Synthesis of Muscleblind-Like Splicing Regulator 1 (MBNL1) protein and CUG^{exp} Toxic RNA biomolecules in Myotonic Dystrophy Type 1 Disease

Christopher Anthony Waugh, BSc, MSc

Thesis submitted to the University of Nottingham for the degree of Master of Research.

September 2025

Abstract

Myotonic Dystrophy type 1 (*dystrophia myotonica* type 1 (DM1)) is a debilitating form of Muscular Dystrophy disease, and the most common form of DM affecting ~1 in 8000 individuals. The disease presents an array of multisystemic, typically neuromuscular symptoms, where complications vary considerably between patients, manifesting in an age-onset manner in adult DM1, or earlier in childhood and congenital DM1. Symptomatic relief of DM1 symptoms is the current treatment route under medical consortium guidelines, with no approved available treatments to cure DM1 disease, although promising novel therapeutics are at various stages of clinical trials.

The cause of DM1 is a trinucleotide repeat CTG ($_{n=50}$ - >3000) sequence located within the 3'-UTR ends of the *dystrophia myotonica protein kinase (DMPK)* gene, with disease severity directly correlated to repeat length. The transcribed mRNA forms highly structured hairpin conformations, preferentially binding splicing factor Muscleblind-Like Splicing Regulator 1 (MBNL1) protein, and other co-factors, forming large RNA-Protein "foci" located typically in the cell nucleus, or cytoplasm. These complexes sequester MBNL1 from its normal function, leading to DM1 symptoms.

Here, the aim of the study presented is to synthesise DM1 representative mRNA and MBNL1 protein constructs for downstream structural analysis of the RNA-MBNL1 complexes. Current structure data depositions relating to MBNL1 domains with toxic RNA are few, and none on intact DM1 disease complexes. Therefore, the ability to do biophysical investigations into DM1 disease complexes in this detail would allow for future Structure-Based Drug Design (SBDD), an approach widely adopted to increase success of novel treatment compound development.

Acknowledgements

Firstly, I would like to thank my supervisors Dr Aditi Borkar and Professor David Brook, for their supervision and help throughout the MRes project. Their insights and guidance have been essential towards all aspects of this project and have provided me with a very in-depth understanding of DM1 disease, and an excellent working understanding of developing key molecular biology processes. I have thoroughly enjoyed being their student and working together!

I would like to thank everyone in Dr Aditi Borkar's Lab group who have helped me on a professional and personal level: Shannon, Alex, Kim, Ceri, Goshen. I also want to thank the wider Vet School research community, A49 laboratory friends, and Office buddies. A special thank you to Andu, Val, Nisha, and Britany for all the good times. I would also like to thank members of Professor David Brooks laboratory, for all their help throughout the MRes.

I would also like to thank Dr Peter E. Wright, at Scripps US, for his help in kindly providing MBNL1 expression plasmids, and guidance with these.

My partner Sim for her constant help, love, and ability to make every discussion relating to work or life more logical, easier to handle, and more fun!

Finally, my family for all their love and support throughout the course duration, and my entire life, and for inspiring me to pursue the things in life that really matter.

Table of Contents

Abstract **2**
Acknowledgements **3**
Table of Contents **4**
Chapter 1: Introduction **6**
Chapter 2: MBNL1 expression and purification. **11**
2.1. Methods **11**
2.2. Results **14**
2.3 Discussion **23**
Chapter 3: Generating DM1 *in-vitro* Trinucleotide Repeat RNA **29**
3.1. Methods **29**
3.2. Results: **34**
3.3. Discussion **49**
Chapter 4: Future perspectives **55**
Chapter 5: Conclusions **57**
References **58**

Table 1) List of Key abbreviations used throughout the dissertation.

Abbreviation	Definition
AEC	Anion-Exchange Chromatography
βME	2-Mercaptoethanol
CiPP	Capture Intermediate Polishing Purification (CiPP) pipeline
C-term	Protein C-terminus
Cryo-EM	Cryo-Electron Microscopy
Cryo-ET	Cryo-Electron Tomography
CUG^{exp}	Expanded triplicate repeats (CUG)
DM1	Myotonic Dystrophy Type 1
DMPK	myotonic dystrophy protein kinase (DMPK) gene
DTT	1,4-Dithiothreitol
GSH	Glutathione
GST	Glutathione S-transferase
IEX	Ion-Exchange Chromatography
IMAC	Immobilised Metal Affinity Chromatography
Luc	Luciferase reporter gene
MBNL1	Muscleblind-like Splicing Regulator 1 protein
N-term	Protein N-terminus
pBSKII	pBlueScript II SK (+) backbone cloning vector, bacterial
pGEX-6P-1	Expression vector for expressing GST-tagged proteins, bacterial
pUC19	pUC19 backbone cloning vector, bacterial
pTRE2	Expression vector with a Tet-responsive promoter, bacterial
RIN	RNA Integrity Number (RIN) score
SEC	Size-Exclusion Chromatography
T7	Bacteriophage T7, origin for T7 promoter
tHDA	thermostable helicase-dependent amplification
Tte-UvrD	DNA Helicase Tte-UvrD
ZnF	Protein Zinc-Finger DNA/RNA Binding domain

Chapter 1: Introduction

1.1. Myotonic Dystrophy Type 1 (DM1) Disease

Myotonic Dystrophy Type 1 (DM1) disease is an inherited autosomal dominant disorder and is the most common form of muscular dystrophy [1, 2], estimated to be affecting around ~1 in 8000 people globally, with considerable regional variation observed[3-5]. Average prevalence estimates are likely also to be underrepresented[6]. DM1 disease is caused by repeat triplicate CTG regions located within the *dystrophia myotonica protein kinase (DMPK)* gene at the 3'UTR end, located on chromosome 19q13.3[7-9]. It is differentiated from the rarer Myotonic Dystrophy Type 2 (DM2) disease by genetic basis, which results from a quadruplicate CCTG repeat in *zinc-finger protein 9 (ZNF9)* intron 1 gene, on chromosome 3q21[10]. Characteristic clinical manifestations of DM1 pathology include myotonia, distal muscle weakness, atrophy, and, like many other neuromuscular inherited autosomal dominant disorders, a plethora of further complications related to repeat variability. This ranges from pulmonary and respiratory complications, which are the leading cause of death amongst DM1 patients, followed by cardiovascular complications, ocular abnormalities, gastrointestinal complications, and neuropsychiatric symptoms[11, 12]. DM1 has even been linked to tumour incidence, and endocrine/metabolic symptoms[13]. Consensus care, as it stands, does not offer any direct DM1 treatment and relies on chronic symptomatic disease management[13-16]. One of the characteristic issues associated with DM1 disease treatment, complication prediction, and symptom onset uncertainty is the high variability of repeat length and inherent instability of repeat regions, expressed in a tissue-specific manner, particularly in musculoskeletal tissue[13].

The presence of toxic CUG repeat mRNA transcripts from the 3' UTR of the *DMPK* gene has been accredited as the main disease driver of DM1 pathology since its discovery[7, 9]. These start to form highly structured mRNA, which preferentially bind key regulator proteins, disrupting regular cellular transcriptional regulation, and leading to symptoms observed during DM1 pathology[7, 8]. One of the hallmarks of the disease is the formation of large, dynamic, and variable Ribonucleoprotein (RNP) Foci [17-22] localised mainly to the nucleus, but also found to be present in the cell cytoplasm[18, 23, 24]. As it stands, key molecules present in CUG^{exp}-protein foci complexes include Muscleblind protein family members; MBNL1-3[25], MBLL, MBXL[26], and CUG-Binding Protein 1 (CUGBP1)[27]. Other factors described include proteins hnRNP H, H2, H3, F, A2/B1, K, L, DDX5, DDX17 and DHX9, although the function of all these factors in nuclei formation remains unclear[28]. Larger RNA nuclear foci have been observed to form with increasing amounts of MBNL1 protein present[29, 30], which aided the target selection criteria for choosing MBNL1-CUG^{exp} interactions in this study. The relative size of nuclear RNP foci formation is highly variable, and has been described as <200nm diameter for single foci, and ~200-1250nm for foci aggregates, where repeats were $CUG_{n=145}$ in length[18, 29]. The main molecular mechanism driving DM1 pathology is the preferential binding of key gene and transcription regulators, specifically MBNL1, to toxic CUG repeat RNA, which

“sequesters” the protein in the nucleus in foci complex. This lowers their availability, misregulates target expression within tissues and exacerbates symptomatic complications in DM1. Examples to try to understand individual pathways, which may be concurrent factors to consider in DM1 pathology have been proposed, such as: miRNA downregulation attributing to cardiac dysfunction in DM1 [31], or mis-regulation of polyadenylation patterns on 3’ UTR’s of pre-mRNA targets reverting to foetal patterns where downregulation of MBNL1 occurred in mice, errors which are linked to a number of human hereditary diseases[32].

Muscleblind-like protein 1 (MBNL1) in DM1 Disease

MBNL1 protein has been widely shown to be a key component in DM1 pathology, and presence in RNP foci formation is well-documented[24, 33]. MBNL1 is an RNA metabolism regulator protein, showing increasing expression and presence during tissue differentiation[34]. It is a pre-mRNA target mediator, and functions by both repressing or activating pre-mRNA splicing. Respective to its normal function, MBNL1 recognises the 5'-YGCU(U/G)Y-3' sequence, and has been shown to bind many pre-mRNA targets, including Cardiac Troponin T (*TNNT2*) exon 5 inclusion inhibition, Insulin Receptor (*INSR*) exon inclusion induction [35, 36], Chloride Channel 1 (*CLCN1*)[37], Calcium Channel voltage-dependent L type alpha 1S subunit (*CACNA1S*)[38], Bridging Integrator 1 *(BIN1*)[39], MBNL2 [40], and is even implicated in MBNL1 autoregulation[41].

The key features in MBNL1 structure are four CCCH Zinc Finger (ZnF) domains, arranged in pairs (1-2/3-4) and separated by a short flexible linker sequence[42]. These domains are in the N-terminus of the amino acid sequence and are the binding elements that recognise 5'-YGCU(U/G)Y-3' RNA sequences. The C-terminus of MBNL1 has been described to have a function around intracellular localisation of the protein in *Drosophila* [43], and in human MBNL1, for example, through the discovery of two Nuclear Localisation Signals (NLS) in *MBNL1*, with alternative splicing of exon 7 leading to “switching” between these, resulting in localisation changes[44]. The C-terminus of MBNL1 has also been shown to have a propensity to self-associate [45], which could provide further evidence of the tendencies for co-localisation and retention within RNP foci. MBNL1 is fairly ubiquitously expressed throughout most tissue types, however, its expression levels vary throughout these, with the highest expression levels found in skeletal muscle and cardiac tissue during myoblast differentiation [26, 46]. Its cellular localisation is found throughout the cell, seen in the cytoplasm, cytoplasmic stress granules [47], and colocalised in the nucleus [26].

Triplicate repeat mRNA transcripts in DM1 Disease

High repeat variability is observed throughout populations, ranging from 50 to >3000 repeats [15, 48], which makes the classification of patients and DM1 type difficult to consolidate into clinical practice guidelines. However, efforts have been made to separate groups based on CTG repeat length[13-15] and their clinical implications:

1) 5-39 CTG repeats, non-DM1.
2) 39-50 CTG Repeats, Pre-mutation / non-DM1 Phenotype.
3) 51->150: Classical, Juvenile, Congenital DM1 (typically observed in >150 repeats).

Research has shown a strong association between repeat length and age/ symptom onset, with disease complications and symptom severity increasing with repeat length. As an unstable, autosomal dominant repeat, expansion of the gene repeats[9] is typically observed in 51->150 repeat lengths, and tends to expand in length through generations, known as genetic 'anticipation'[49]. Expansion has been observed to be driven in a maternal, sex-dependent fashion, observed in congenital forms of DM1 (generally >1,000s repeats) where repeat expansion may be attributed to pre-meiotic or post-zygotic events[50-53]. This gene instability has also been observed to vary in a tissue-dependent manner, observed in DM1 cell-line fibroblasts during proliferation[54], and in mice models showing an age-dependent, tissue-specific expansion (highest observed in kidneys, also with observed mosaicism in an age-dependent fashion[55]). All these factors above have made the concise outcome prediction of DM1 implications for each patient difficult. The structure of CUG^{exp} repeats has been shown to form stable stem-loop hairpin conformations[56], and proposed to bind RNA-binding proteins at the stem conformation regions[25]. Short CUG RNA transcripts $r(CUG)_{n=6}$ adopt an A-form helical structure, with U-U mismatches not bound by hydrogen bonding, but rather G-C base-stacking events[57]. Increasing mismatches of these bases could distort the RNA from a typical A-form helical structure, promoting the creation of RNA-protein binding sites for existing cellular machinery processes. Research has also proven that binding partners, such as those in foci, CUGBP1 and MBNL1, preferentially bind 3'-UTR mRNA targets[58], and have a strong binding affinity for CUG repeat regions[59].

The treatment landscape for DM1 Disease

DM1 disease remains a chronic, incurable disease that relies on symptomatic treatment of patient-to-patient disease manifestation through continuous care and monitoring, further complicated by factors such as symptomatic age-onset variation in DM1 patients [13, 60]. Nonetheless, many efforts within the drug development world have been progressing to develop orphan drugs to either treat the disease at its core or achieve better symptomatic treatment during DM1 pathology. This is reflected in the twenty preclinical and clinical candidate drugs in current pipelines, with another three first-in-man interventional Phase III trials started from 2021-2022[60, 61]. Main classes of candidate molecules include: Repurposed and novel small molecule drugs[62], Nucleic acid therapies[63-65], and Gene editing/ engineering therapies[60, 66-69]. A lot of interesting, varied approaches are in development to disrupt different stages of a very complicated and dynamic condition at the molecular level. Associated DM1 disease targets are increasingly becoming the focus of drug development for treatments, for example, CDK12 inhibition by small drug molecules, showing promise by reduction of RNP foci aggregates[70, 71].

Existing structural studies into DM1 RNA-protein complexes:

The work presented in this project has been focused on setting up expression and reconstitution of the MBNL1-CUG^{exp} RNA-protein (RNP) complex, to decipher a high-resolution structure of DM1 core complex elements in a disease-representative fashion. Having this kind of structural data can allow for Structure-Based Drug Design (SBDD). This approach has been widely adopted throughout the field of medicinal chemistry[72], reflected in drug target discovery and implementation into pharma R&D pipelines, trends showing a dramatic increase in molecular targets, for example, 324 human or pathogen targets in 2006[73], sharply rising to 893 targets by 2016[74], in part due to consistent advances in structural biology enabling new target determination and discovery.

When reviewing current high-resolution structures deposited for MBNL1 and DM1 toxic RNA, separately or in combination, there is very little near-atomic resolution data available. Crystal structures derived from X-ray crystallography methods have been deposited for MBNL1 N-term Zinc finger domains, specifically tandem zinc fingers 1 and 2 (PDBe 3D2N) at 2.70Å, tandem zinc fingers 3 and 4 (PDBe 3D2Q) at 1.50Å, and tandem zinc fingers 3 and 4 in complex with RNA sequence 5'-CGCUGU-3' (PDBe 3D2S) at 1.70Å resolution[75]. These structures do not provide a disease-representative example of intact MBNL1 with DM1 toxic RNA, however, they provide a good understanding of the near-atomic architecture of the separate tandem zinc finger pair domains, as well as the binding observed with an RNA sequence containing one CUG repeat. Nuclear magnetic resonance (NMR) data has also been made available for deciphering solution structures of MBNL1 tandem zinc fingers 1 and 2 (PDBe 5U6H), tandem zinc fingers 3 and 4 (PDBe 5U6L), and tandem zinc fingers 3 and 4 in complex with cardiac Troponin T pre-mRNA sequence 5'-GUCUCGCUUUUCCCC-3' (PDBe 5U9B)[76]. Although the latter data did not yield a near-atomic resolution structure, in-solution structure data has provided valuable insights into binding patterns and residue contacts of ZnF domains with pre-mRNA targets. Crucially, lowest-energy NMR structure determined for zinc finger domains ZnF1-2 (structure residues 11-85 (NMR), 11-86 (X-ray)) and ZnF3-4 (structured residues 180-252 (NMR), 180-245 (X-ray) overlap very well in apo state[76], giving confidence in the architecture of these domains within MBNL1 by two powerful methods for structure determination.

