 60 | CLAU532 | | Seq. |
| **oFS211** | TGCTGCTAAAGCAATGTG | 60 | CLAU534 | | Seq. |
| **oFS212** | TTTATAGCTCACAATGCCATG | 60 | CLAU534 | | Seq. |
| **oFS213** | ACCTCGAAGCCTTAACTG | 61 | CLAU1794 | | Seq. |
| **oFS214** | CATTAACTCATGCACTAGAAGC | 61 | CLAU1794 | | Seq. |
| **oFS215** | AATCAATGCACGATGCAG | 60 | *pyrE* LHA | | Seq. |
| **oFS216** | AGTCTAAGGATGCAGCAAG | 62 | *pyrE* RHA | | Seq. |

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS232** | cagttagaaaaaatatagatttttggacgtccccaatgtctagctgggagttttagagctagaaatagcAAGTTAAAATAAGGCTAGTCCGTTATC | 62 | vFS57 | oFS233 | vFS94 |
| **oFS233** | cttattttaacttgctatttctagctctaaaactacctgctaaacattgtgCTCGAGTACCATATTTATATTATCATATTTTTGCTAATTTTTAAAGTATTTAATATCTC | 62 | vFS57 | oFS232 | vFS94 |
| **oFS234** | cagagcttttaaatattttgatacaacaattgatcgtaaagtatatagaagtACAAAAGAAGTTTTAGATGCC | 60 | vFS57 | oFS28 | vFS72 |
| **oFS235** | cgatcaattgttgtatcaaaatatttaaaagctctGGGAGCTCCAAGATTCGTCAACGTAAATAAATG | 62 | vFS57 | oFS236 | vFS72 |
| **oFS236** | cggtcgtaaacggatgctggctagtgccagatttTTACAAAAAGGAAATGAGCTGGC | 63 | vFS57 | oFS235 | vFS72 |
| **oFS237** | aaatctggcactagccAGCATCCGTTTACGACCGTTTTC | 65 | vFS57 | oFS238 | vFS72 |
| **oFS238** | caaggagtcaattagaccaaaaagaaattcggacaagcttattgctcgtaaaaaagactgggatccaaaaaaatatggtggttttgtaAGTCCAACGGTAGCTTATTCAG | 64 | vFS57 | oFS237 | vFS72 |
| **oFS239** | cgagcaataagcttgtccgaatttctttttggtctAATTGACTCCTTGGAGAATCCGCCTGTCTGTACTTCTGTTTTCTTG | 63 | vFS57 | oFS240 | vFS72 |
| **oFS240** | acaaattgtgccttctgcctgacgtcCAAAAATCTATATTTTTTCTAACTGATTTAATTATAACCATAAATTTACTAGCTTTC | 66 | vFS57 | oFS239 | vFS72 |
| **oFS241** | gatttttggacgtcaggcagaaggcacaatttgtgttttagagctagaaatagcAAGTTAAAATAAGGCTAGTCCGTTATC | 62 | vFS57 | oFS242 | vFS72 |
| **oFS242** | tattttaacttgctatttctagctctaaaacgaatttgggccagattggatctcgagTACCATATTTATATTATCATATTTTTGCTAATTTTTAAAGTATTTAATATCTC | 62 | vFS57 | oFS241 | vFS72 |
| **oFS245** | gaatatgcatcatatactatctcactagtTTTTTAACAAAATATATTGATAAAAATAATAATAGTGGGTATAATTAAGTTGTTAGAGAAAAC | 65 | pMTL-IC101 | oFS246 | vFS62 |
| **oFS246** | ctctgcgacatcgtataacgttactggtttcacatgAACTAACCTCCTAAATTTTGATACGGG | 64 | pMTL-IC101 | oFS245 | vFS62 |
| **oFS247** | gtaggcaaggcaagggactaaatataaatctagaGAGAGCGTTCACCGACAAACAAC | 65 | vFS35 | oFS248 | vFS62 |
| **oFS248** | GTGAAACCAGTAACGTTATACGATGTC | 65 | vFS35 | oFS247 | vFS62 |
| **oFS249** | gatttccacatatataaatcttaaggaggagttttcgtcgacagacaaaaagctaaatttgtgttttagagctagaaatagcAAGTTAAAATAAGGCTAGTCCGTTATC | 62 | vFS57 | oFS250 | vFS74 |
| **oFS250** | catacctgctaaacattgtgagacgtcCAAAAATCTATATTTTTTCTAACTGATTTAATTATAACCATAAATTTACTAGCTTTC | 62 | vFS57 | oFS249 | vFS74 |
| **oFS251** | gatttttggacgtctcacaatgtttagcaggtatgttttagagctagaaatagcAAGTTAAAATAAGGCTAGTCCGTTATC | 62 | vFS57 | oFS252 | vFS74 |

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS252** | tattttaacttgctatttctagctctaaaactaaaacaggagctgtatggcctcgagTACCATATTTATATTAT CATATTTTTGCTAATTTTTAAAGTATTTAATATCTC | 62 | vFS57 | oFS251 | vFS74 |
| **oFS253** | AGTAGAGCAGATCCTAAGAAGAAAAGAAAAG | 65 | vFS36 | oFS254 | vFS83 |
| **oFS254** | catatataaatcttaaggaggagttttcgtcgacgcaaaatggaaatatgactagttttagagctagaaatagc AAGTTAAAATAAGGCTAGTCCGTTATC | 65 | vFS36 | oFS253 | vFS83 |
| **oFS257** | TTCTGTTTTCTTGACAATATTGACTTGGG | 65 | vFS36 | oFS258 | vFS83 |
| **oFS258** | CGTGAAGATTTGCTGCGC | 65 | vFS36 | oFS257 | vFS83 |
| **oFS259** | caaaggtccgttgcttgcgcagcaaatcttcacgtttTAGTTTCACCAATAATTCCTCAGTACC | 60 | vFS36 | oFS260 | vFS83 |
| **oFS260** | cgattctttctgcacgattgagtaaatcaagaaaaTTAGAAAATCTCATTGCTCAGCTCC | 60 | vFS36 | oFS259 | vFS83 |
| **oFS261** | TCTTGATTTACTCAATCGTGCAGAAAG | 65 | vFS36 | oFS262 | vFS83 |
| **oFS262** | GAAAACTCCTCCTTAAGATTTATATATGTGGAAATC | 65 | vFS36 | oFS261 | vFS83 |
| **oFS278** | ctaataatttaattgtcaattctgcatcgtg | 62 | *pyrE* | oFS105 | Seq. |
| **oFS279** | gcagacaaattagagaatgttgactatg | 63 | *pyrE* | | Seq. |
| **oFS284** | gataatataaatatggtaCTCGAGcacaatgtttagcaggtagttttagagctagaaatagcAAGTTAAAATAA GGCTAGTCCGTTATC | 60 | vFS36 | oFS145 | vFS94 |
| **oFS285** | cttgtaggcaaggcaagggactaaatataaatctagaTTTTTAACAAAATATATTGATAAAAATAATAATAGTG GGTATAATTAAGTTGTTAGAGAAAAC | 65 | vFS42 | oFS286 | vFS96 |
| **oFS286** | CTATAAGATTTTTTTTTGATACTGTGGCGGTCTGTATTTC | 68 | vFS42 | oFS285 | vFS96 |
| **oFS287** | ctaaaattggtttgataaatttataaaattcttcttggctagc | 66 | vFS36 | oFS288 | vFS96 |
| **oFS288** | GAAATACAGACCGCCACAGTATCAAAAAAAAATCTTATAG | 68 | vFS36 | oFS287 | vFS96 |
| **oFS303** | GCCTCTATTATGACTGCCAG | 62 | hprK | | Seq. |
| **oFS304** | CCATGCAGCCTGAACTTAG | 63 | hprK | | Seq. |
| **oFS305** | CGCAATAGTTATTCTTTTAGCATAC | 59 | CLAU3142 | | Seq. |
| **oFS306** | GGTAGTATATATACAATAAAGGCTG | 57 | CLAU3143 | | Seq. |
| **oFS307** | GCAGTATTCTCAACGATTATCTTTTAAATC | 62 | SpoA | | Seq. |
| **oFS308** | CCTTTTATATGTGCTGGAACAC | 61 | SpoA | | Seq. |
| **oFS309** | GCGATTGTACATCCTCTAGC | 62 | CLAU532 | | Seq. |
| **oFS310** | TCTGCACCATCAAATACATGAG | 62 | CLAU532 | | Seq. |

| Label | Sequence | Tm (°C) | Annealing template | Use with | Purpose |
|---|---|---|---|---|---|
| **oFS313** | cactaacaatatactctgagacttatcatc | 62 | *pyrE* | | |
| **oFS314** | gaaataagaggaataatttaggaggac | 60 | *pyrE* | | |