A series of direct Transmission Electron Microscope (TEM) images were made available by Yuan et al., (2007)[45], visualising intact full-length MBNL1 in complex with stable $dsCUG_{n=136}$, with MBNL1 under these conditions forming a ring-like structure with a visual central cavity, and dsRNA forming a long rod-like structure. At molar ratios of 1:10 RNA: Protein, RNA structures are barely visible due to occlusion by multiple MBNL1 proteins binding. Further images with ssRNA to represent disease-state DM1 RNA would be ideal, particularly for direct observation of hairpin-like RNA conformations, and binding patterns with MBNL1.

Experimental Strategy and Aims:

To produce a library of DM1 representative CUG^{exp} RNA repeats and MBNL1 protein constructs, the core of the work was divided into two main sections: Firstly, heterologous expression of MBNL1 protein constructs in *E. coli*, and purification of the expressed protein using liquid chromatography techniques. Full-length MBNL1 protein as well as individual C-terminal and Zinc finger domains (ZnF) 1-4 were expressed from rationally designed, commercial plasmids containing affinity purification tags (GST and 6*His). In addition, plasmids containing expression sequences for MBNL1 ZnF1-2, MBNL1 ZnF3-4, MBNL1 ZnF1-4 functional domains were donated by the Wright Laboratory at Scripps, US, and Wright's purification pipeline was adapted for producing these constructs in the present work.

The second aspect of the project focused on creating DM1 RNA transcripts. Three DM1 classes of RNA repeats: $CUG_{n=11}$ (non-DM1), $CUG_{n=54}$ (asymptomatic, pre-mutation DM1), $CUG_{n=166}$ (classic DM1) were shortlisted and created by *in-vitro transcription (IVT)* techniques using a variety of commercially offered kits. These required linear DNA templates containing T7 promoter regions upstream of the transcription sequence of choice. Commercially synthesised plasmids were designed for this purpose and linearised by restriction digestion. PCR amplification techniques were also trialled, both to modify existing plasmids containing DNA sequences of interest and to amplify amount of linearised DNA template material for optimal IVT yield. Various RNA analysis techniques were used to identify the transcript species. These were also optimised to overcome difficulties when working with triplicate repeat-only RNA constructs.

Each chapter hereon includes all experimental findings, methodology, discussions on problems encountered, alternative options and process improvements, and future work. An in-depth discussion of future routes for structural and biophysical study of RNA-Protein complexes in DM1 disease is included to help guide future application of the work presented below.

Chapter 2: MBNL1 expression and purification.

Summary

The aim of this chapter was to optimise the heterologous bacterial expression and purification of MNBL1 protein domains such that the sample yield, stability and homogeneity is amenable to downstream structural characterisation.

2.1. Methods

A. Construct Design, Synthesis and Validation:

Canonical isoform 1 (UniProt Entry Q9NR56[77],) of the MBNL1 protein was selected, and in line with previous studies [45, 75, 76], constructs were designed (Table 2 and Figure 1) for expression in *E. coli*. All domain boundaries are numbered according to the canonical isoform described. Isoelectric point (pI) and molecular weight (Mw) were calculated using the Expasy webtool[78, 79], and included all construct design features.

Table 2) Summary of MBNL-1 protein constructs investigated in this project.

Construct	Amino Acids	Tags	Plasmid Backbone	Mw (kDa)	Isoelectric Point (pI)
ZnF1-4 (Zn finger containing N-terminal domain)	1-255	N-terminal GST and C-terminal 6x His	pGEX-6P-1	56.33	7.48
Full-length (FL)	1-388	GST, C-terminal 6x His		71.12	8.18
C-term	256-388	GST		43.37	6.27
ZnF1-2	1-92		pET21a(+)	10.56	9.13
ZnF3-4	173-255			9.66	8.83
ZnF1-4	1-255			28.06	8.98

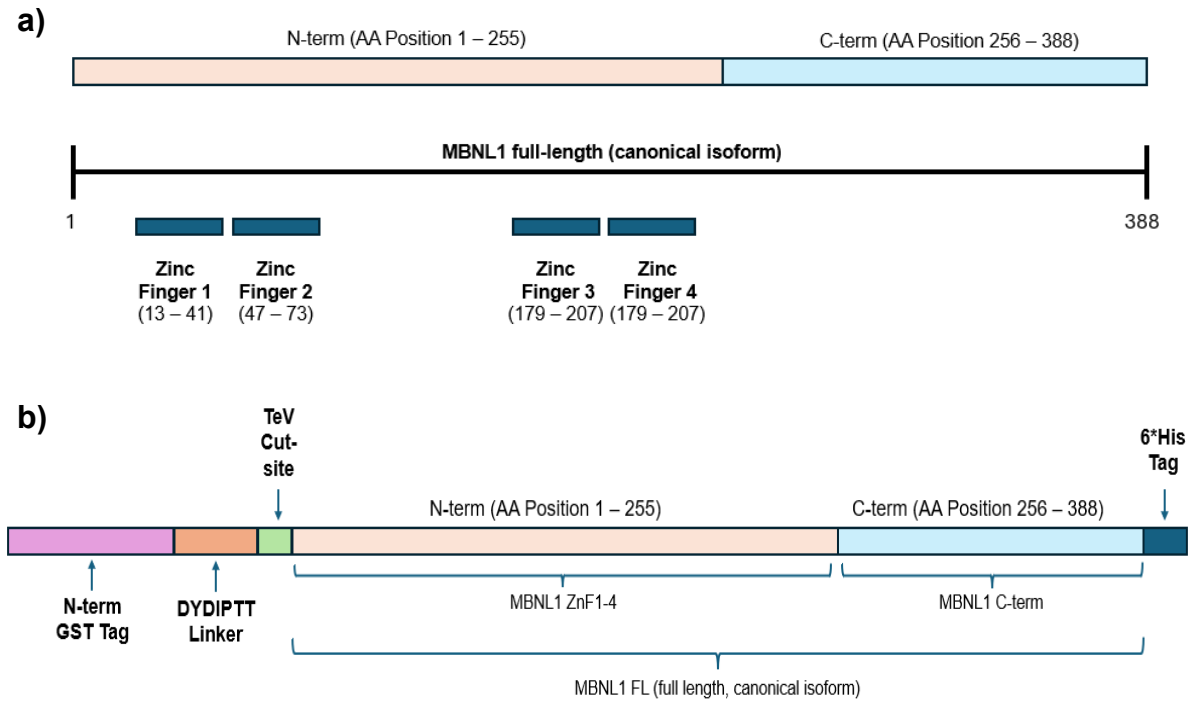


Figure 1a) Schematic of the MBNL1 domain organisation. **b)** MBNL1 constructs for heterologous expression. This contained an N-term affinity tag, a short linker and protease site upstream of the canonical sequence to assist affinity purification.

MBNL1-Cterm, FL and ZnF1-4 constructs were commercially synthesised (GenScript Biotech, NJ, USA) and cloned into pGEX-6P-1 vectors. N-term MBNL1 ZnF1-2, 3-4 and 1-4 construct plasmids were donated by Professor Peter E. Wright (Scripps Research, CA, USA) and were prepared by DNA sequence PCR amplification of MBNL1 gene from a full-length human liver cDNA library, and cloned into a pET21(a) backbone[76]. These plasmids were provided in lyophilised form and resuspended with 10 mM Tris-Cl, 0.1 mM EDTA pH 8.0.

Plasmids were cloned into Subcloning Efficiency™ DH5α Chemically Competent Cells (Invitrogen, MA, USA) and BL21(DE3) Competent Cells (Thermo Fisher Scientific, MA, USA), for plasmid library expansion and protein expression respectively. Plasmids were verified by restriction digestion using BamHI-HF®, EcoRV-HF®, and NotI-HF® (NEB LTD. Ltd, MA, USA), and agarose gel electrophoresis, and confirmed by supplier provided sequencing data.

B. Protein Expression and Purification:

Temperature of incubation, time of incubation and IPTG concentration was optimised for heterologous expression of all constructs in Table 2. Cultures were grown at 37°C, 200rpm in Luria-Bertini broth (and Terrific Broth for FL) supplemented with 100 µg/mL

Ampicillin and induced with either 0.3, 0.5, 1 and 1.5 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) at OD_{600} 0.6-0.8. For optimising time and temperature of incubation, upon induction, cultures were grown at 37 °C (2 and 3 hrs), 30 °C (3 and 4 hrs), and 21 °C (4 and 16 hrs). Uninduced cultures were used as controls. Cell pellets from 1.5 mL aliquots were resuspended in phosphate-buffered saline (PBS) and lysed by sonication (2 mins; 15's on, 5's off, amplitude 39%, on ice). Soluble and insoluble fractions were separated by centrifugation at 16000 x g, 4 °C for 30 mins and protein overexpression in each fraction was analysed by SDS-PAGE. Solubility of the FL construct was additionally tested by varying the composition of the lysis (binding buffer): 25 mM Tris-Cl, 5% Glycerol (pH 8.0), 50 mM MES (pH 6.5) or 50 mM HEPES (pH 7.6). All were supplemented with 500 mM NaCl, 10 mM βME and Protease Inhibitor Cocktail (with or without 1% Triton X-100). For each expression test, the intensity of the induced band in the soluble and insoluble fractions was compared to the intensity in corresponding region in the uninduced samples to determine the optimal expression condition.

Once the optimal expression conditions were identified, batch expression was scaled up to 4.5 (or 9) L cultures in LB (or TB) broth supplemented with 100 µg/mL Ampicillin and 0.15 mM $ZnSO_4$, and the recombinant protein was purified using tandem affinity, ion exchange, and size exclusion chromatography methods. All purification steps were performed at 4 °C using an ÄKTAxpress chromatography system (Cytiva Life Sciences, MA, USA) in buffers as summarised in Table 3. For affinity chromatography purification, the clarified cell lysate in binding buffer was loaded on pre-equilibrated GSTrap™ Fast Flow (or HisTrap™ HP) columns (Cytiva Life Sciences, MA, USA) followed by wash with 10 CV of wash buffer and 10 CV of binding buffer. The bound protein was eluted using a gradient of 1-100 mM GSH or 1-500 mM imidazole in binding buffer. For GST affinity, due to the slow binding kinetics of GST tag to the GSH beads, the first flow-through from the loading step was rerun through the column at reduced rates of 0.5 mL/min. Similarly, while eluting, the flow rate was reduced to 0.1 mL/min to allow the exchange of the bound protein with the reduced glutathione in the elution buffer. The GST tag was cleaved from the eluted protein using TeV protease (10 U/uL) at 40 U/mg of protein overnight at 4 °C.

For anion exchange chromatography, the eluted fractions from affinity chromatography were pooled, concentrated, and dialysed against 20 mM Tris-Cl, 0.5 mM DTT (pH 8.0) and exchanged with a 1 mL HiTrap™ Q SP column (Cytiva Life Sciences, MA, USA).

Size exclusion chromatography was performed on a 24 mL Superdex® 200 10/300 GL Gel filtration column, and fractions were eluted at 0.2 mL/min over 1.5 CV of 25 mM Tris-Cl, 500 mM NaCl, 2 mM DTT, 10% glycerol (pH 8.0). For improved separation of the GST tag from the protein of interest, we considered including a 1 mL GSTrap™ Fast Flow column in tandem with the SEC column.

Table 3) MBNL1 chromatography pipeline and buffer composition at each step.

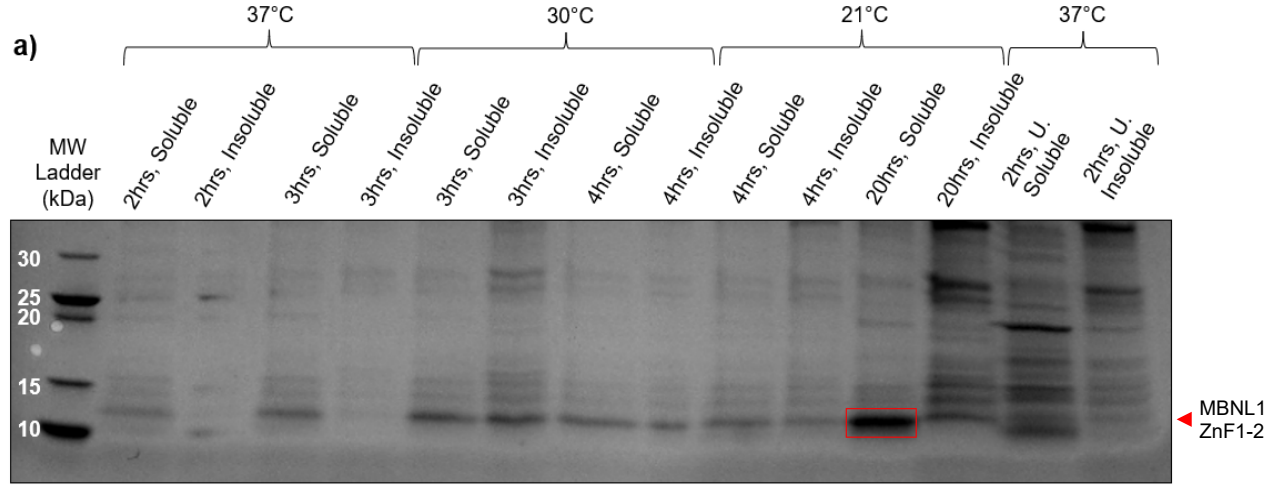
Construct	Chromatography		
	Affinity	**IEX**	**SEC**
ZnF1-4	GST & IMAC	NA	Yes
FL	GST & IMAC	NA	Yes
C-term	GST & IMAC	NA	Yes
ZnF1-2	IMAC	Yes	Yes
ZnF3-4	IMAC	Yes	Yes
ZnF1-4	IMAC	Yes	Yes
GST	**Buffer**		
Lysis	25 mM Tris–Cl, 500 mM NaCl, 10 mM βME, 5% glycerol, protease inhibitors, pH 8.0		
Equilibration	25 mM Tris–Cl, 500 mM NaCl, 10 mM βME, 5% glycerol, pH 8.0		
Wash	25 mM Tris–Cl, 500 mM NaCl, 10 mM βME, 5% glycerol, pH 8.0		
Elution	25 mM Tris-Cl, 20 mM reduced glutathione (GSH), 10 mM βME, 0.5M NaCl, pH 8.0		
IMAC			
Lysis	20 mM Tris-HCl, 100 mM NaCl, 2 mM DTT (pH 8.0) with Protease Inhibitor Cocktail.		
Equilibration	20 mM Tris-HCl, 100 mM NaCl, 2 mM DTT, pH 8.0		
Wash	20 mM Tris-HCl, 500 mM NaCl, 2 mM DTT, 50 mM Imidazole, pH 8.0		
Elution	20 mM Tris-HCl, 100 mM NaCl, 2 mM DTT, 500 mM Imidazole, pH 8.0		
AEC			
Equilibration	20 mM Tris-Cl, 0.5 mM DTT, pH 8.0		
Wash	20 mM Tris-Cl, 0.5 mM DTT, pH 8.0		
Elution	20 mM Tris-Cl, 0.5 mM DTT, 1M NaCl, pH 8.0		
SEC			
Equilibration and elution	25 mM Tris-Cl, 500 mM NaCl, 2 mM DTT, 10% glycerol (pH 8.0).		

2.2. Results

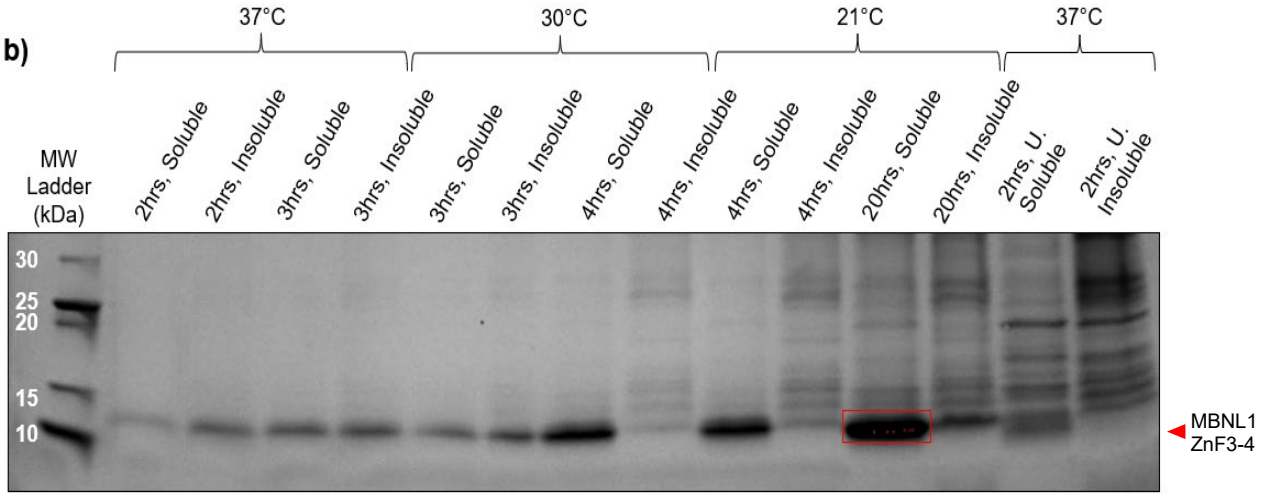
The results presented below detail efforts to optimally express, scale-up and purify MBNL1 constructs for reconstitution with DM1 RNA. To optimise the expression yield for each construct, the effects of culture media type (LB or TB), temperature of expression (21, 30 or 37), time of incubation (30 mins or few hours), and concentration of inducing agent (IPTG 0.5 vs 1 mM) was determined through a series of expression tests. Following this, batch culture scale up at 4.5 and 9 litres was attempted. However, only MBNL1-FL scale-up was successful, whereas C-term and N-term constructs encountered difficulties. A detailed description of these findings is presented below.

N-term MBNL1 and MBNL1-FL expression test results:

MBNL1 ZnF1-2 (10.56kDa):



MBNL1 ZnF3-4 (9.65kDa):



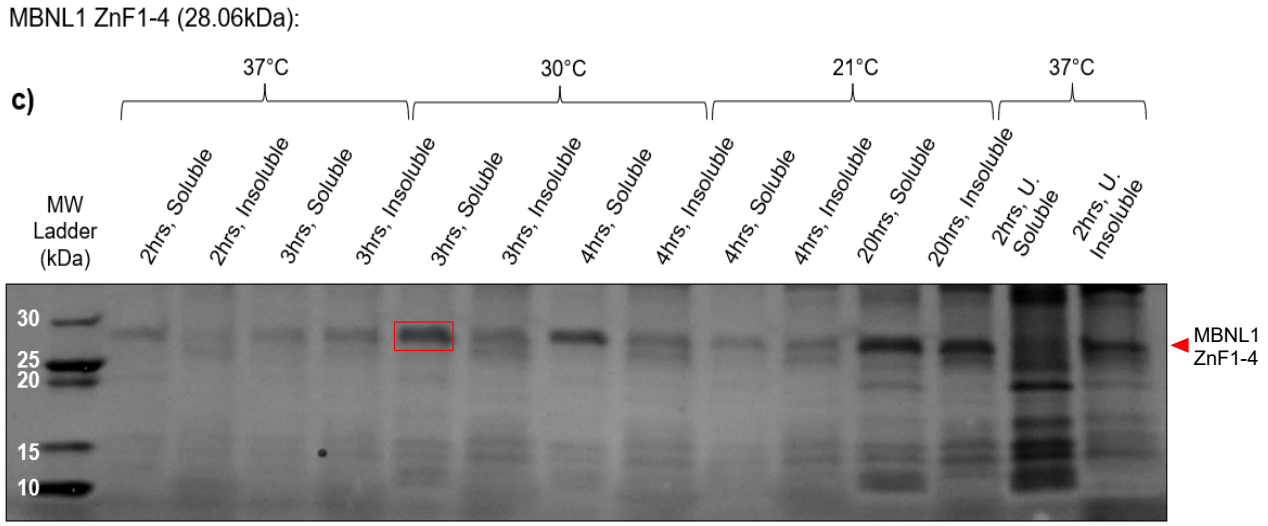


Figure 2) SDS-PAGE images of MBNL1 ZnF1-2 (**a**), MBNL1 ZnF3-4 (**b**), and MBNL1 ZnF1-4 (**c**) expression in BL21 (DE3) *E. coli* in LB-amp with 0.5mM IPTG Induction. Uninduced fractions are included for overexpression comparison. Red boxes highlight induction conditions selected for expression scale-up.

MBNL1 ZnF1-2, ZnF3-4 and ZnF1-4 expression test showed the presence of strong soluble overexpression bands at each predicted molecular weight location in figure 2, with ZnF3-4 appearing strongest. This visual overexpression compared to band weight location on uninduced samples allowed for easy condition selection to progress to the batch scale-up stage. MBNL1 ZnF1-4 differed from the other two constructs in selected conditions for scale-up (3 hrs, 30°C) due to the strong overexpression band presence in the insoluble fraction at 20 hrs, 21 °C, with little improvement to soluble fraction overexpression seen at 3 hrs, 30 °C, possibly resulting from inclusion body formation.

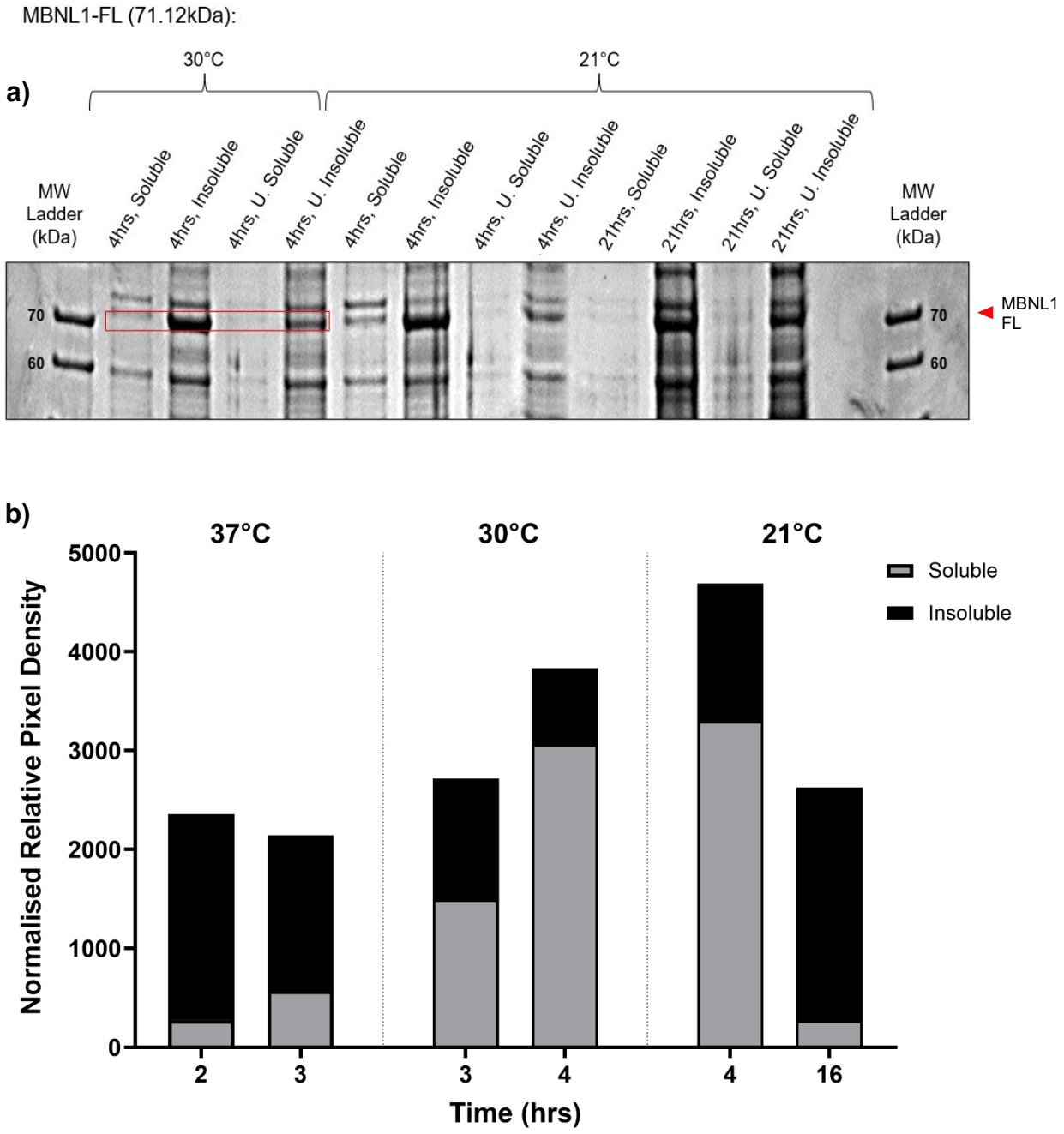


Figure 3) a) SDS-PAGE MBNL1-FL expression in LB-amp media with 1mM IPTG induction. Soluble and insoluble protein fractions are compared with uninduced culture fraction expression (n=1). The red boxes indicate overexpression band and region for MBNL1-FL. **b)** ImageJ analysis results were normalised at conditions of 37°C; 2/3hrs, 30°C; 3/4hrs, 21°C; 4/16hrs.

MBNL1-FL did not show overexpression at 0.5 mM IPTG, with visual improvements at 1mM IPTG (presented in figure 3 a-b). However, it displayed poor solubility, with most expressed protein in the insoluble fraction. Visual distinction of overexpression proved

difficult and required quantitation, therefore an ImageJ banding analysis comparison (figure 3b) helped aid the decision to choose the best condition for progressing to scale-up. For this, each band of interest location for each MBNL1-FL construct condition was selected and normalised against the whole background lane pixel intensity value. Then, uninduced values were subtracted from induced pixel density values. Higher soluble/ insoluble ratio values for 4 hrs at 30 ºC were quantified and were preferable to overnight growth to reduce background host-cell protein presence when taken to the purification stage.

MBNL1-FL protein lysis buffer solubility testing:

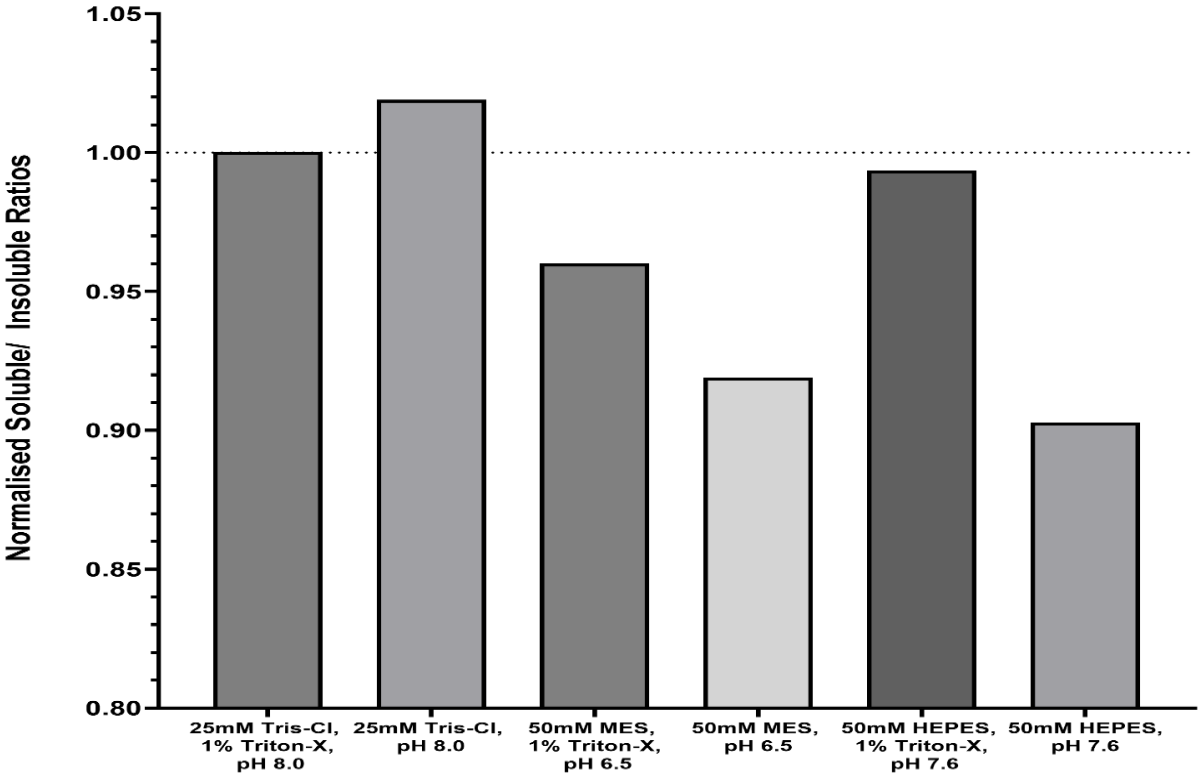


Figure 4) ImageJ analysis of MBNL1-FL SDS-PAGE solubilisation image displaying normalised soluble/ insoluble fraction pixel density ratios.

Protein solubility ImageJ analysis of SDS-PAGE images (Figure 4) showed the best solubility buffer composition for sonication lysis of *E. coli* MBNL1-FL cell pellets as 25 mM Tris-Cl, 500 mM NaCl, 10 mM βME, 5 % glycerol (pH 8.0). To improve on solubility issues observed with this construct, cell pellets were resuspended in equal volume in each buffer composition, and equal lysis conditions and SDS-PAGE sample loading. Here, ImageJ results are shown instead of the SDS-PAGE image as quantification of banding between lanes was required to best interpret subtle solubility differences. To achieve this, bands were normalised as described for figure 3 above, and ratios calculated from soluble divided by insoluble normalised pixel density values. The

uninduced sample used for this analysis was resuspended in 25mM Tris-Cl, 500mM NaCl, 5% glycerol buffer (pH 8.0).

MBNL1 C-term expression:

MBNL1 C-term induction with 1mM IPTG showed no overexpression bands (data not shown). With 0.3 mM and 1.5 mM IPTG induction, no protein band was observed at the expected molecular weight (43.4 kDa), although some additional high MW bands were observed. A bead-scale GST purification (Figure 5) was performed followed by Western blot analysis to check for the presence of MBNL1 C-term protein alone or within the high MW bands, outlined in red.