### VIII.2.4. List of vectors

**Table S. 7: Summary of all vectors used in this project. For each vector, its name, culture collection reference number (CC#), usage and source are presented. All vectors use *catP* (chloramphenicol resistance gene), except the ones followed by *(Ampicilin resistance) or ^(Erythromycin).**

| Label | CC# | Usage | Source |
|---|---|---|---|
| **vFS04_ pMTL83151_nCas9_trthl** | 6520 | Amplify spCas9n | Daphne Groothuis |
| **vFS05_ Pfacoid_CatP^** | 6276 | Amplify Pfacoid | Hengzeng |
| **vFS06_FacOID_cas9_placIQ_lacI** | 6521 | Amplify PlacIq-LacI | Ryan Hope |
| **vFS07_ pmCDA1_COOL_Genescript*** | 6522 | Amplify pMCDA1 to make the AID of Target-AID | This study |
| **vFS08_pMTL83151_nCas9_trthl_RSKO** | 6524 | Amplify spCas9n_RSKO | This study |
| **vFS11_pMTL43151_pta_Peter** | 6526 | Host vector for Bookmark complementation of *pyrE* locus (currently deletes pta) | [285] |
| **vFS13_vFS30_vFS45_col3** | 6528 | Host vector for Target-AID protein | This study |
| **vFS17_vFS11_oFS66-oFS67_col3** | 6529 | Host vector for Bookmark complementation of *pyrE* locus (currently cuts *pta* but has *pyrE* complementation LHA/RHA) | This study |
| **vFS20_ LacI_RSKO** | 6532 | Amplifying a *lacI* without *C. autoethanogenum* restriction site to replace LacI in vFS19 | This study |
| **vFS21_pMTL43151_pBM4_col1** | 6533 | Complementing C auto D*pyrE* with *Cas9 by targeting BM4 | This study |
| **vFS22_pMTL43151_pBM5_col2** | 6534 | Complementing C auto D*pyrE* with *Cas9 by targeting BM5 | This study |
| **vFS23_pMTL43151_pBM6_col1** | 6535 | Complementing C auto D*pyrE* with *Cas9 by targeting BM6 | This study |
| **vFS24_pMTL43151_pBM7_col1** | 6536 | Complementing C auto D*pyrE* with *Cas9 by targeting BM7 | This study |
| **vFS25_pMTL43151_pBM8_col3** | 6537 | Complementing C auto D*pyrE* with *Cas9 by targeting BM8 | This study |
| **vFS26_pMTL43151_pBM9_col1** | 6538 | Complementing C auto D*pyrE* with *Cas9 by targeting BM9 | This study |
| **vFS27_pMTL43151_BM10_col1** | 6539 | Complementing C auto D*pyrE* with *Cas9 by targeting BM10 | This study |

| Label | CC# | Usage | Source |
|---|---|---|---|
| **vFS28_pMTL43151_BM11_col3** | 6540 | Complementing C auto D*pyrE* with *Cas9 by targeting BM11 | This study |
| **vFS29__pMTL43151_BM12_col4** | 6541 | Complementing C auto D*pyrE* with *Cas9 by targeting BM12 | This study |
| **vFS30_ Pfacoid_LacI^** | 6542 | Promoter assay and template for Target-AID assembly | This study |
| **vFS31_vFS13_vFS08_vFS07_col13** | 6568 | Final Target-AID construct targeting *bdh* in *C. autoethanogenum* | This study |
| **vFS32_ pMTL-IC111-E** | 6569 | Amplify rb3 repression system to replace LacIq in Target-AID. | [245] |
| **vFS35_TA_Placiq_LacI_*pyrE*** | 6574 | Test Pfacoid-LacI repression system with Target-AID. | This study |
| **vFS36_ TA_$P_{fdxE}$_*pyrE*** | 6575 | Delete *pyrE* using Target-AID | This study |
| **vFS37_ NosgRNA** | 6676 | Control for sgRNA toxicity/off-target mustagenesis | This study |
| **vFS38_pMTL83151_*pyrE*_ACE** | 6682 | Replace homology arms for Bookmark array integration | [169] |
| **vFS39_dTA** | 6683 | Control for nCas9 toxicity | This study |
| **vFS40_ nCas9_$P_{fdxE}$_HDR** | 6684 | Control for Target-AID efficiency | This study |
| **vFS42_ Cas9_*pyrE*_BM4-12** | 6686 | Replace *pyrE* with Bookmark array in *C. autoethanogenum* | This study |
| **vFS44_pMTL83151** | 1231 | Cut out Cloning site to reconstitute a RSKO pMTL83151. Conjugation control. | [168] |
| **vFS45_pMTL83151_RSKO** | 7545 | RSKO backbone. | This study |
| **vFS46_pMTL8225x_TT_Cpf_fdx_catP_CACthlRBS^** | 4077 | Amplify Ppffdx with oFS142-143 | [246] |
| **vFS47_pMTL8225x_TT_Cpf_thl_CACthlRBS_catP^** | 4097 | Amplify Ppfthl with oFS138-139 | [246] |
| **vFS48_ TA_mtRNA** | 7539 | Knock out CLAU532,534,1794 in one go; test tRNA as a multiplexing tool | This study |
| **vFS49_Truncated_ TA_msgRNA** | | Knock out CLAU532,534,1794 in one go; Truncated, do not use | This study |
| **vFS50_TA_msgRNA** | 7540 | Knock out CLAU532,534,1794 in one go; | This study |
| **vFS51_TA_mCRISPR** | 7541 | Knock out CLAU532,534,1794 in one go; test CRISPR DR and traRNA as a multiplexing tool | This study |
| **vFS52_ TA_CA532A** | 7542 | Knock out CLAU532; control with single target for multiplexing experiment. | This study |
| **vFS53_ TA_CA534A** | 7543 | Knock out CLAU534; control with single target for multiplexing experiment. | This study |
| **vFS54_ TA_CA1794A** | 7544 | Knock out CLAU1794; control with single target for multiplexing experiment. | This study |

| Label | CC# | Usage | Source |
|---|---|---|---|
| **vFS56_mTHRtRNA** | 7547 | Troubleshoot vFS48 with alternative CLAU1794sgRNA and same tRNA twice in the array | This study |
| **vFS57_ TA_msgRNA_CA1794B** | 7548 | Troubleshoot vFS50 with alternative CLAU1794sgRNA | This study |
| **vFS58_TA_CA1794B** | 7549 | Knock out CLAU1794; control with single target for multiplexing experiment. | This study |
| **vFS60_TA_mCRISPR_HHDV** | 7551 | Troubleshoot vFS51 with alternative CLAU1794sgRNA and traRNA flanked by Hammerhead and HDV ribozymes | This study |
| **vFS61_ pMTL-IC101** | 7759 | Clone Pacthl into my new TA-LacI construct | [245] |
| **vFS62_TA_Pth_LacI_*pyrE*** | 7760 | Lac inducible target-AID, troubleshoot vFS35. | This study |
| **vFS66_truncated_TA-msgRNA_short** | | Troubleshoot vFS57 with shorter sgRNAs to increase selectivity; Do not use, truncated scaffold. | |
| **vFS70_nCas9_ RSKO** | 7761 | See if removing multiple native methylation sites increases conjugation efficiency | This study |
| **vFS71_nCas9_ trthl** | 7762 | Control for test of removing methylation sites on conjugation efficiency. | This study |
| **vFS72_mTA-NG** | 7763 | Troubleshoot vFS57 with broader PAM recognition | This study |
| **vFS74_ mTA-NG_CA1794** | 7764 | Test three additional gRNAs for CLAU1794 in case the one in vFS72 fails. | This study |
| **vFS75_mTA-UGILVA** | 7765 | Troubleshoot vFS57 with UGI | This study |
| **vFS83_ iSpymac_*pyrE*** | 7766 | Test if Spymac can be used to target C. auto. | This study |
| **vFS94_mTA_trsgRNA** | 7767 | Test if using shorter sgRNAs can shift editing window | This study |
| **vFS96_ trTA_*pyrE*** | 7768 | See if I can get rid of theophylline | This study |
| **vFS100_TA-UGI-NoL** | 7779 | See if UGI improves TA  without LVA tag | This study |
| **vFS101_dTA-UGI-NoL** | 7770 | See if UGI improves dTA  without LVA tag | This study |

### VIII.2.5. List of all gblocks

**Table S. 8: Sequence of all gblocks used in this project.**

| Label | Sequence |
|---|---|
| gFS04 _BM4-12 | GCATCGTGCATTGATTTTTTCGACTTATTTAGGGTTGTGGGTTGTACGGAAGGATTTCTGATATTACTGTCACAGGACCGATACCGTTTA<br>CGAAATAGGTGAAGATCAGGCTATCACTGAGGTCCGGAGCTCCGATAAAAAATGGTATTGATTCTCTTCAAGTAGAGGCCATTGTACTAT<br>CATGCTAGAGGATGCAGTCGGCTGTAGAAAGAGGCGACTGCATTTTATTATGTAAGGATCCATAACTGTCCTCCTAAATTATTCCTC |
| gFS05_UGI-GLVA_genescript | ATGACAAATTTATCAGATATTATAGAAAAAGAAACAGGAAAACAGTTAGTTATACAAGAATCAATACTTATGTTACCAGAAGAAGTTGAA<br>GAAGTAATAGGAAATAAACCTGAAAGTGATATATTAGTACATACTGCTTATGATGAAAGTACAGATGAAAATGTAATGTTACTTACTTCT<br>GATGCACCTGAATATAAACCTTGGGCTTTAGTAATACAGGATAGTAATGGAGAAAATAAAATAAAAATGTTAGGTCTTGTTGCATAA |
| gFS06_Spymac-Pam_genescript | ATTCAGACTGTTGGTCAGAATGGTGGTCTTTTTGATGATAATCCTAAATCTCCTCTTGAAGTAACTCCTAGTAAACTTGTACCTCTTAAA<br>AAAGAACTTAATCCTAAAAAATATGGAGGATATCAAAAACCTACTACAGCATATCCAGTATTACTTATAACTGATACAAAACAATTAATA<br>CCTATATCTGTAATGAATAAAAAACAATTTGAACAAAATCCAGTAAAATTTTTAAGAGATAGAGGATATCAACAAGTAGGTAAAAATGAT<br>TTTATAAAACTTCCAAAATATACTTTAGTAGATATAGGAGATGGAATAAAAAGACTTTGGGCAAGTTCTAAAGAAATACATAAAGGAAAT<br>CAATTAGTAGTAAGTAAAAAATCTCAAATACTTCTTTATCATGCTCATCATTTAGATTCAGATTTAAGTAATGATTATTTACAAAATCAT<br>AATCAACAATTTGATGTACTTTTTAATGAAATAATATCATTTAGTAAAAAATGTAAATTAGGAAAAGAACATATACAAAAAATAGAAAAT<br>GTATATTCTAATAAGAAAAATTCTGCATCAATAGAAGAATTAGCTGAATCATTTATAAAACTTCTTGGATTTACTCAATTAGGAGCTACA<br>AGTCCTTTTAATTTTCTTGGAGTAAAACTTAATCAAAAACAATATAAAGGAAAGAAAGATTATATACTTCCATGTACTGAAGGAACATTA<br>ATAAGACAGAGTATTACAGGACTTTATGAAACAAGAGTAGATTTATCAAAAATTGGAGAAGAT |

## VIII.2.6. Standard parts library

### *VIII.2.6.1) Promoter library (Pander,2017)*

Copied from [246].