However, due to the lack of MBNL1 C-term band in Figure 5 elution fractions and poor-quality Western blot results, we were unable to confirm GST-MBNL1-Cterm protein identity neither in the predicted region nor for the high Mw band sizes observed.

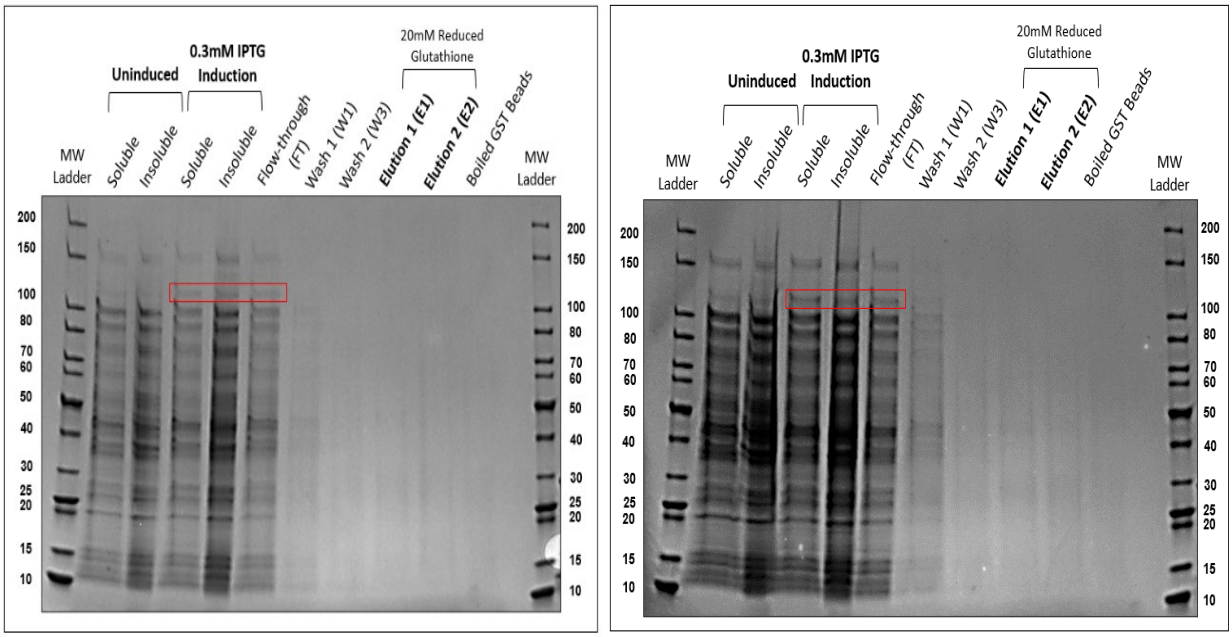


Figure 5) SDS-PAGE image of Glutathione bead-scale purification of **a)** 2 mL, and **b)** 5 mL MBNL1 C-term *E. coli* TB-amp lysed cell pellets (n=1 purification attempts), induced with 0.3 mM IPTG, and with soluble fraction loading from FT fraction onwards.

Overall, expression of soluble N-term MBNL1 constructs, and shake-flask expression scale-up experiments were successful. The best protein expression conditions for MBNL1-FL for scale-up expression determined, and scale-up experiments in LB and Terrific Broth were successful. Initial solubilisation trials to determine lysis and GST affinity purification loading buffer of choice were taken to downstream purification. With this in mind, MBNL1-Cterm and MBNL1 ZnF1-4 (from pGEX-6P-1 backbone) required more development work for achieving expression and identity validation. MBNL1

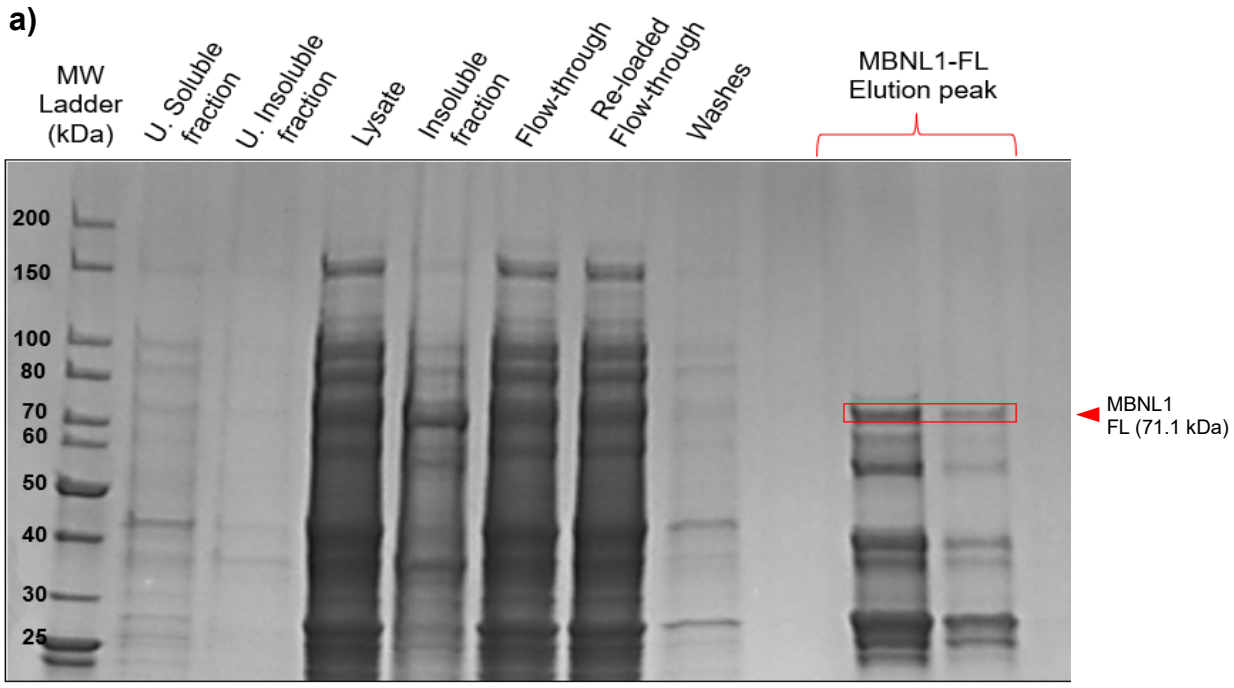
ZnF1-4 from pET21a(+) backbone (donated by the Wright Laboratory) construct expressed well in comparison (Figure 2c) and was selected from the two plasmid options for this construct to further pursue in downstream purification.

Purification of N-term MBNL1 constructs

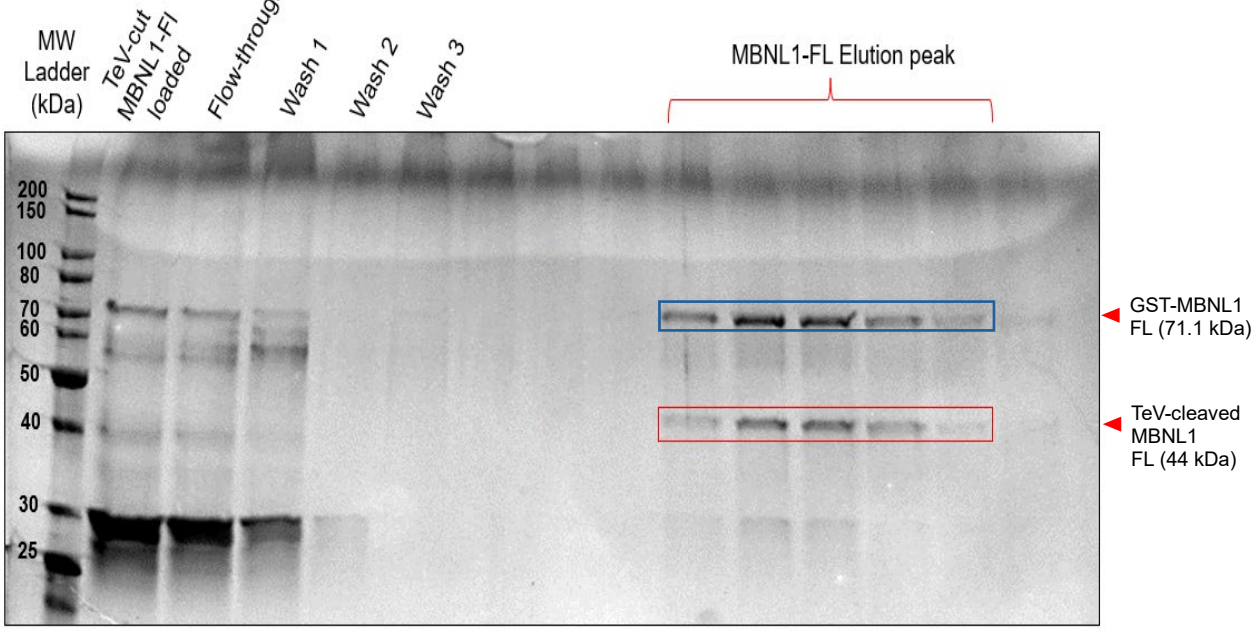
MBNL1 ZnF1-2, 3-4 and 1-4 domains, provided by the Wright laboratory, expressed well, as previously described in Figure 2. Scale-up pellets previously described for MBNL1 ZnF1-2, and MBNL1 ZnF1-4 were taken through to automated ÄKTA purification, first separated through an initial IMAC affinity step. MBNL1 ZnF3-4 purification was not attempted at scale here due to project time and resource constraints. As all purifications displayed consistent soluble protein loss in FT, Washes 1-3 and elution fractions (data not shown), the Wright laboratory submitted plasmids were checked, and revealed these did not have a 6x histidine tag, and the elution profile observed was likely from non-specific binding of overexpressed protein to the Ni-NTA column. MBNL1 ZnF1-4 elution fractions were still taken through to AEC, but the final yield was too low to proceed to SEC.

Further process development efforts were prioritised around the purification of MBNL1-FL, and all N-term MBNL1 expression plasmids and scale-up pellets were stored for future use.

MBNL1-FL purification:



b)


MW Ladder (kDa)
TeV-cut MBNL1-FL loaded
Flow-through
Wash 1
Wash 2
Wash 3
MBNL1-FL Elution peak
200
150
100
80
70
60
50
40
30
25
GST-MBNL1 FL (71.1 kDa)
TeV-cleaved MBNL1 FL (44 kDa)

c)

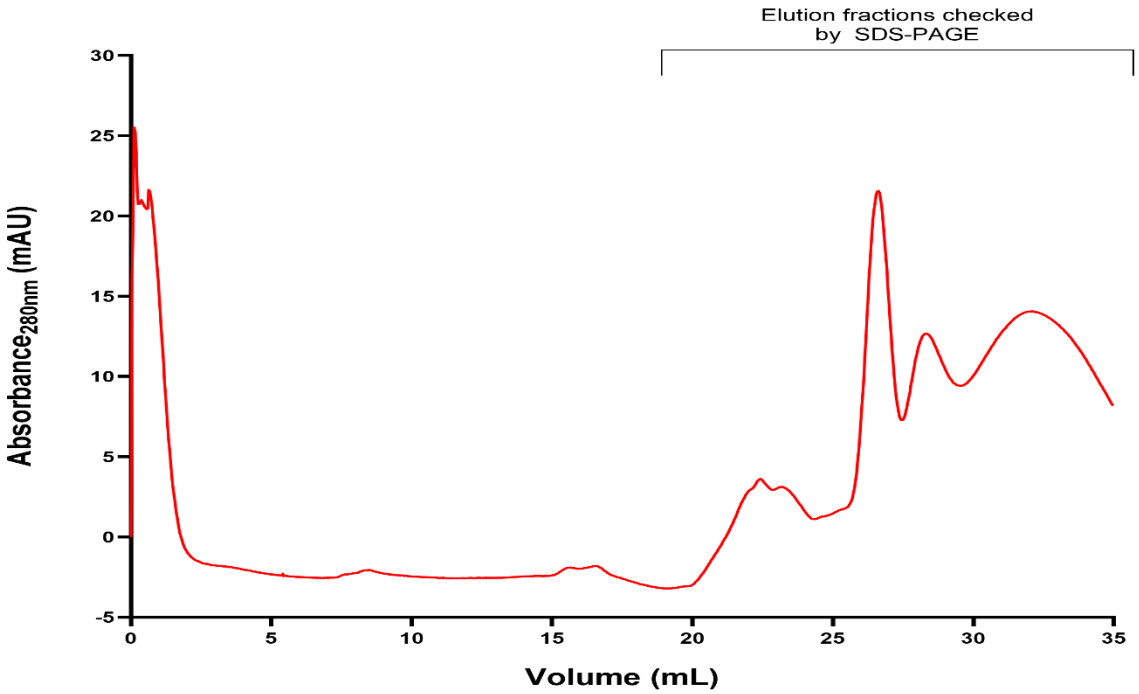

Elution fractions checked by SDS-PAGE
Absorbance280nm (mAU)
30
25
20
15
10
5
0
-5
0
5
10
15
20
25
30
35
Volume (mL)

Figure 6) a) SDS-PAGE gel image showing 10 mL GSTrap™ Fast Flow column purification, with 100 mL lysate loaded and 2 mL elution fractionation. The red boxes contain bands of the expected size for MBNL1-FL (71.1 kDa). **b)** IMAC Purification (1 mL column) with TeV-cleaved MBNL1-FL eluted under a 0.5 M imidazole gradient. The red box outlines cleaved MBNL1-FL (44 kDa) construct in the elution profile, and the blue box highlights the uncleaved MBNL1-FL. **c)** S200 SEC column purification (no GSTrap™ column attached) curves showing elution trace of 0.374 mg loaded cleaved MBNL1-FL.

The results shown above in figure 6 cover MBNL1-FL purification attempts taken through the designed CiPP pipeline. This pipeline aimed to achieve initial GST-MBNL1-FL separation through the initial affinity step from lysate, followed by GST-tag overnight cleavage by TeV, and separation of both cleaved tag and TeV protease from MBNL1-FL by IMAC. SEC then followed as a final polishing step for high purity sample recovery. GST purifications from LB-amp pellets gave poor observable recoveries of soluble MBNL1, and further purifications were discontinued. The GST purification result shown is from TB-amp pellets, which supported higher *E. coli* cell growth for better protein recovery.

Figure 6a displays MBNL1-FL being separated during GST purification, as identified in the elution gradient fractions. There was a strong presence of unknown lower molecular weight bands, which could be due to free-GST formation (27 kDa), GST dimerization (54 kDa), and truncation of MBNL1 protein constructs. Non-specific banding may also be expected at a stage with high concentrations of protein-rich lysate. Previous GST-purification attempts (not shown) followed a similar protein separation profile. However, when compared, AKTA A_{280} curves showed a noticeable difference in the elution peak symmetry, likely from increased free or dimerised GST tag presence. Nanodrop concentration values in elution fractions showed low target recoveries, approximately 0.12 and 0.14% of total lysate protein from 50 mL and 100 mL of soluble lysate loaded.

Figure 6b SDS-PAGE results show separation of cleaved MBNL1-FL in elution fractions (which contained the 6x His tag for column binding), and free-GST present at ~27 kDa in flow-through and wash fractions. Despite optimisation attempts, partial cleavage was seen as GST-MBNL1-FL was present in elution fractions. MBNL1-containing elution fractions were quantified at this stage by Qubit 4 Protein Assay to avoid imidazole interference when measuring samples at A_{280nm} wavelength. Here, an additional overnight TeV cleavage step was introduced before SEC to improve tag cleavage. Due to poor initial protein recovery from purification, the samples were spin concentrated to ~0.5 mg/mL to minimise potential aggregation.