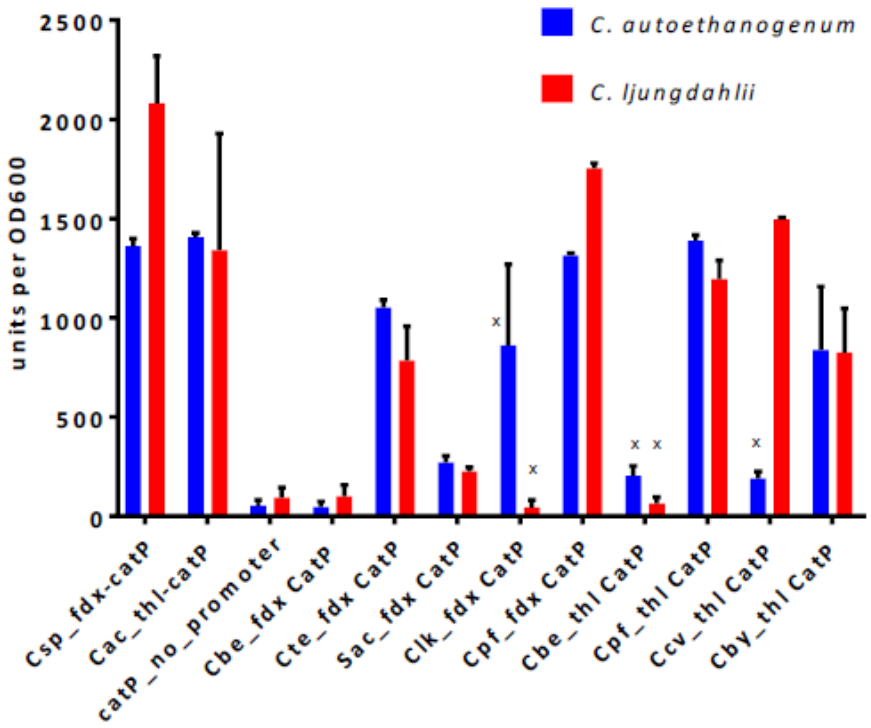



**Figure 8-1 The units of CatP activity per OD equivalent of promoter-catP constructs of *C.ljungdahlii* and *C.autoethanogenum*.**

For every construct there are three replicates per species. Error bars represent standard error of the mean. *Csp_fdx* & *Cac_thl*, *Cpf_fdx* and *Cpf_thl* apparently are of equal high strength, *Cby_thl* and *Cte_fdx* are somewhat weaker. The weakest promoter with significant results for both species was *Pfdx* of *C. saccharoperbutylacetonicum*. The x indicates that sequencing revealed mutations in essential parts of the plasmid, mostly caused by insertion elements.

#### VIII.2.6.1)(1) original promoter sequences

##### (a) fdx promoters

###### *(i) Sporogenes fdx*

1. **Csp_fdx**

gtgtagtagcctgtgaaataagtaaggaaaaaaaagaagtaagtgttatatatgat
gattattttgtagatgtagataggataatagaatccatagaaaatataggttatac
agttatataaaaattactttaaaaattaataaaaacatggtaaaatataaatcgta
taaagttgtgtaatttttaaggaggtgtgttacat

###### *(ii) Beijerinckii fdx (GeneID:5291299)*

2. **Cbe_fdx**

Aagaaaaattcttttccttatactatatttatttcatcatttaaattatattatt
aaatatacataacttaatatttcttgggaaataataagcttgtaatttgatattag
aggagatgaagtaaacc

###### *(iii) Saccharoperbutylacetonicum fdx (GeneID:14618708)*

3. **Sac_fdx**

aaatatagtggttttataattgaaaattatataattatattaatacataacttaat
atttcttaggaaa taataagttatgtaattttatattgaggagatgaattaaatt

*(iv) Clostridium perfringens "GeneID:990809"*

**4. Cpf_fdx**

AATTAAATTTTTAATACGGTATAGGGGTATTCTTTAGCATGTTAATTCTAATTTTA
CTAGAATAGGCTAAATATGCTTAAAAGAGATATTAAATACTTTAAAAATTAGCAAA
AATATGATAATATAAATATGGTATTTTTATACCAACATTTTAGGAGGTGTTTTATT

*(v) Clostridium tetani GeneID1060240*

**5. Cte_fdx**

TATTAAAACTGTTGAGAATTAATTTTCTCAGCAGTTTTTTATTTTGTACGTAACAT
TCGTTACAGACCTATACCTTCTATTTTAATATAATGTACTTATAAAATCAATAAAT
AGCAGTGAATAATTAGGAGGATAGATTT

*(vi) Kluyveri fdx (YP_001397146.1)*

**6. Clk_fdx**

AAAATTATCTTTAAAAATTATAATAAATATGATAAAATGTTATAGGATATAAAAAC
ATATTGATTCTTAAAATCTA**AGGAGG**TGTAGTAT

(b) thl_promoters

*(i) Beijerinckii_thl (GeneID:5294796)*

**7. Cbe_thl**

AGTTTGTTTAAATTTTAACAATTTATGCTTGATAAAAGAAATAATAAAAGGTATAA
TTTAGTTATGTTAATAATTTAACAAAAGTTAATAAAATAATCTATAAAATTTTAGG
AGGTCAAAC

*(ii) Cellulovorans thl 2 (Clocel_3058)*

**8. Ccv_thl**

AATATTTTATTGGATAAAGAAAATAATACATAAATTAGTTAAAATTTTAACAATTT
ATATTGATAAAAAAGCTTTAAATAGTTATAATAAATATGTTGACAAATGTAAATTC
GTTTTAAAGCAATAAGCTTAAATAAATTTAGGAGGTAATTCC

*(iii) Kluyveri thlA3 (CKL_3698)*

**9. Clk_thl**

GTATACTATGTTTTTAGGCTAGTAAGTTAATTATAGAAGTGTAAAAAACCTTTGTG
TTTTTTTTAACAACAACATTGACAACGTTATGATTACTTAGTATAATTAGCTTGTC
ATCAGGATATTAAACATCGGATAGGTTAAATTTTGGTTAAAAATAAAATAAACTTT
TAGGAGGATTTACT

*(iv) Butyricum thl (CBY_1290)*

**10. Cby_thl**

Aattatatactttgataaaaaataaaactttgtctaaaatttaacaataaactat
tgataaaaaagttaaaaatggtataattatcttatgttaataattaacgaaaagt
aaataaaaaacataattataatttaaaataatttatgatattttaggaggtaaa
tt

*(v) Perfringens thl (GeneId: 990519)*

**11. Cpf_thl**

GACTTTGTTAAAAAAGTTTAATAAATATAATTTGAATAAATGGTATAAATAGACAG
ATATTTAGAATATTATAGAAATTTTAATAAAAGACTTCTATAATAAAGCTAAATTA
TCTGTCTTTTTTTTTCGAAAAGAGAAAAAATAATAAAAAAGATTGTTTAAAATTTAA
CAAAAAATATTGAAAGCTAGTAAATTTATGGTTATAATTAAATCAGTTAGAAAAAA
TATAGATTTTTGATAAATTTTTACCTAAATGGGAGGTTTTTGAT

Promoter prediction software

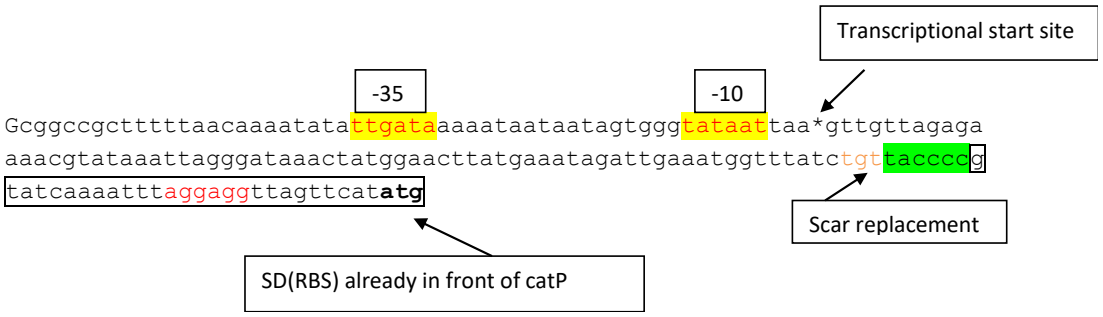
http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb

Terminator prediction software

http://rna.igmors.u-psud.fr/toolbox/arnold/index.php#Results

### VIII.2.6.1)(2) Example for BB2 assembly

**Pthl *Clostridium acetobutylicum***




**Pthl_BB2_noRBS**

BB2 prefix

GAATTCGCGGCCGCACTAGTttttaacaaaatatattgataaaaataataatagtgggtataattaa
gttgttagagaaaacgtataaattagggataaactatggaacttatgaaatagattgaaatggtttat
ctgtGCTAGCGCGGCCGCTGCAG

BB2 suffix

**Cbe_fdx**

Aagaaaaattcttttccttatactatatttatttcatcattttaaattatattatt
aaatatacataacttaatatttcttgggaaataataagcttgtaatttgatattag
aggagatgaagtaaacc

### VIII.2.6.1)(3) New promoters in bb2 format

RBS (shine dalgarno) has been removed.