Figure 6c shows SEC curves with peaks likely corresponding to monomeric cleaved and un-cleaved MBNL1-FL. Attaching the GSTrap™ 1 mL column to SEC did not yield visible MBNL1-FL when analysed by SDS-PAGE, and this method variation was discontinued. As SEC analysis did not yield gel-detectable MBNL1-FL, the full purification process would need to be repeated with higher soluble MBNL1-FL expression to obtain a pure protein sample for reconstitution with DM1 RNA.

Overall, both 3-step CiPP pipelines yielded insufficient amounts of pure MBNL1 constructs for downstream reconstitution with DM1 RNA. MBNL1-FL recovery was hindered by initial low soluble protein expression, cumulative effects of inadequate TeV cleavage, and protein loss through polishing steps, despite many optimisation efforts throughout. Nonetheless, process progression was achieved from expression of tagged MBNL1-FL to cleaved higher-purity MBNL1-FL. N-term MBNL1 constructs required an alternative purification strategy design due to the absence of the His-tag, and options for future improvements based on all the results obtained will be discussed further below.

2.3. Discussion

***Expression of MBNL1 constructs*:**

MBNL1 N-term and full-length (FL) construct expression was achieved as shown in figures 2 and 3, also observed in various studies where MBNL1 constructs are expressed with these same affinity tags[80], including MBNL1 full-length constructs produced from a pGEX-6P-1 backbone plasmid [45, 81]. Expression of N-term constructs donated by the Wright laboratory differed in use of minimal media and isotope labelling in preparation for NMR experiments [76]. Despite this variation, solubility did not present a problem, demonstrated in figure 2 soluble fraction overexpression.

Previous studies have used alternative affinity tag strategies to express and purify MBNL1. Botta et al. (2013) used only Histidine tags when expressing isoforms 40 & 43[82]. Yuan et al., (2007)[45] Ni-NTA employed expression from a pGEX-6P-1 backbone, but IMAC purification preceded the Glutathione affinity purification step, which included an on-column protease tag cut step. Published purification procedures also included a DNAse I step to remove bound nucleotides, potentially important for reconstitution with RNA.

Constructs from pGEX-6P-1 plasmid backbones initially posed some difficulties. Figure 3 shows lower overexpression and poor solubility post-sonication compared to N-term MBNL1 constructs. Of these, the most successfully expressed construct was MBNL1-FL. Strong MBNL1 presence in the insoluble fraction could be from inclusion body formation. Botta et al. (2013)[82] addressed this issue through sample preparation with 7M Urea to disrupt inclusion bodies before refolding, unlike the native conditions employed in this study. Research alluding to proteins having a role in self-association has also been described, with MBNL1 homotypic interactions confirmed *in vivo* by Two-hybrid yeast assay, during MBNL1-FL IP immunoprecipitation from transfected 293T cell lines with MBNL1-FL/ MBNL1-FL and MBNL1-FL/ MBNL1 C-term constructs[45]. This supports the observed difficulty in achieving MBNL1 C-term overexpression.

GST as an affinity tag has known properties for acting as a chaperone for protein folding and promoting protein solubility by reducing inclusion body formation[83]. GST-tag purifications for proteins used in structural studies have been extensively published[84, 85], and gave confidence in choosing this tag. A possible solution here could be using alternative *E. coli* strains, such as Rosetta (DE3) strain[86], for reducing premature translation termination. Alternative purification tags could be a good way to improve solubility and protein recovery[87], for example, Maltose-Binding-Protein (MBP) tags to reduce insoluble protein inclusion body formation [88, 89]. The linker used in this study, 5'- DYDIPTT-3' is a straight linker designed to provide spacing between GST-tag and MBNL1. However, there are other published linker sequences to consider during design[90], particularly to provide better stability and folding, for example, flexible linker $(GGGS)_3$ [91] or rigid linker $(EAAAK)_3$ [92]. To further improve solubility and folding, cytosolic expression could include co-expressing molecular chaperones to help guide cytosolic nascent chain for well-folded protein production[93, 94]. To reduce inclusion body formation, plasmid design could swap to secretion peptide sequences[95], at the cost of decreased expression levels. Considering these, redesigning constructs would need further expression and solubilisation trailing.

To improve high-yield expression in *E. coli*, a switch to Terrific-Broth was decided upon for growth post-induction, as TB contains higher Yeast extract to support nutritional demands, buffering capacities, and glycerol addition. This resulted in an improved biomass compared to LB (~15 mL PCV(TB) vs ~7 mL PCV (LB)). In addition to this, if suspected protein toxicity is hindering production and growth, which may contribute to observed low titres of FL and C-term constructs, strains BL21-AI, or BL21(DE3)pLys S, and E, can be selected to minimise low-level background expression before induction[96, 97]. Yield may be further improved through scaling expression inside a controlled bioreactor environment, which has higher aeration rates and tight pH control, allowing continual oxygen transfer for protein production and avoiding culture acidification. A fed-batch approach could also be adopted to avoid nutrient availability being a rate-limiting step on growth and protein synthesis[98].

To achieve the high yields needed for structural biology work, *E. coli* remains the preferred option intracellular recombinant protein expression[99] without target protein post-translational modifications. However, an alternative expression organism should be considered as a backup option for high-level heterologous expression of MBNL1. This includes, but is not limited to, yeast (e.g. *P. pastoris, S. cerevisiae*)[100], insect cell culture (e.g. Baculovirus expression systems in Sf21 insect cell-lines)[101], mammalian cell-culture [102], transgenic plant expression [103], and transgenic animal expression[104]. A cell-free expression system is another option to produce the constructs, which benefit from rapid protein production and minimal handling before downstream processing[105]. The drawback, however, is the scale-up limitations from the substantially higher cost. The general rule of thumb around organism up-scaling is that the yield of active compounds will decrease with increasing molecule quality[106]. *S. cerevisiae* or *P. pastoris* are often a good alternative organism as they share many of the benefits of prokaryotic suspension culture systems, in addition to well-studied eukaryotic PTM machinery and guided biomolecule secretion [100, 107, 108]. If new

expression designs and bioreactor control of *E. coli* protein expression do not yield desired MBNL1 constructs, *S. cerevisiae* would be a suitable alternative to try.

Solubilisation trial:

Persistent insolubility of MBNL1 FL and C-term led to the solubility testing trial presented in figure 4 to improve solubilisation of MBNL1 constructs at the lysis stage. The buffers shortlisted are well-established in protein biology research [109], and were shortlisted to 25 mM Tris-Cl, 50 mM MES and 50 mM HEPES. Each of the buffers selected contained 500 mM NaCl to reduce non-specific intramolecular and GST-resin interactions[110], 10 mM βME to reduce unwanted disulfide bond formation between cysteines to reduce aggregation, and 5% glycerol aimed at reducing aggregation and promoting solubility[111, 112]. These buffers were trialled with and without 1% Triton-X100 for potential improvement in solubility, a common strategy employed with molecules containing highly hydrophobic segments[113]. A pH of 8.0 was chosen based on previous publication success in the expression and purification of MBNL1 constructs[45, 76] and is within a physiological range. The findings saw minimal improvements between buffer components, with the most favourable composition being 25 mM Tris-Cl, 500 mM NaCl, 10 mM BME, 5% glycerol (pH 8.0). Interestingly, the addition of 1% Triton X-100 detergent did not significantly improve MBNL1-FL solubility. Due to time constraints, this small trial (n=1) was not expanded further, and this buffer composition was selected for purification. The lysing procedure was performed with 1 mL spun culture; however, to avoid potential issues with error due to over-sonication, at least 100 mL shake-flask culture pellets should be retested. To improve the result accuracy, the repeat of this experiment should be analysed by loading equal amounts of quantified lysate to the SDS-PAGE gel and performing a western blot with a corresponding positive control loaded at a known concentration, for an accurate pixel density comparison between buffer compositions. These results further reinforce the difficulty in overexpressing soluble MBNL1-FL protein construct in *E. coli* BL21(DE3) expression strain.

Purification of N-term MBNL1 constructs

The proposed purification strategy differed from the original Park et al. (2017) method which did not include an affinity purification step. In their method the authors described initial cation-exchange chromatography followed by AEC to remove lysate contaminants from N-term MBNL1 constructs, before SEC polishing. Nonetheless, assumption of a His-tag located on the constructs, from the pET21a(+) backbone Addgene entry (https://www.addgene.org/vector-database/2549/) was the basis for adapting the purification strategy to the available laboratory capabilities. The issue with this approach here was the presence of non-specific protein in the elution gel images for MBNL1 Zn-F1-4 and ZnF1-2 during IMAC, with strong MBNL1 ZnF1-4 expression band presence in flow-through and wash fractions. Plasmid checks with the Wright Laboratory confirmed the 6xHis tag was not present in these constructs and observed binding is likely non-specific interaction of overexpressed MBNL1 construct

with the column resin. As this is an enrichment step, elution presence should concentrate the construct which would appear as the strongest protein band present, considering overexpression levels were high. This was also observed in preliminary bead-scale purifications with N-term MBNL1 constructs. The elution pool with MBNL1 ZnF1-4 contained measurable amounts of protein (0.1842 mg/mL) and was dialysed through spin column buffer exchange to remove NaCl and imidazole salts present in the elution buffer. The computed isoelectric points of N-term MBNL1 [78] constructs meant at buffer pH 8.0 these constructs would hold a net positive charge, and will bind to the strong quaternary ammonium anion exchange resin. This was also the reasoning behind eluting with a strong salt concentration (1M NaCl) step[114], however not enough protein was loaded and eluted for SEC separation. Due to time constraints, this purification process optimisation was discontinued to place focus on purifying MBNL1-FL.

In future attempts, the purification strategy design must be revisited to generate large quantities of protein for further structural biology studies. The authors performed an initial streptomycin sulfate cut to remove host-cell nucleic acids bound to the MBNL1 expressed constructs, followed by clarification by centrifugation and dialysis before AEC and SEC steps. The present study presumed this as gratuitous, as prior published work indicated that preferential binding of MBNL1 ZnF binding domains to YGCU(U/G)Y (Y equals a pyrimidine) pre-mRNA targets showing a higher preference for hairpin-structured single-stranded RNA targets [42, 115, 116], and further re-constitution with DM1 representative molecules (i.e., $CUG_{n=54}$, $CUG_{n=166}$) in a native-stoichiometry format, would preferentially bind MBNL1 ZnF-containing targets. Nevertheless, when revisiting this adapted purification strategy, an initial clarification and dialysis step could replace the first affinity chromatography step used in the work outlined. Ideally, the original paper purification workflow should be assessed in parallel with new propositions, such as following a 2-step only purification approach post-lysate clarification of Ion-Exchange chromatography and a size exclusion polishing step, or if final purity is not >95%, a double Ion-exchange chromatography process[76]. To verify the identity of MBNL1 constructs devoid of common molecular tags or markers, monoclonal antibodies should be sourced or generated that are specific to N-terminus region of MBNL1 and tested with HRP-conjugated secondary antibodies to validate a reproducible western blot antibody pair specific to N-term MBNL1 constructs. This requires time and resources which fell outside of the scope of this project and should be considered going forward. Final protein in a pure form (>95%) should be validated at the amino acid sequence level, possibly through approaches such as Tandem Mass Spectrometry (MS/MS)[117].

Purification of MBNL1 C-term, Full-length, and ZnF1-4 constructs:

The majority of purification process development in this project focused on the production of pure MBNL1-FL. The process was designed to apply to all three molecules where expression of correctly folded, soluble protein is present. The CiPP process rationale followed an orthogonal approach aiming to maximise target molecule purification in the first affinity step. Unfortunately, the described protein

solubility benefits of the molecular tags chosen for purification compatibility were not observed with MBNL1 constructs used in this study. The presence of GST-associated purification contaminants is observed in all Glutathione purification experiments, with comparable or even higher expression level to MBNL1-FL as seen in figure 6a, most likely resulting in protein translation events deviating from MBNL1-FL expression. Western blot detection with anti-GST antibodies and Mass Spectrometry analysis for amino acid sequence verification would be beneficial quality control additions for future expression scalability. Due to the homodimeric nature of GST[83], tag cleavage should enable MBNL1 constructs to remain monodisperse post-purification. This property may explain the presence of high MW expression bands of unknown identity outlined in figure 5 during MBNL1 C-term expression. The MBNL1-FL results show a low overall percentage recovery of target protein when quantified in pooled eluate fractions, a limiting factor for further downstream processing. Efforts to optimise the initial GST-tag affinity chromatography included re-loading the flow-through to maximise the binding of any MBNL1-FL, and introducing elution pauses to allow timely displacement of GST-tagged MBNL1-FL from the resin when applying a GSH elution buffer gradient, yielding moderate improvements shown in figure 6a. When reviewing this step, another beneficial approach would be to lower the flow rate of lysate loading, to as low as 0.1 mL/min instead of 0.5 mL/min, due to the slow binding kinetics of GST-tag protein with GSH resin[118].

The intermediate purification step (IMAC) was introduced here aimed to separate any free-cleaved GST tag and TeV protease, which contained a His-tag for purification[119]. This step relied on efficient cleavage of MBNL1-FL, and as observed in figure 6b, this was not always the case, possibly due to cut-site accessibility by the protease due to intrinsic MBNL1 construct folding properties. Further optimisation, for example of protease amounts added, time and temperature conditions, is required. Alternatively, re-design of the cut-site could include recognition sequences for other cheaper, and efficient proteases, such as Thrombin or Factor Xa[120]. A check of the MBNL1 construct sequences via PeptideCutter expassy tool[78] confirms these do not contain intrinsic cleavage sites for the mentioned proteases.

The final polishing SEC stage aimed at achieving sample purity of >95% by separation of proteins based on molecular size. In practice, this step should separate any contaminants and uncleaved MBNL1-FL, resulting in a clean, sharp elution peak at a retention time specific to only MBNL1-FL shape and size molecules. Unfortunately, SDS-PAGE gels did not show any detectable MBNL1-FL in fractions, to be expected from a low protein concentration loaded onto a separation technique that has a diluting effect. The three purification steps would greatly benefit from improvements in soluble MBNL1-FL construct expression, as described previously, to have enough starting material for reproducible process assessment to yield a >95% purity protein construct in solution. An additional step that could be introduced to assess purification success would be to introduce western blots specific to MBNL1 construct backbone, or the 6x His-C-tag for detection at each stage with N-term tag removal. Running native gels for protein elution fractions would also help assess the potential for GST-MBNL1 proteins in different oligomeric states.

Conclusions:

Expression of Wright Laboratory submitted MBNL1 Znf1-2, 3-4, and 1-4, showed good overexpression banding patterns and good protein solubility. Expression of MBNL1 ZnF1-4, C-term, and Full-length did not show obvious overexpression banding patterns in comparison when produced from a pGEX-6P-1 backbone plasmid in *E. coli* BL21 (DE3). Only MBNL1-FL was successful in giving evidence of overexpression suitable for downstream processing; however, retention of the protein within insoluble lysate fractions remains a large caveat in producing high yields. Optimising this step for better protein solubility could allow for a good yield improvement in the purification strategy proposed and enable pure MBNL1-FL protein to be re-constituted with DM1 RNA transcripts. The overall protein purification process would benefit from re-designing at the expression level, with a different affinity N-term tag, and other features described, which does not re-direct so much protein expression into free or dimerised tag. Evidence of the process being successful in increasing protein purity through each step is shown and should be re-attempted with redesigned expression constructs and *E. coli* protein expression under bioreactor conditions, or failing this, within a different host organism. The project would also benefit from specific, robust protein identification steps such as western blotting against MBNL1 constructs and molecular tags, and amino acid sequence determination by mass spectrometry of the final purified protein constructs.

Chapter 3: Generating DM1 *in-vitro* Trinucleotide Repeat RNA

Summary

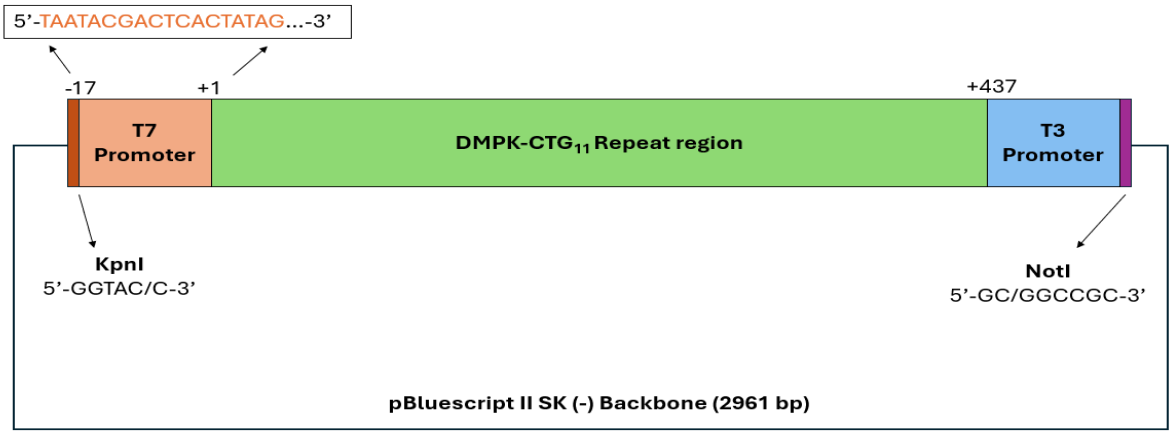
This chapter was aimed at generating CUG^{exp} RNA transcripts by *in vitro* transcription and analyse their yield and quality for downstream reconstitution into RNA-protein complexes akin to those observed in healthy (<39 repeats), Classical (50-150 repeats), and Juvenile, Congenital (DM1 >150 repeats) patients. $CUG_{n=11}$, $CUG_{n=54}$, $CUG_{n=160}$, and $CUG_{n=166}$ RNA transcripts were shortlisted to model representative DM1 mRNA species. For all RNA constructs generated, linear template preparations and *in-vitro* transcription conditions were systematically optimised to maximise RNA production.

3.1. Methods

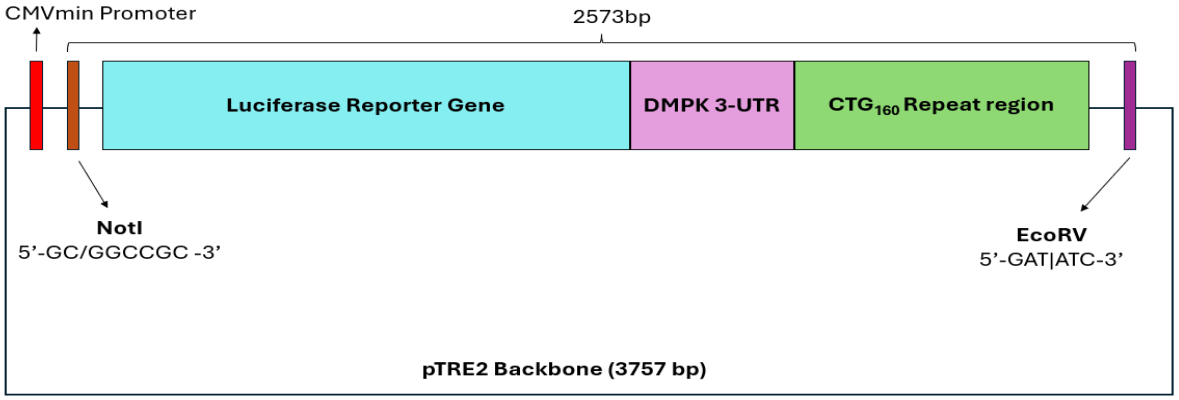
A. RNA template design, preparation for IVT, and validation

Plasmids containing $CTG_{n=11}$ and $CTG_{n=160}$ in the parent DMPK gene construct within pBluescript II SK (-) and pTRE2 backbone vectors, respectively, were made available from the David Brook laboratory, and these constructs were originally cloned directly from patient DNA into a bacterial vector. $CTG_{n=160}$ was similarly cloned in a mammalian expression vector that lacked a T7 promoter. $CTG_{n=11}$ is contained within a 437-nucleotide-long insert in conjunction with bases from the *DMPK* gene. Additional $CTG_{n=11}$, $CTG_{n=54,}$ and $CTG_{n=166}$ repeat sequences were commercially synthesised and cloned into pUC19 vectors for T7 promoter *in vitro* transcription of the trinucleotide sequences only, and RNA transcript termination on the final CTG repeat from a 3' EcoRI cut site.

a) pBSKII-DMPK-CTG$_{n=11}$



b) pTRE2-Luc-CTG$_{n=160}$



c) pUC19_CTG$_{n=11}$, pUC19_CTG$_{n=54}$, pUC19_CTG$_{n=166}$

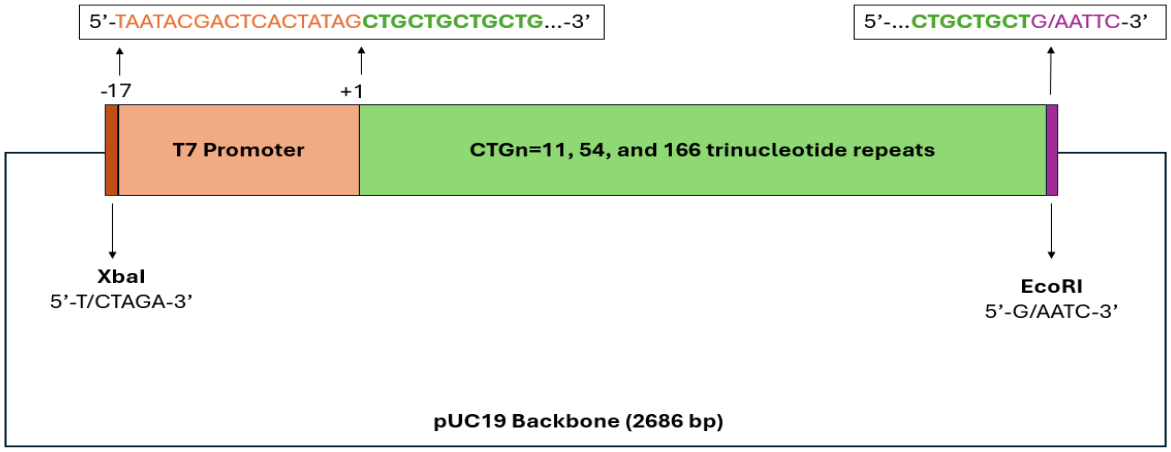


Figure 7) Illustrations showing cloned insert features for DM1 plasmid templates for RNA creation by *in-vitro* transcription, in their respective plasmid backbones.

Sanger Sequencing (Eurofins Genomics Europe Shared Services GmbH, DE) was performed for the David Brook laboratory donated plasmids (pBSKII-DMPK-$CTG_{n=11}$ and pTRE2 Luc-$CTG_{n=160}$ plasmids), with designed oligonucleotide sequencing primers listed in table 6 in appendix I. pBSKII-DMPK-$CTG_{n=11}$ and pTRE2 Luc-$CTG_{n=160}$ plasmids were checked by double-digestion of NotI® and EcoRV® (NEB LTD. Ltd, MA, USA) and SacII® and BbvCI® (NEB LTD. Ltd, MA, USA) respectively and visualised by TAE-agarose gel electrophoresis. All plasmids were cloned into Subcloning Efficiency™ DH5α Chemically Competent Cells (Invitrogen, MA, USA) for plasmid library expansion.

To prepare linear templates for IVT reactions, plasmid constructs were treated with single-restriction enzyme digestion, heat inactivation, and clean-up before adding to IVT reactions. pBSKII-DMPK-$CTG_{n=11}$ was cut with NotI-HF® and pUC19 backbone plasmids with EcoRI-HF® (NEB LTD. Ltd, MA, USA). Linearisation conditions were tested at 37°C for timepoints of 1 hr, 1.5 hrs, 4 hrs, and 16 hrs. All reactions contained QuickCIP® (NEB LTD. Ltd, MA, USA). Reactions were checked by Nanodrop 8000 and visualised by gel electrophoresis. Linear template concentrations that were too dilute for optimal kit-specific *in-vitro* transcription reactions were spin-concentrated with Vivaspin® 500 10K MW Centrifugal Concentrators (Sartorius, Göttingen, DE).

To circumvent a low starting template molarity in IVT reactions, and to generate large amounts of linear template, PCR reactions were set up for pUC19_$CTG_{n=11}$ and pUC19_$CTG_{n=54}$ plasmids, using primer pair sequences described in Table 7 in Appendix I. PCR was attempted to incorporate a T7 promoter site into plasmid pTRE2-Luc-$CTG_{n=160}$ for IVT kit compatibility, using primer pairs shown in Table 8 in Appendix I. All reactions were prepared with the Phusion® high-fidelity PCR kit (NEB LTD., MA, USA), including 0.5 µM Primer pair and 10 ng DNA template in 50 µL volumes.

Further PCR optimisation approaches were tried for T7 promoter site introduction with pTRE2-Luc-$CTG_{n=160}$ as follows: adding DNA template from gel-extracted SacII/BbsvCI-digested fragments (purified using PureLink™ Quick Gel Extraction Kit (Invitrogen, MA, USA)), adding 1-100 ng uncut plasmid template per reaction, varying primer concentrations at 0.2 µM and 0.05 µM, reducing annealing gradient (55-65 °C), 1X GC™ buffer and 3 % DMSO additions. Negative controls were set up at 58.0 °C or 55 °C. Amplification products were checked by 1 % TAE-agarose gel electrophoresis and quantified.

All PCR reactions were run in a Bioer Genetouch Thermal Cycler TC-E-96GA with the following cycle conditions shown below:

Table 4) PCR reaction conditions for DNA template amplification.

Step	Temperature	Time		
Initial Denaturation	95 °C	30 s		
30x Cycles	95 °C	30 s	Denaturation	
	60 °C	60 s	Annealing	Gradient from 58-68 °C
	68 °C	30 s	Extension	
Final Extension	68 °C	5 min		
Hold	4 °C	-		

A tHDA reaction strategy was tested for T7 promoter site incorporation into pTRE2-Luc-CTG$_{n=160}$ as an alternative method to PCR, using the IsoAmp® II Universal tHDA Kit (NEB Ltd., MA, USA). Oligonucleotide primer design followed kit recommendations, described in table 8 in appendix I. Two-step tHDA reactions were set up according to kit instructions: 25 µL of reaction mix 'A' tubes prepared with primer concentration at 75 nM, 100 nM and 150 nM, 0.4 ng DNA, and 25 µL of reaction mix tubes 'B'. Negative control tubes had no DNA added and positive control tubes contained 0.04 ng pCNG1 (+) plasmid DNA. Reaction mix 'A' tubes were heated to 95 °C for 2 mins, placed on ice, mixed 1:1 with 'B', and run at 65 °C for 1.5 hrs in a Bioer Genetouch Thermal Cycler TC-E-96GA with 70 °C heated lid. After 1.5 hrs, reactions were placed on ice. Samples were run on a 2 % TAE-agarose gel and quantified.

B. RNA In-Vitro Transcription

Three commercial IVT kits were selected to transcribe RNA: HiScribe® T7 High Yield RNA Synthesis Kit (NEB LTD. Ltd, MA, USA), MEGAshortscript™ T7 Transcription Kit (Thermo Fisher Scientific, MA, USA), and the T7 RiboMAX™ Express Large Scale RNA Production System (Promega Corporation, WI, USA). An In-house assembled IVT reaction was also tested as a feasible and scalable IVT alternative.

Table 5) Overview of *in vitro* transcription conditions tested for generating CUG^{exp} RNA transcripts.

RNA Transcript	Kit tested	DNA Template (nM)	Repeats	Reaction Runtime (Hrs)	Timepoint testing (Hrs)
DMPK-CUG$_{n=11}$	HiScribe® T7 High Yield RNA Synthesis Kit	25.22	2	20	3, 20
CUG$_{n=54}$	MEGAshortscript™ T7 Transcription Kit	40.31, 85.3	1	4	4

	T7 RiboMAX™ Express Large Scale RNA Production System	11.21,	3	26	0.5, 1, 2, 4, 6, 8, 20, 22, 24.5, 26
		7.83*	2		
		109.1, 100.93**,	1	4	0.5, 1, 2, 4
$CUG_{n=166}$	T7 RiboMAX™ Express Large Scale RNA Production System	14.22	3	20	0.5, 1, 2, 4, 20

***1.5 µL thermostable helicase Tte-UvrD was included in the reaction.*

***DNA template used was a $CTG_{n=54}$ PCR amplicon.*

$CUG_{n=54}$ RNA transcription was also attempted by custom IVT reaction consisting of 200 µL reactions containing: 40 mM HEPES-KOH, 28 mM $MgCl_2$, 5 mM NaCl, 20 mM DTT, 2 mM Spermidine, 2.5µL RiboLock™ RNase Inhibitor (40 U/µL), 50 ng/µL linear DNA template (reactions run for pUC19_$CTG_{n=54}$ and pTRI-RNA 18S Control DNA), 5 mM rNTP ACGU Mix (from HiScribe T7 High Yield RNA Synthesis Kit), and 0.025 mg/mL T7 RNA Polymerase (50,000 U/mL) (NEB LTD. Ltd, MA, USA), and nuclease-free H_2O. Reaction conditions were 37°C for 20 hrs, with RNA quantification at 4 hrs and 20 hrs time-points.

All IVT reactions were assembled as outlined by each kit-specific method in 20 µL volumes, with additions of DNA template in table 5 above. These were all finalised with 1 U/1 µg DNA TURBO DNase I at 37 °C for 15 mins. RNA was cleaned via Monarch® RNA Cleanup Kit (50 µg) (NEB LTD. Ltd, MA, USA), with RNA elution in 20 µL nuclease-free H_2O. Aliquots were taken for RNA quantification and analysis, and the remaining RNA was flash-frozen and stored at -80 °C. RNA was analysed by Agilent 4200 TapeStation system (Agilent Technologies Inc., CA, USA) automated electrophoresis, at the University of Nottingham Deep Seq facility (Nottingham, UK). Where indicated, transcripts were also checked via gel electrophoresis on Novex™ TBE-Urea Gels, 15% (Invitrogen, MA, USA). Samples were prepared by sample dilution in 2X Novex™ TBE-Urea Sample Buffer (Invitrogen, MA, USA), with heating to 95 °C for 3 mins before loading. Quantification of RNA transcripts generated during IVT, and post clean-up samples was performed by Qubit® RNA BR Assay Kit measured on Qubit® 4 Fluorometer (Invitrogen, MA, USA). All steps during RNA handling, outside of the IVT reaction and storage, and during RNA clean-up, were performed at 4 °C. All transcription and RNA processing steps were performed under RNAse-free conditions to avoid degradation.