#### (a) fdx promoters

1. **Cbe_fdx_BB2 (131)**

GAATTCGCGGCCGCACTAGTaagaaaaattcttttccttatactatatttatttca
tcattttaaattatattattaaatatacataacttaatatttcttgggaaataata
GCTAGCGCGGCCGCTGCAG

2. **Sac_fdx_BB2**

GAATTCGCGGCCGCACTAGTaaatatagtggtttataattgaaaattatataatt
atattaatacataacttaatatttcttaggaaataataagGCTAGCGCGGCCGCTG
CAG

3. **Cpf_fdx_BB2**

GAATTCGCGGCCGCACTAGTAATTAAATTTTTAATACGGTATAGGGGTATTCTTTA
GCATGTTAATTCTAATTTTACTAGAATAGGCTAAATATGCTTAAAAGAGATATTAA
ATACTTTAAAAATTAGCAAAAATATGATAATATAAATATGGTAGCTAGCGCGGCCG
CTGCAG

4. **Cte_fdx_BB2**

GAATTCGCGGCCGCACTAGTTATTAAAACTGTTGAGAATTAATTTTCTCAGCAGTT
TTTTATTTTGTACGTAACATTCGTTACAGACCTATACCTTCTATTTTAATATAATG
TACTTATAAAATCAATGCTAGCGCGGCCGCTGCAG

5. **Clk_fdx_BB2**

GAATTCGCGGCCGCACTAGTAAAATTATCTTTAAAAATTATAATAAATATGATAAA
ATGTTATAGGATATAAAAACATGCTAGCGCGGCCGCTGCAG

(b) thl promoters

6. **Cbe_thl_BB2**

GAATTCGCGGCCGCACTAGTAGTTTGTTTAAATTTTAACAATTTATGCTTGATAAA
AGAAATAATAAAAGGTATAATTTAGTTATGTTAATAATTTAACAAAAGTTAATAAG
CTAGCGCGGCCGCTGCAG

7. **Ccv_thl_BB2**

GAATTCGCGGCCGCACTAGTAATATTTTATTGGATAAAGAAAATAATACATAAATT
AGTTAAAATTTTAACAATTTATATTGATAAAAAAGCTTTAAATAGTTATAATAAAT
ATGTTGACAAATGTAAATTCGTTTTAAAGCAGCTAGCGCGGCCGCTGCAG

8. **Clk_thl_BB2**

GAATTCGCGGCCGCACTAGTGTATACTATGTTTTTAGGCTAGTAAGTTAATTATAG
AAGTGTAAAAAACCTTTGTGTTTTTTTTAACAACAACATTGACAACGTTATGATTA
CTTAGTATAATTAGCTTGTCATCAGGATATTAAACATCGGATAGGTTAAATTTTGG
TTAGCTAGCGCGGCCGCTGCAG

9. **Cby_thl_BB2**

GAATTCGCGGCCGCACTAGTAattatatactttgataaaaaataaaactttgtct
aaaatttaacaataaactattgataaaaaagttaaaaatggtataattatcttatg
ttaataattaacgaaaagtaaataaaaaacataattataatttaaGCTAGCGCG
GCCGCTGCAG

10. **Cpf_thl_BB2**

GAATTCGCGGCCGCACTAGTGACTTTGTTAAAAAAGTTTAATAAATATAATTTGAA
TAAATGGTATAAATAGACAGATATTTAGAATATTATAGAAATTTTAATAAAAGACT
TCTATAATAAAGCTAAATTATCTGTCTTTTTTTTCGAAAAGAGAAAAAATAATAAA
AAAGATTGTTTAAAATTTAACAAAAAATATTGAAAGCTAGTAAATTTATGGTTATA
ATTAAATCAGTTAGAAAAAATATAGATTTTTGGCTAGCGCGGCCGCTGCAG

### *VIII.2.6.2) RBS library (Rowe, 2017)*

Copied from [285].

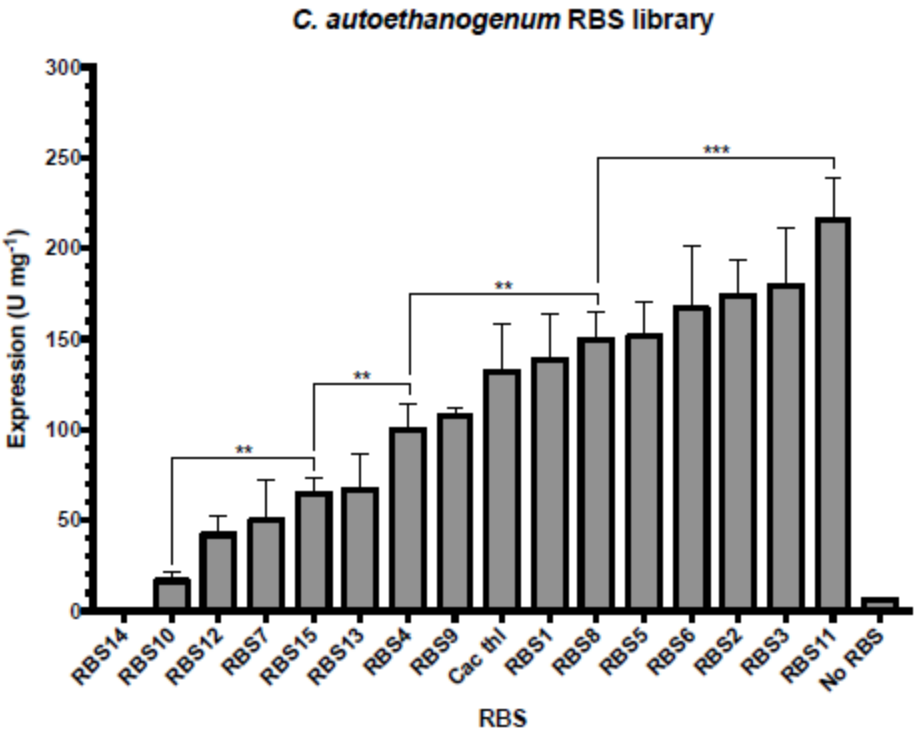



**Figure 3.10:** ***C. autoethanogenum*** **RBS library analysis**
Expression of CatP from vector library consisting of pMTL82254 backbone with C. acetobutylicum Pthl driving catP and the original Pthl RBS substituted for the RBS sequences detailed in Table 3.3. Expression is given as CatP units per mg of total protein. Statistical significance of ≤ 0.01 represented by two asterisks and ≤ 0.001 represented by three asterisks.

**Table 3.3: RBS sequences for study in *C. autoethanogenum***
RBS sequences for analysis in *C. autoethanogenum*, start codons shaded in grey. RBS identifiers are provided, along with the *C. autoethanogenum* gene with which they are associated, with "Cac thl" derived from the *C. acetobutylicum* $P_{thl}$ RBS present of the pMTL82254 vector.

| RBS | Associated gene | Sequence |
|---|---|---|
| 1 | CLAU 2296 | CAAACTACAAAAAGGAGGTGTTCATATG |
| 2 | *rplO* | TAATACAAAGGAGGTGTAAGTACATATG |
| 3 | *rplK* | TAACCACATATAGGAGGTGTATCATATG |
| 4 | *rhaB* | GTTCTAAAAAGAGGTGTTAAATCATATG |
| 5 | CLAU 2414 | TTTTATCATACAAGGAGGTTTACATATG |
| 6 | CLAU 3014 | AGAAAGGAGTTGCCTATATATTCATATG |
| 7 | CLAU 0011 | CATGATAAGGAGGTAATTTAAACATATG |
| 8 | *fabH* | AAAATTGGATCAAAGGAGAAGACATATG |
| 9 | CLAU 1812 | TATGAGGAGTTGATTAATACAACATATG |
| 10 | CLAU 2282 | TAAGATGTGATTAAGGAAGGTTCATATG |
| 11 | *pyrE* | AGAGGAATAATTTAGGAGGACACATATG |
| 12 | CLAU 1740 | AAAATAATGGGGAGGGTTAAAACATATG |
| 13 | *pth* | ATAGTATTGAGATTGGAGTGAACATATG |
| 14 | CLAU 0410 | ATGCTCAGATTCCGGGATATAACATATG |
| 15 | CLAU 0818 | TATATTTTACAATTCCTGTAAATCATTG |
| Cac thl | pMTL82254 | ATCAAAATTTAGGAGGTTAGTTCCATTG |

### *VIII.2.6.3) Terminators library (Willson, 2016)*

Copied from [244].