3.2. Results:

The results shown here describe efforts to optimise the production of large quantities of heterogenous DM1 CUG^{exp} repeat RNA for future reconstitution with MBNL1 constructs. Initially, plasmid validation and DNA template preparation through restriction digestion and PCR amplification approaches enabled template compatibility with various commercial IVT kits used in this study. Next, RNA transcript production through IVT was attempted and optimised for each representative RNA transcript species. This included methods to assess the identity and purity of reaction products obtained.

Plasmid Characterisation

The sanger sequencing data obtained showed the presence of the $CTG_{n=11}$ repeat region within the pBSKII-DMPK-$CTG_{n=11}$ plasmid, confirming the presence and identity of the *DMPK* gene flanking sequence following a NCBI BLAST alignment check, which returned a high sequence identity similarity to Homo sapiens protein kinase mRNA, partial cds (Sequence ID: M94203.1). Multiple file sequence alignment from sequencing of pTRE2-Luc-$CTG_{n=160}$ plasmids indicated the presence of repeat CTG regions, preceded by the DMPK 3'-UTR gene region.

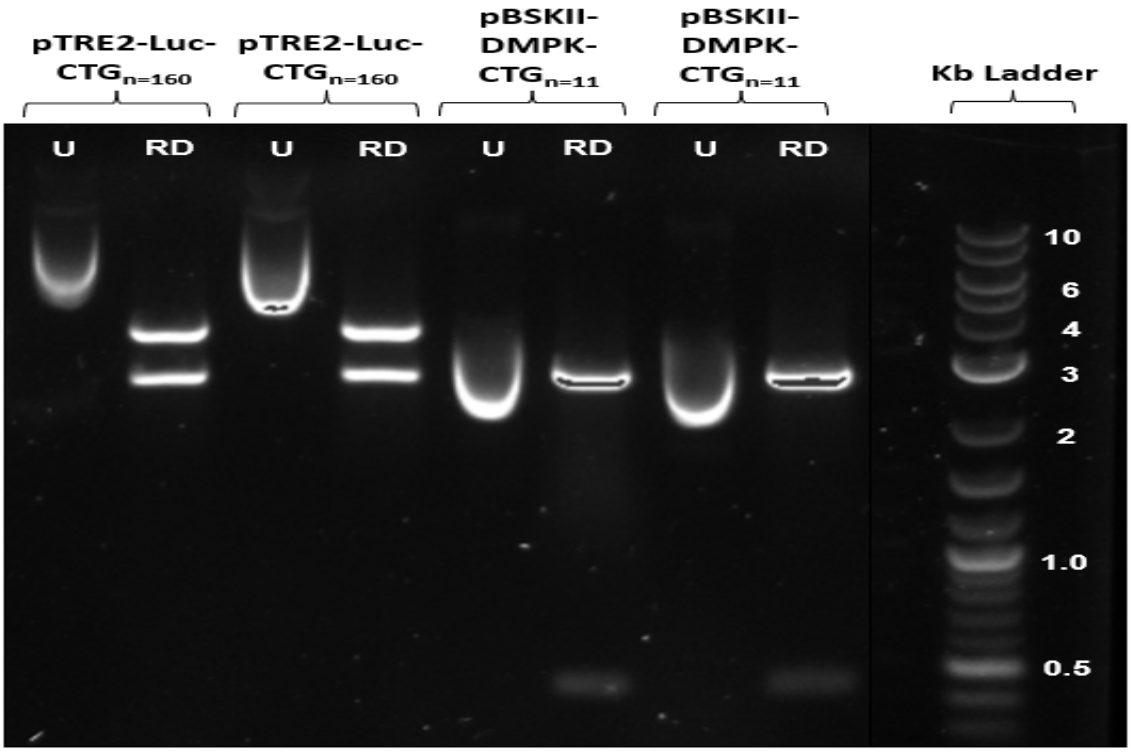


Figure 8) 1% TAE-agarose image showing restriction digestion of pTRE2-Luc-$CTG_{n=160}$ and pBSKII-DMPK-$CTG_{n=11}$ plasmids, in uncut (U) and cut (RD) form (n=2).

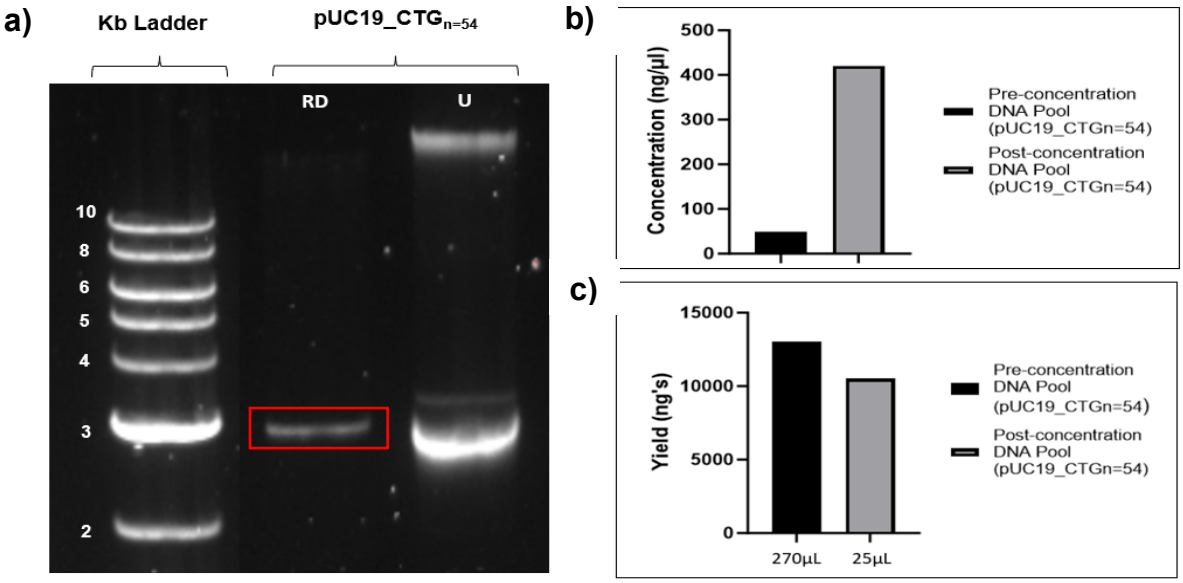
Double restriction enzyme digested validated plasmids displayed above in figure 8 showed the presence of expected banding sizes relative to uncut plasmid conformations: pTRE2-Luc-$CTG_{n=160}$ digested as 2562 bp and 3730 bp; pBSKII-DMPK-$CTG_{n=11}$ digested as 2961 bp and 443 bp fragments. Uncut plasmid migration follows a typical pattern observed in supercoiled plasmid DNA conformation relative to linear DNA fragment forms of similar sizes when run through an agarose gel. The uncut supercoiled plasmid, although heavier, is packed more tightly than similarly sized linear fragments, moving through the gel matrix faster.

The combined results confirmed correct plasmid identity. The supplier data for synthesized pUC19_$CTG_{n=11}$, pUC19_$CTG_{n=54}$, and pUC19_$CTG_{n=166}$ confirmed plasmid identity, and the data was reliable for validation purposes.

Template preparation for *In-vitro* Transcription reactions

Plasmid linearisation and concentrating

Plasmid linearisation by single-cut restriction enzyme digestion was key for correct transcription results using the T7 in-vitro transcription kits throughout this study. For each kit, transcription is initiated at nucleotide position +1 after the T7 promoter sequence and finalises at the 3’ end of the linear DNA template. It was important to test linearisation conditions to ensure optimal enzyme-cutting, resulting in a sample of completely linear DNA without uncut plasmid, which would lead to long, non-specific RNA transcripts containing plasmid elements beyond the desired multi-cloning site features in the reaction.



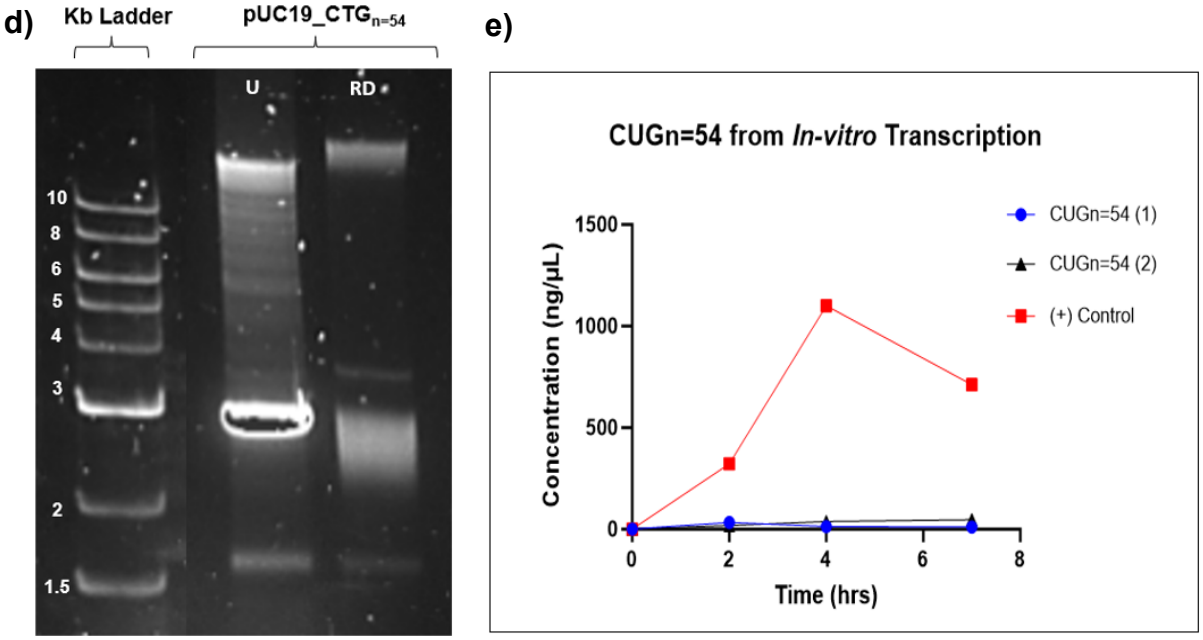


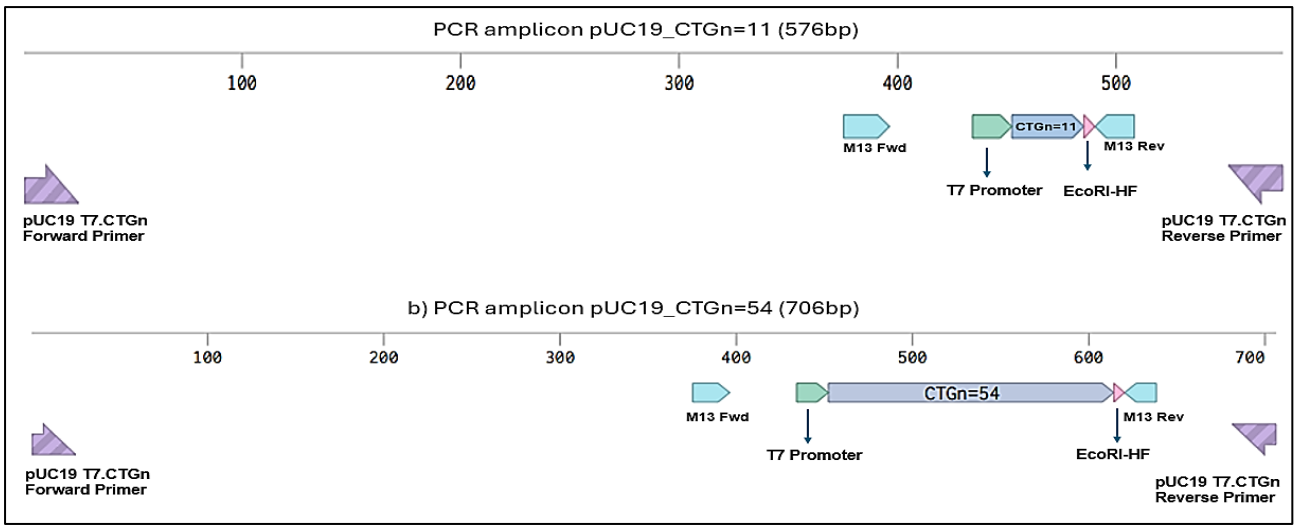
Figure 9) a) 1% TAE-agarose gel image showing complete cutting (red box) of template 10 x 1 ug DNA plasmid reactions run for 1 hr at 37 °C for IVT (n=10). **b-c)** Concentration values and yield results of 270 µL pooled, alkaline phosphatase-treated linear plasmid template before and after concentrating to 25 µL via Vivaspin® 500 10K MW Centrifugal Concentrator. **d)** Incomplete cutting of template in a 1x 50 µL 10 ug DNA plasmid reaction run for 4 hrs at 37 °C for IVT. **e)** Graphical representation of concentration (ng/µL) values for each $CUG_{n=54}$ RNA replicate MEGAshortscript™ T7 Transcription Kit reaction against the kit-provided (+) RNA control using poorly cut template from (d).

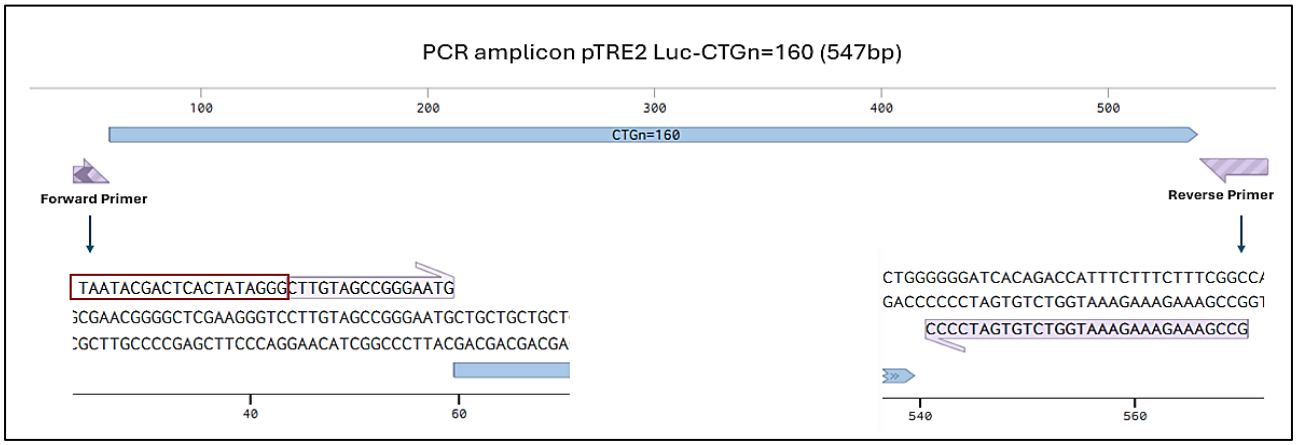
The above figure 9a shows good digestion, where the expected single band fragment size for the 2877 bp pUC19_$CTG_{n=54}$ is at the correct position. The absence of other plasmid forms confirms a complete restriction digestion result taken to IVT. The multiple bands observed in the uncut lane are typical of plasmid gel images due to the supercoiled (fastest), relaxed, and open circular (both slower) conformations of plasmid migrating through the gel. The effect of heterogenous DNA template is exemplified in figure 9 (d-e), which shows a poor IVT yield when compared to the kit-provided positive linear control plasmid. Here, the concentration of DNA per IVT reactions following the 4-hr restriction digest was 61.68 nM, and the yield should have been higher than the experimental values observed. However, as the template digestion observed was so poor, showing possible star activity reflected in considerable smearing present, the amount of target linear DNA template would have been reduced drastically. Template preparation testing experiments showed the best conditions for template preparation derived from individual 1 ug DNA cuts per reaction, not exceeding reaction volumes of 25 µL for 1 hr, which were then pooled and concentrated for IVT.