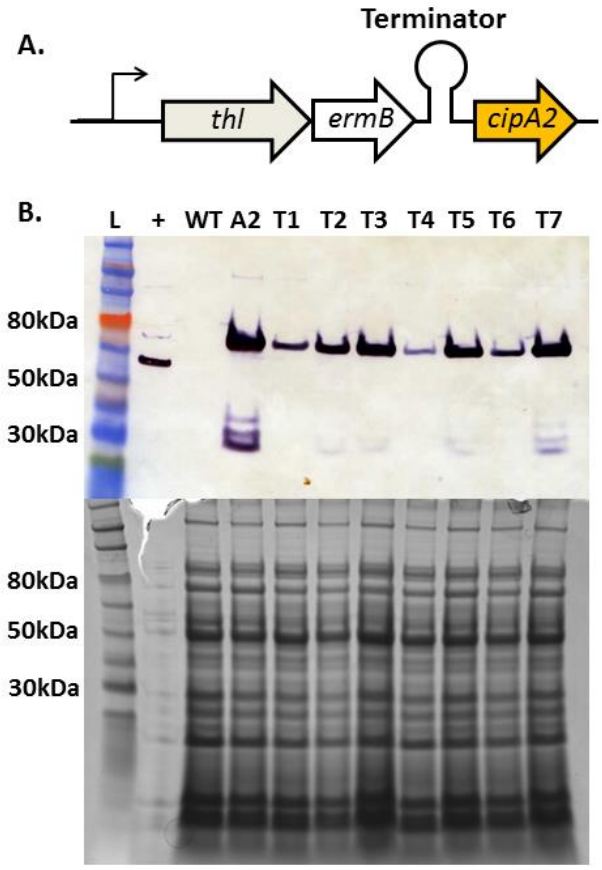



**Figure S. 1: Western analysis of different Rho-independent terminators in *C. acetobutylicum*. Section A: Diagram showing the gene arrangement at the thiolase locus after integration. Section B: Western blot (above) and Coomassie-stained gel (below) of TCA-precipitated supernatants from wild-type *C. acetobutylicum* ATCC 824 (WT) and from strains integrating CipA2 with no terminator (A2) or with the L. lactis pepN terminator (T1), *L. acidophilus slpA* terminator (T2), *C. difficile slpA* terminator (T3), *E. coli rrnB* terminator T1 loop (T4), *B. subtilis* Φ29 phage late TD1 terminator (T5), *B. subtilis tyrS* tRNA terminator (T6), or *B. subtilis gyrA* terminator (T7). Copied from** [244]**.**

- T1: TpepN (*L. lactis pepN* terminator)

TAATTTATAAATAAAAATCACCTTTTAGAGGTGGTTTTTTTATTTATAAATTA

- T2: TslpA_LA (*L. acidophilus slpA* terminator)

TGAAAAAGGCAGAGCGAAAGCTCTGTCTTTTTT

- T3: TslpA_CD (*C. difficile slpA* terminator)

AAATATAAAAAGACTTCTCAGATGAGAAGTCTTTTTTGTGAAA

- T4: EcoT1 (*E. coli rrnB* terminator T1 loop)

CCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGT
TTGTCGGTGAACGCTCTC

- T5: phiTD1 (*B. subtilis* phage Φ29 late TD1 terminator)

AACAATCAAAAGAAAAGCCTATCGTCTGAGGAACGGTAGGCTCTTTTGTAGCATATAGTTG

- T5: TtyrS (*B. subtilis tyrS* tRNA terminator)

ATAATCAATCGTCCCTTCGTGTAAACGAAGGGGCGTTTTTTATTT

- T7: TgyrA (*B. subtilis* gyrA terminator)

AAGAAGAAGTGTGAAAAAGCGCAGCTGAAATAGCTGCGCTTTTTTGTGTCATAA

## VIII.3. Supplementary data

### VIII.3.1. Bookmark

#### *VIII.3.1.1) Raw data*

**Table S. 9: Raw data from the calculation of the complementation efficiency of each bookmark protospacer, across three independent conjugations.**

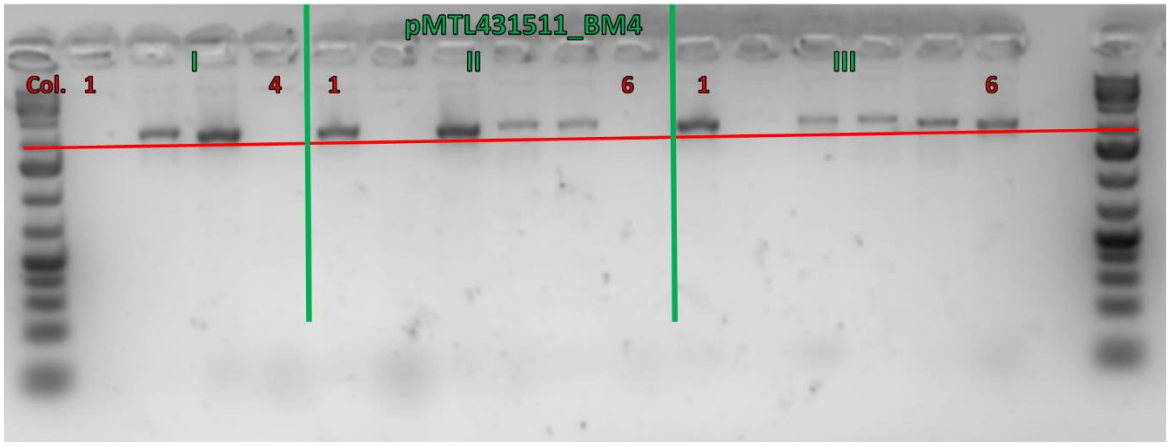
| Bookmark protospacer | Replicate | Colony count (CFU) | Complemented | Screened | Efficiency |
|---|---|---|---|---|---|
| **BM4** | I | 1132 | 6 | 8 | 75% |
| **BM4** | II | 750 | 8 | 8 | 100% |
| **BM4** | III | 1154 | 9 | 9 | 100% |
| **BM5** | I | 1616 | 8 | 9 | 89% |
| **BM5** | II | 1475 | 7 | 7 | 100% |
| **BM5** | III | 536 | 11 | 11 | 100% |
| **BM6** | I | 414 | 12 | 12 | 100% |
| **BM6** | II | 946 | 6 | 7 | 86% |
| **BM6** | III | 113 | 7 | 7 | 100% |
| **BM7** | I | 849 | 3 | 5 | 60% |
| **BM7** | II | 615 | 3 | 3 | 100% |
| **BM7** | III | 971 | 8 | 9 | 89% |
| **BM8** | I | 1178 | 3 | 3 | 100% |
| **BM8** | II | 308 | 8 | 9 | 89% |
| **BM8** | III | 433 | 4 | 4 | 100% |
| **BM9** | I | 412 | 7 | 7 | 100% |
| **BM9** | II | 690 | 7 | 7 | 100% |
| **BM9** | III | 481 | 5 | 5 | 100% |
| **BM10** | I | 410 | 5 | 6 | 83% |
| **BM10** | II | 308 | 2 | 2 | 100% |
| **BM10** | III | 436 | 6 | 7 | 86% |
| **BM11** | I | 183 | 5 | 5 | 100% |
| **BM11** | II | 156 | 7 | 8 | 88% |
| **BM11** | III | 259 | 9 | 11 | 82% |
| **BM12** | I | 235 | 2 | 6 | 33% |
| **BM12** | II | 280 | 4 | 4 | 100% |
| **BM12** | III | 568 | 4 | 4 | 100% |
| **WT** | I | 0 | N/A | N/A | N/A |
| **WT** | II | 0 | N/A | N/A | N/A |
| **WT** | III | 0 | N/A | N/A | N/A |

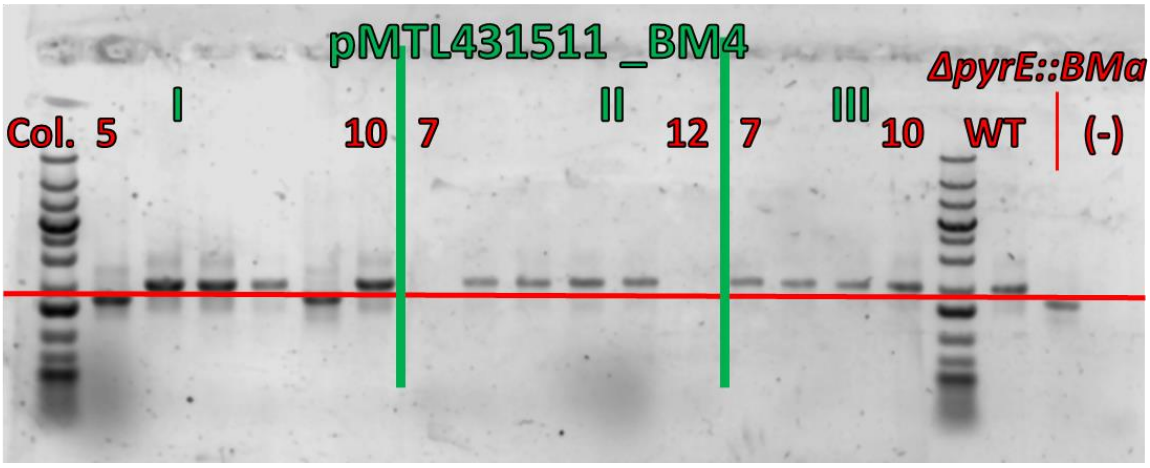
### *VIII.3.1.2) Complementation efficiency gels*

For each gel, a red line has been drawn in between the 2 kb bands of the DNA ladder. The strains complemented to the WT genotype should give a 2 kb amplicon while the Δ*pyrE*::*BM4-12* strain should give a 1.7 kb amplicon. The different replicates are separated by green vertical lines and labelled with roman numerals (I for first replicate, II for second replicate and III for third replicate). Many colonies gave no amplicons at all, which is a limitation of my colony PCR protocol. These colonies were excluded from the calculation of the complementation efficiency.

#### VIII.3.1.2)(1) pMTL83151_BM4

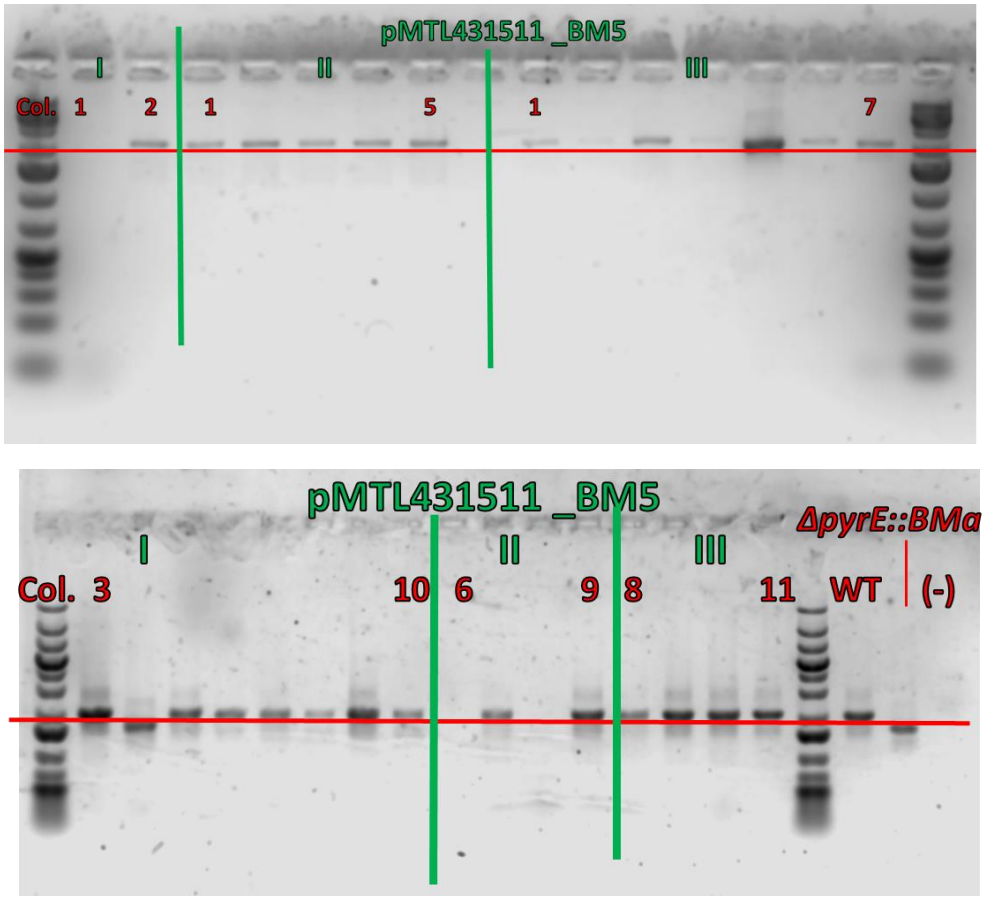
**Figure S. 2: Electrophoresis gel of *C. autoethanogenum ΔpyrE::BM4-12* colonies obtained after conjugation of pMTL431511_BM4.** The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% or 2% (w/v) agarose gel alongside Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific. The expected size of the amplicon of a successfully complemented *pyrE* locus is 2 kb, versus 1.7 kb for the *ΔpyrE::BM4-12* background. The different replicates are separated by green vertical lines and labelled with roman numerals. I = first replicate; II = second replicate; III = third replicate; WT= wild-type *pyrE* locus (2 kb); (-) = negative control of colony PCR without DNA template; Col.: colony.



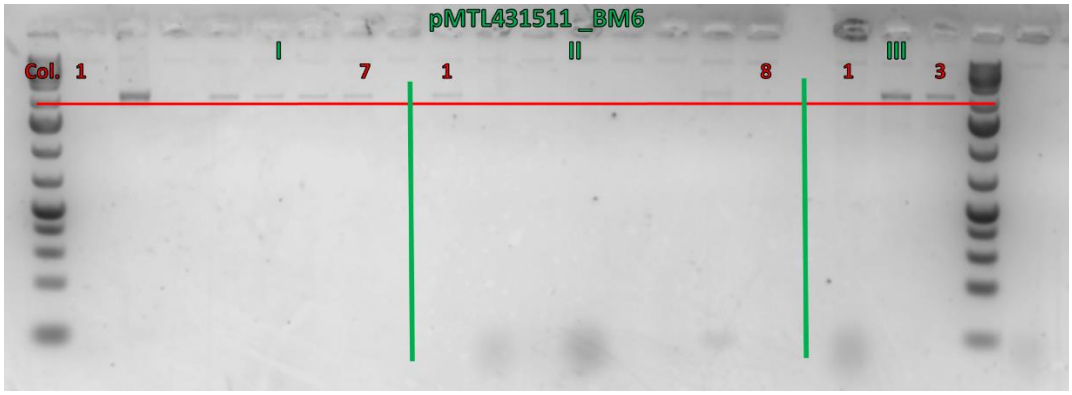
VIII.3.1.2)(2) pMTL83151_BM5

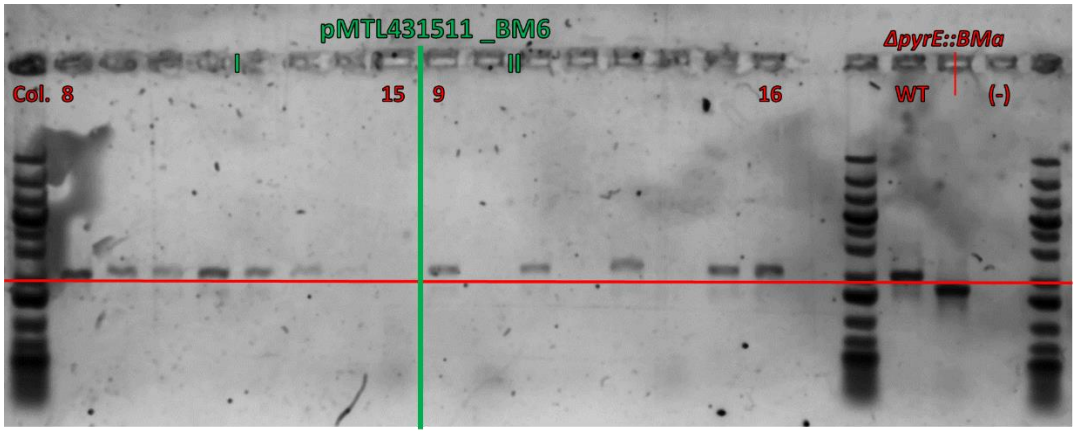
**Figure S. 3: Electrophoresis gel of *C. autoethanogenum ΔpyrE::BM4-12* colonies obtained after conjugation of pMTL431511_BM5.** The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% or 2% (w/v) agarose gel alongside Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific. The expected size of the amplicon of a successfully complemented *pyrE* locus is 2 kb, versus 1.7 kb for the *ΔpyrE::BM4-12* background. The different replicates are separated by green vertical lines and labelled with roman numerals. I = first replicate; II = second replicate; III = third replicate; WT= wild-type *pyrE* locus (2 kb); (-) = negative control of colony PCR without DNA template; Col.: colony.



### VIII.3.1.2)(3) pMTL83151_BM6

**Figure S. 4: Electrophoresis gel of *C. autoethanogenum* ⊿*pyrE::BM4-12* colonies obtained after conjugation of pMTL431511_BM6.** The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% or 2% (w/v) agarose gel alongside Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific. The expected size of the amplicon of a successfully complemented *pyrE* locus is 2 kb, versus 1.7 kb for the *⊿pyrE::BM4-12* background. The different replicates are separated by green vertical lines and labelled with roman numerals. I = first replicate; II = second replicate; III = third replicate; WT= wild-type *pyrE* locus (2kb); (-) = negative control of colony PCR without DNA template; Col.: colony.



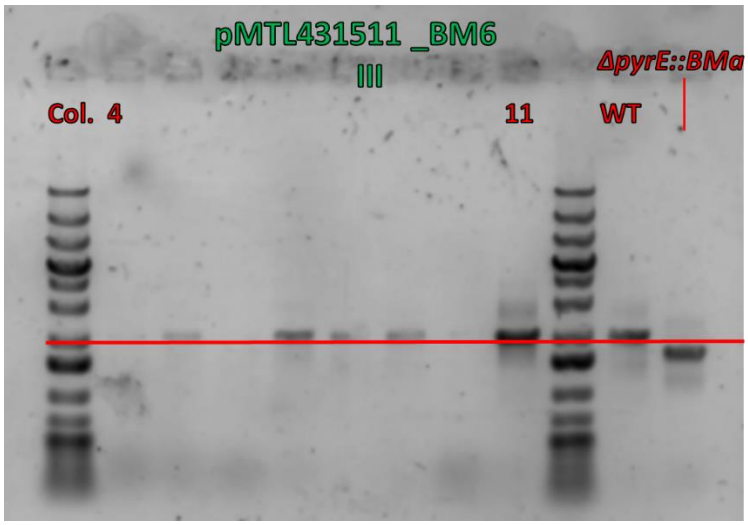
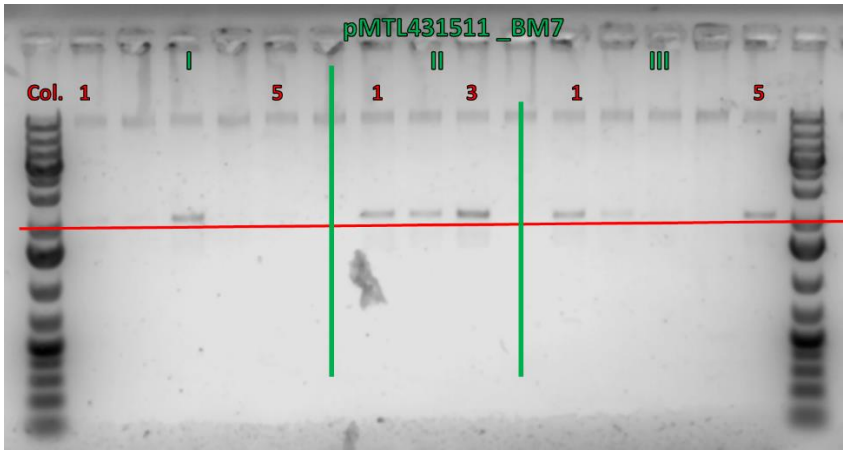
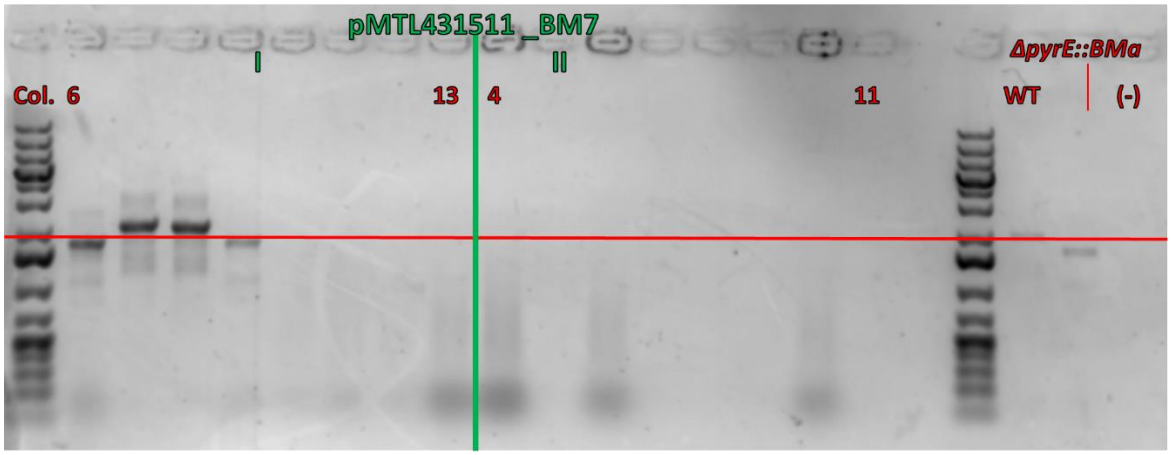



**Figure S. 5: Electrophoresis gel of *C. autoethanogenum* ⊿*pyrE::BM4-12* colonies obtained after conjugation of pMTL431511_BM7.** The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% (w/v) agarose gel alongside Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific. The expected size of the amplicon of a successfully complemented *pyrE* locus is 2 kb, versus 1.7 kb for the ⊿*pyrE::BM4-12* background. The different replicates are separated by green vertical lines and labelled with roman numerals. I = first replicate; II = second replicate; III = third replicate; WT= wild-type *pyrE* locus (2kb); (-) = negative control of colony PCR without DNA template; Col.: colony.



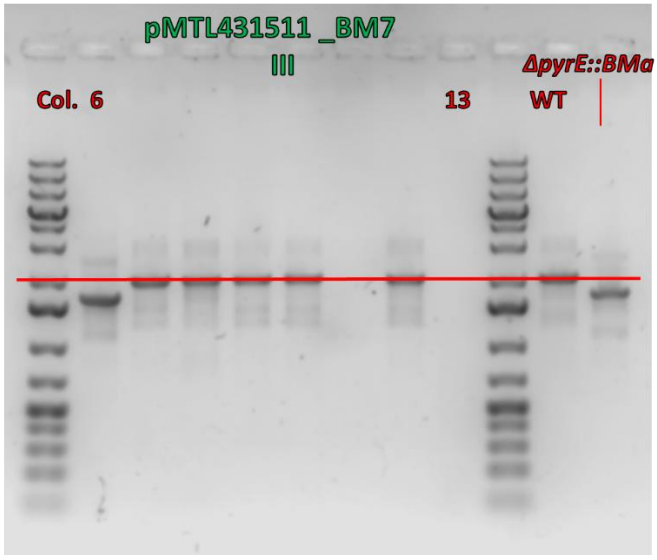
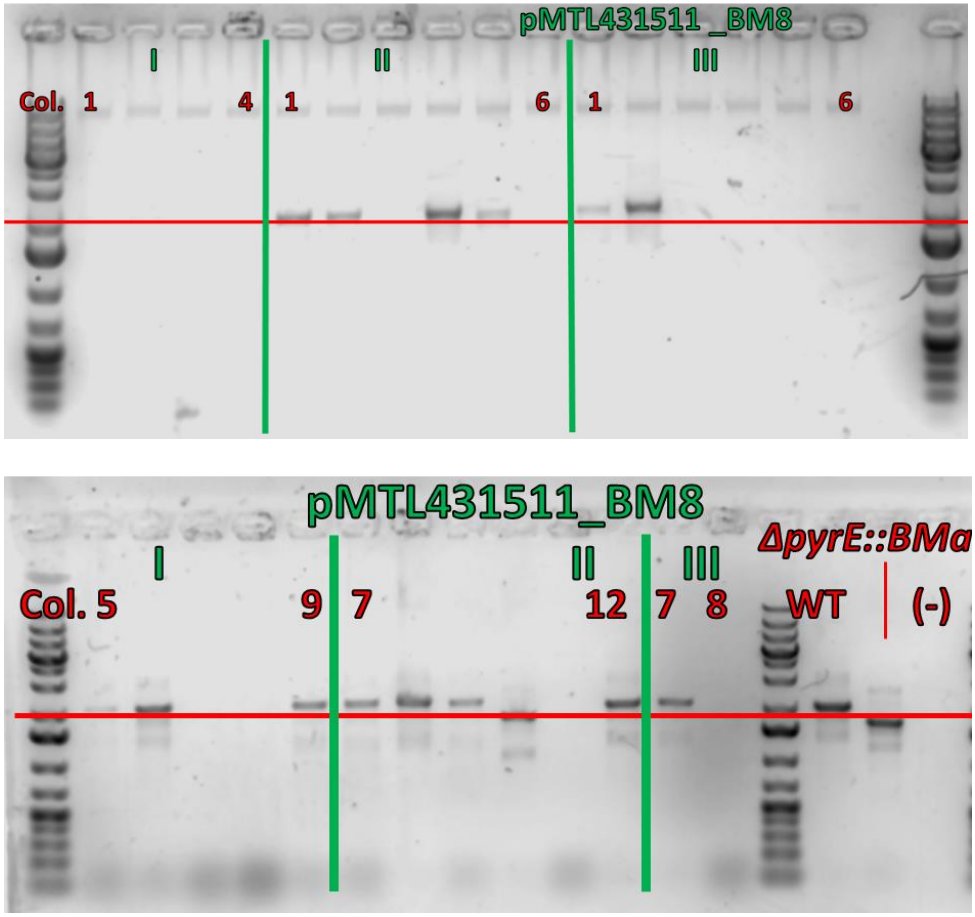
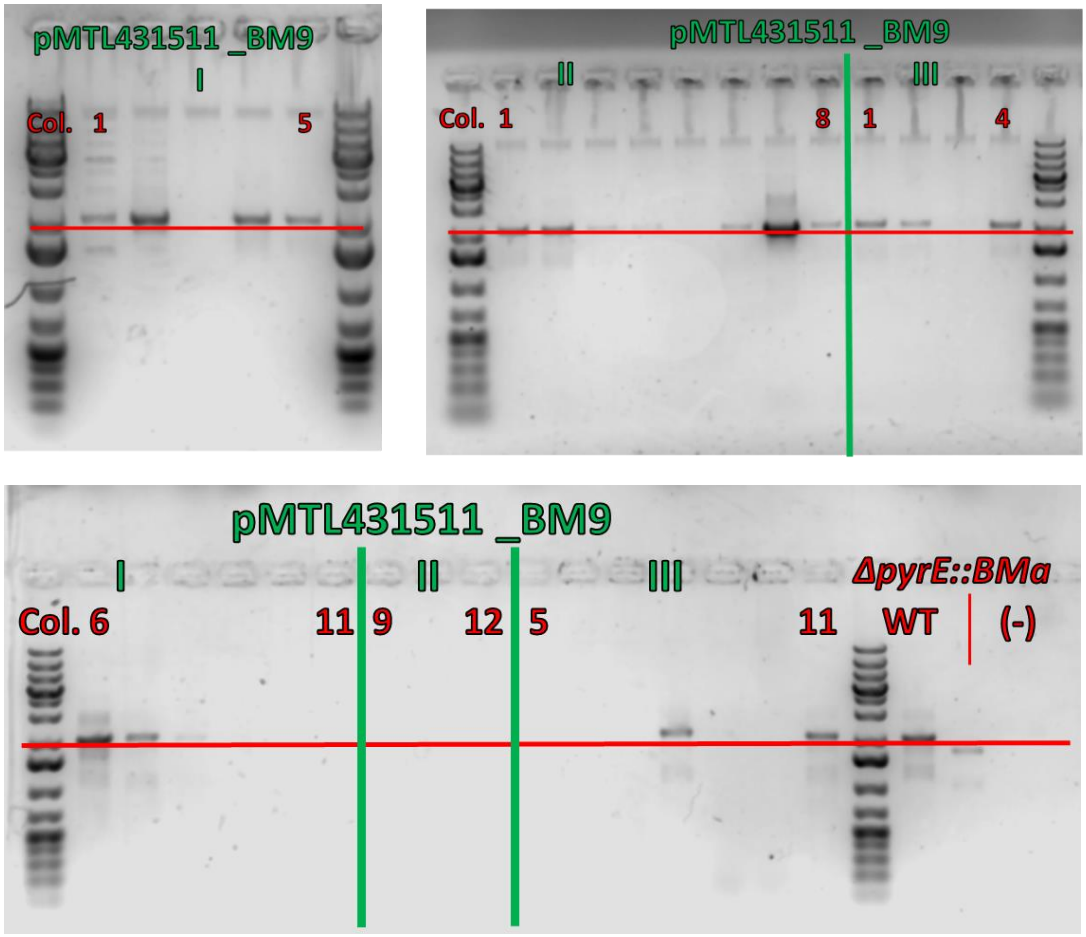



**Figure S. 6: Electrophoresis gel of *C. autoethanogenum ΔpyrE::BM4-12* colonies obtained after conjugation of pMTL431511_BM8.** The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% (w/v) agarose gel alongside Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific. The expected size of the amplicon of a successfully complemented *pyrE* locus is 2 kb, versus 1.7 kb for the *ΔpyrE::BM4-12* background. The different replicates are separated by green vertical lines and labelled with roman numerals. I = first replicate; II = second replicate; III = third replicate; WT= wild-type *pyrE* locus (2 kb); (-) = negative control of colony PCR without DNA template; Col.: colony.



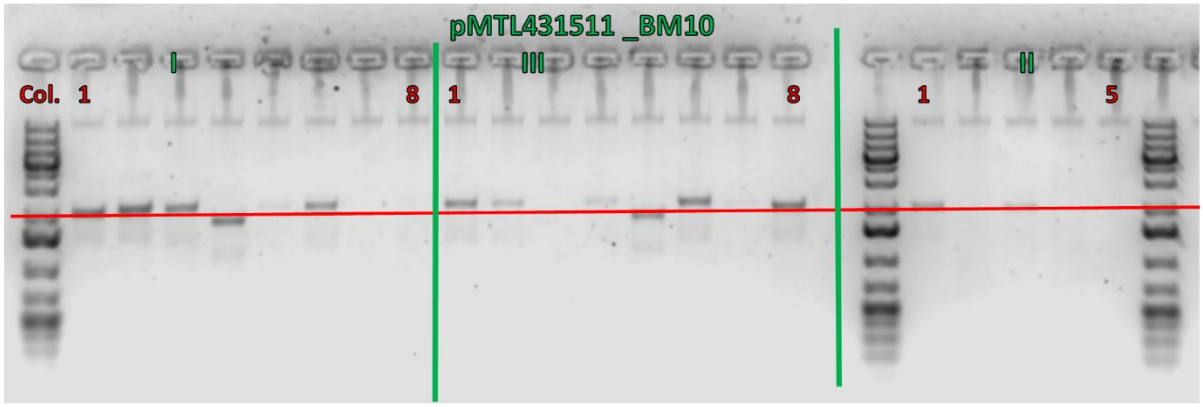
### VIII.3.1.2)(6) pMTL83151_BM9

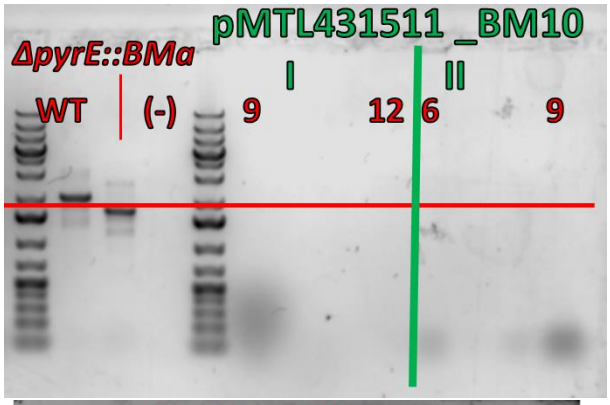
**Figure S. 7: Electrophoresis gel of *C. autoethanogenum ΔpyrE::BM4-12* colonies obtained after conjugation of pMTL431511_BM9.** The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% (w/v) agarose gel alongside Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific. The expected size of the amplicon of a successfully complemented *pyrE* locus is 2 kb, versus 1.7 kb for the *ΔpyrE::BM4-12* background. The different replicates are separated by green vertical lines and labelled with roman numerals. I = first replicate; II = second replicate; III = third replicate; WT= wild-type *pyrE* locus (2 kb); (-) = negative control of colony PCR without DNA template; Col.: colony.



### VIII.3.1.2)(7) pMTL83151_BM10

**Figure S. 8: Electrophoresis gel of *C. autoethanogenum ΔpyrE::BM4-12* colonies obtained after conjugation of pMTL431511_BM10.** The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% (w/v) agarose gel alongside Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific. The expected size of the amplicon of a successfully complemented *pyrE* locus is 2 kb, versus 1.7 kb for the *ΔpyrE::BM4-12* background. The different replicates are separated by green vertical lines and labelled with roman numerals. I = first replicate; II = second replicate; III = third replicate; WT= wild-type *pyrE* locus (2 kb); (-) = negative control of colony PCR without DNA template; Col.: colony.



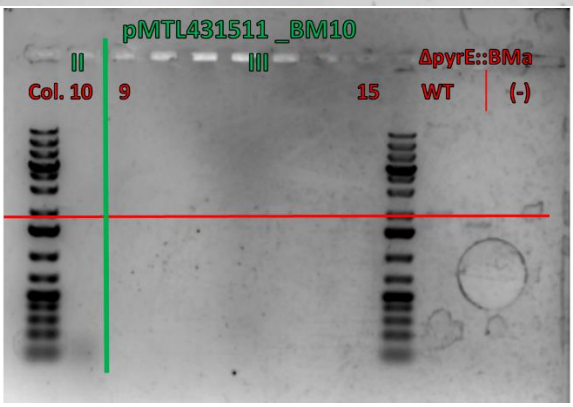
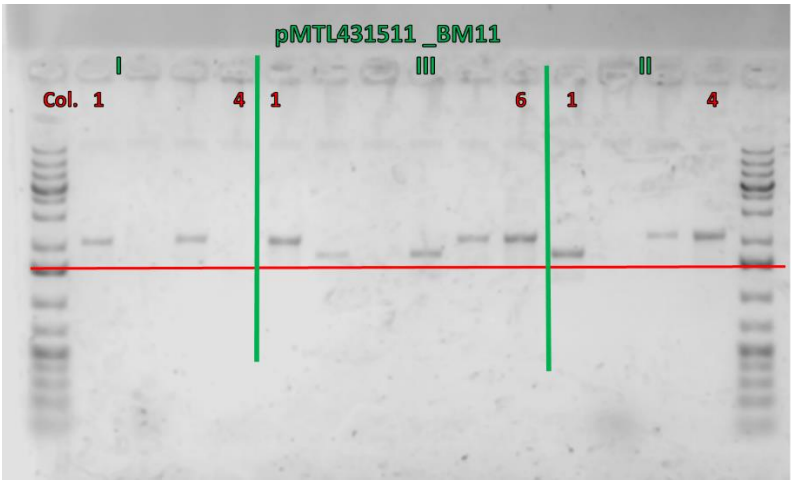
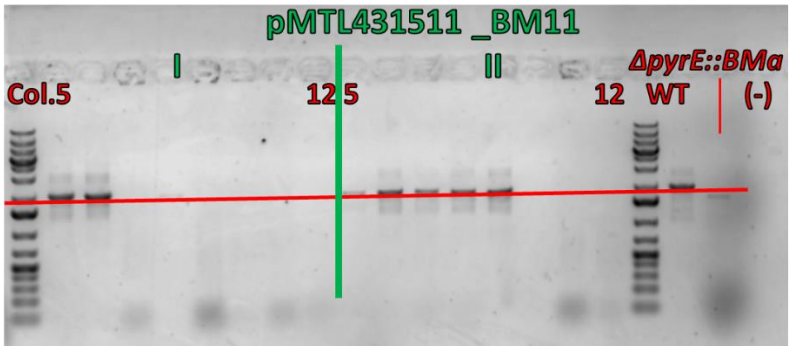



**Figure S. 9: Electrophoresis gel of *C. autoethanogenum* *ΔpyrE::BM4-12* colonies obtained after conjugation of pMTL431511_BM11.** The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% (w/v) agarose gel alongside Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific. The expected size of the amplicon of a successfully complemented *pyrE* locus is 2 kb, versus 1.7 kb for the *ΔpyrE::BM4-12* background. The different replicates are separated by green vertical lines and labelled with roman numerals. I = first replicate; II = second replicate; III = third replicate; WT= wild-type *pyrE* locus (2 kb); (-) = negative control of colony PCR without DNA template; Col.: colony.



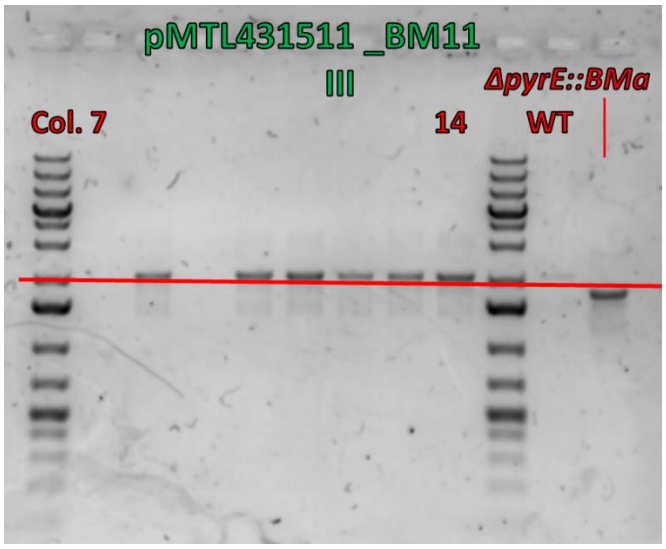



**Figure S. 10: Electrophoresis gel of *C. autoethanogenum ΔpyrE::BM4-12* colonies obtained after conjugation of pMTL431511_BM12.** The *pyrE* locus of each colony was amplified using the primers oFS105 and oFS106 then run on a 1% (w/v) agarose gel alongside Generuler 1 Kb Plus DNA ladder, ThermoFisher Scientific. The expected size of the amplicon of a successfully complemented *pyrE* locus is 2 kb, versus 1.7 kb for the *ΔpyrE::BM4-12* background. The different replicates are separated by green vertical lines and labelled with roman numerals. I = first replicate; II = second replicate; III = third replicate; WT= wild-type *pyrE* locus (2 kb); (-) = negative control of colony PCR without DNA template; Col.: colony.