IVT transcription reactions require varying degrees of DNA linearised template to reach optimal concentration requirements for a successful IVT reaction. Shorter transcription product reactions posed the most challenges by requiring highly concentrated linearised template material, due to intrinsic enzyme kinetics of T7 polymerase action when initiating and completing transcription events. This was a crucial step when combining multiple 25 µL restriction enzyme reactions, which consistently resulted in a diluted DNA template sample below supplier recommended molarity values. The MEGAshortscript™ T7 Transcription Kit, for example, required a concentration of ideally 125 nM DNA per reaction when tried with our transcript of interest product $CUG_{n=11}$ from pUC19_$CTG_{n=11,}$ and many failed attempts with it could be attributed to low DNA template molarity. However, use of centrifugal concentrators overcame this hurdle, as exemplified in figure 9b-c, with minimal template loss. Introducing this step now allowed for optimal template concentration addition to IVT reactions.

Amplification Strategies for IVT Template Preparation

Another strategy tested for increasing already linearised IVT template availability was to amplify T7 promoter and CTG repeat regions within plasmids by conventional PCR. Scaling up PCR reactions would provide a more rapid, resourceful alternative method to re-synthesising or extracting plasmids from bacterial cultures. In the case of pTRE2-Luc-$CTG_{n=160}$, PCR and thermostable helicase-dependent amplification (tHDA) were tried as a means of introducing a T7 promoter upstream of the $CTG_{n=160}$ repeats region, to make the template DNA IVT kit-compatible.





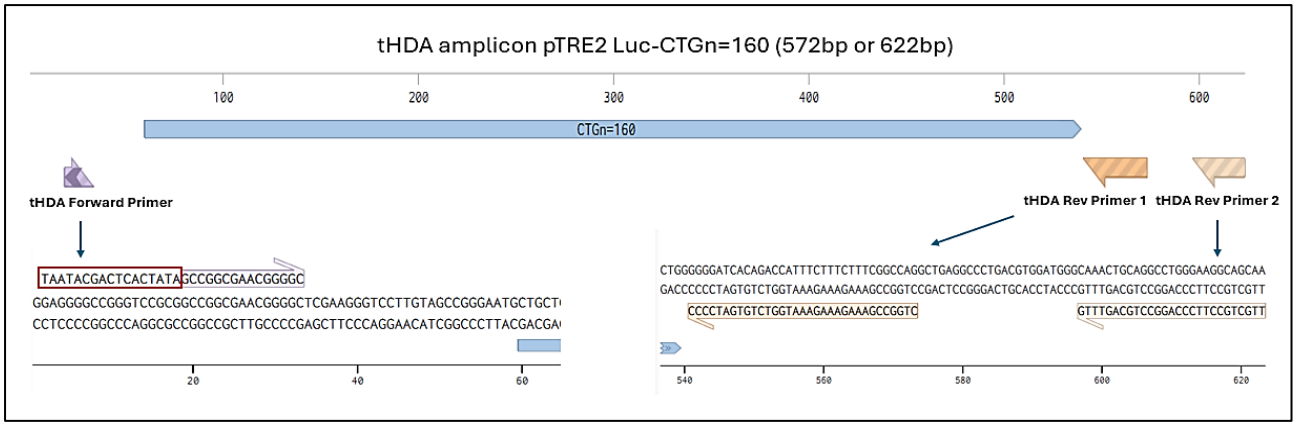


Figure 10) Expected PCR and tHDA amplicon products and their features for use in IVT reactions with pUC19_CTG$_{n=11}$, pUC19_CTG$_{n=54}$, and pTRE2-Luc-CTG$_{n=160}$.

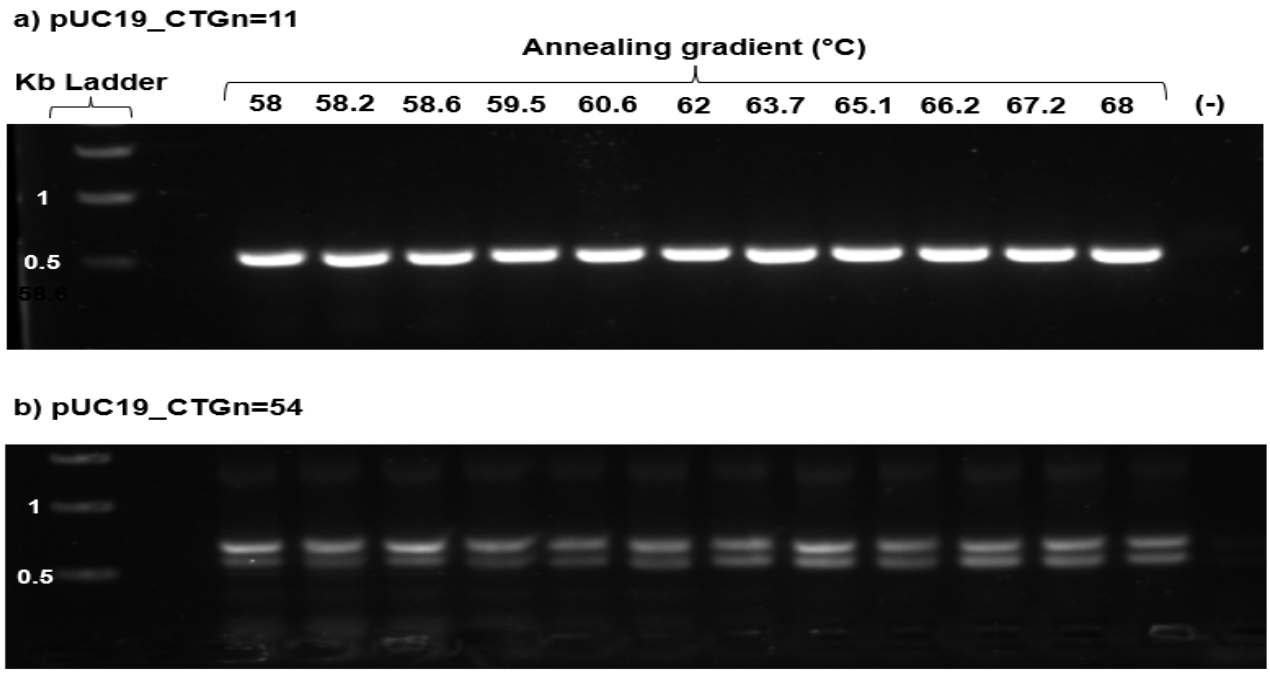


Figure 11) a) PCR annealing gradient reactions with 10 ng pUC19_CTG$_{n=11}$ and **b)** pUC19_CTG$_{n=54}$ template using identical NEB Phusion Kit components, primer pairs and cycle conditions (n=1).

PCR results from amplifying pUC19_CTG$_{n=11}$ and pUC19_CTG$_{n=54}$ show good amplification in figure 11a with pUC19_CTG$_{n=11}$, at the correct predicted fragment size of 576 bp, with no (-ve) bands present in the reaction. However, the pUC19_CTG$_{n=54}$ plasmid amplicons in figure 11b showed a consistent double-banding pattern, under the same kit conditions and primer pair. The top band roughly corresponds to the correct product size of 706 bp. As both templates used the same reaction components, primers, and running conditions, the observed result could be due to DNA secondary structure bias interfering with polymerase action. The PCR product verification by sequencing methods was not performed in this work, however, as part of the verification process, these amplicon products were taken to IVT stage to check if expected RNA transcripts could be generated.

Before receiving the pUC19_CTG$_{n=166}$ plasmid, many attempts were made to introduce a T7 promoter site into plasmid pTRE2 Luc-CTG$_{n=160}$, aimed at making this construct process compatible with RNA IVT kits used in this study, as the backbone plasmid contains a CMVmin mammalian expression promoter, and PCR provides a quicker alternative for plasmid adjustment compared to conventional clonal, if successful. Therefore, efforts focused on introducing the T7 promoter site into pTRE2 Luc-CTG$_{n=160}$ plasmid through conventional PCR approaches (predicted products resulting from primer design are illustrated in figure 10).

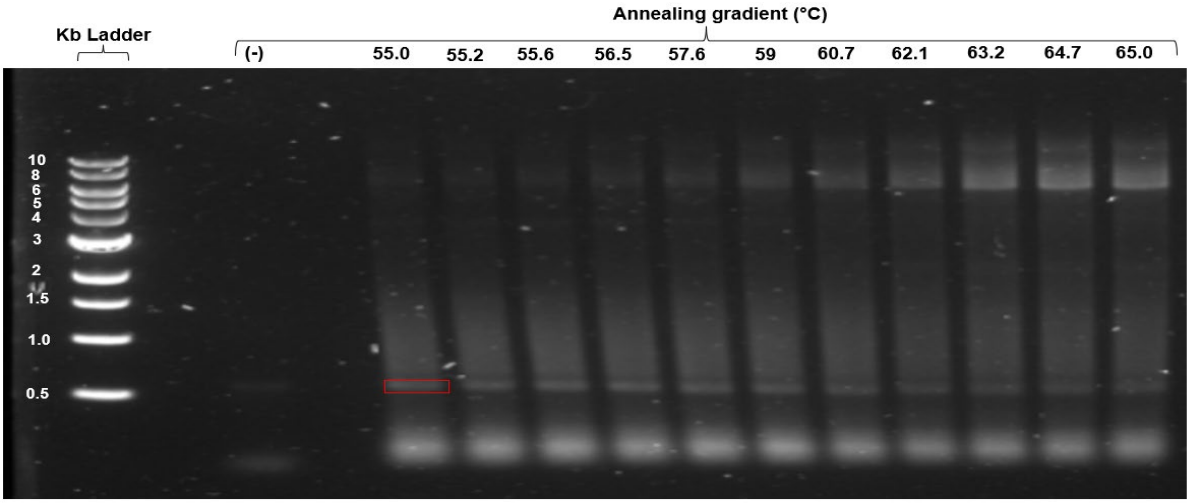


Figure 12) 1% TAE-agarose image of a PCR reaction result for introducing a T7 promoter site into pTRE2-Luc-CTG$_{n=160}$ plasmid backbone using the Phusion® high-fidelity PCR kit, with primer pair described in the method section, and 10ng of pTRE2-Luc-CTGn=160 with reduced anneal gradient from 55-65°C, and additions of GC™ buffer and 3% DMSO per reaction. (n=1).

Figure 12 above was selected to exemplify the banding pattern consistently seen across all gels and optimisation attempts to introduce a T7 promoter site into the PCR amplicon. The red box indicates the correct approximate band size (547bp) seen in a slightly stronger, distinct band signal amongst the background smearing observed. The PCR optimisation strategies attempted here were selected to reduce non-specific band formation as recommended by the Phusion kit manual literature. All PCR optimisation approaches without modifying the proposed primer sequences were unsuccessful, unlike the discrete PCR amplicon products observed in pUC19_CTG$_{n=11}$ PCR figure 11a results. Another alternative step in the PCR optimisation plan could be to design a different T7 primer set, potentially including some 3' DMPK-UTR sequence found in the pTRE2 Luc-CTG$_{n=160}$ as it is DM1-specific RNA, which could also include sequencing primer sites to allow for PCR product validation.

The possibility of a high degree of secondary structure DNA formation from CTG-rich areas led to the decision to test a thermostable Helicase-Dependent Amplification (tHDA) kit approach to introduce the T7 promoter site into the existing pTRE2-Luc-CTG$_{n=160}$ plasmid with a new set of primers flanking the CTG repeat region. The tHDA reaction utilises a thermostable helicase Tte-UvrD that will unwind DNA in the presence of single-strand DNA binding protein (SSB), enabling designed primer annealing at a set reaction temperature of 65°C [121].

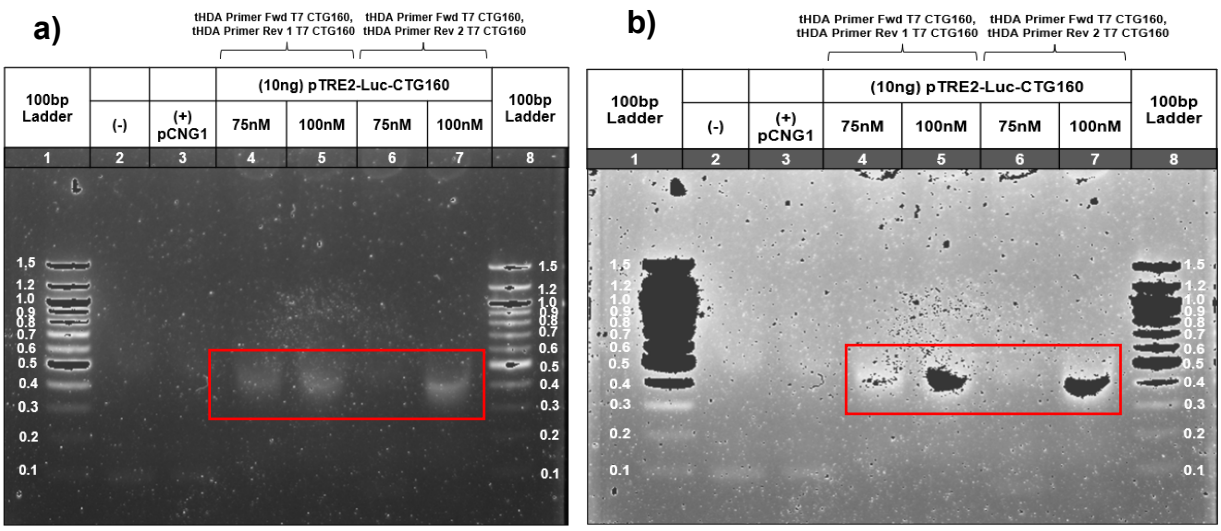


Figure 13) a-b) 2% TAE-agarose gel image of tHDA reactions with 10ng pTRE2-Luc-CTG160 with both sets of primer pairs at concentrations of 75nM and 100nM, including no DNA (-ve) and pCNG1 (+ve) controls with primer pairs at 75nM of kit provided Forward/ Reverse primer. Reactions shown were n=1.

The tHDA reaction results shown in Figure 13 show lower non-specific DNA fragments compared to previous PCR strategies. However, the amplicon band sizes, highlighted with red boxes, did not correspond to the expected band sizes of 573 bp and 623 bp for both primer sets. In this gel image, the (+ve) control band is not seen; however, a previous attempt under the same conditions yielded a correct (+ve) band size product, which could indicate an experimental error in the presented tHDA attempt. To further increase the yield of amplicon and investigate this product by IVT, the tests were optimised by varying primer concentrations, as shown in figure 13. ImageJ analysis and normalisation against the known 45 ng 400 bp fragment from the 100 bp NEB ladder quantified tHDA product concentrations of ~6 ng/µL, too low for IVT reactions or for sequencing analysis. With the arrival of the pUC19 $CTG_{n=160}$ plasmid, which already included a T7 promoter sequence downstream of CTG repeats, further attempts to introduce a T7 promoter site into pTRE2-Luc-$CTG_{n=160}$ with these approaches were discontinued.

DM1 trinucleotide repeat RNA transcription results

The following section outlines IVT transcript creation results for DM1 RNA CUG^{exp} repeat regions, using the successfully linearised DNA templates already described.

DMPK-CUG$_{n=11}$ RNA transcription:

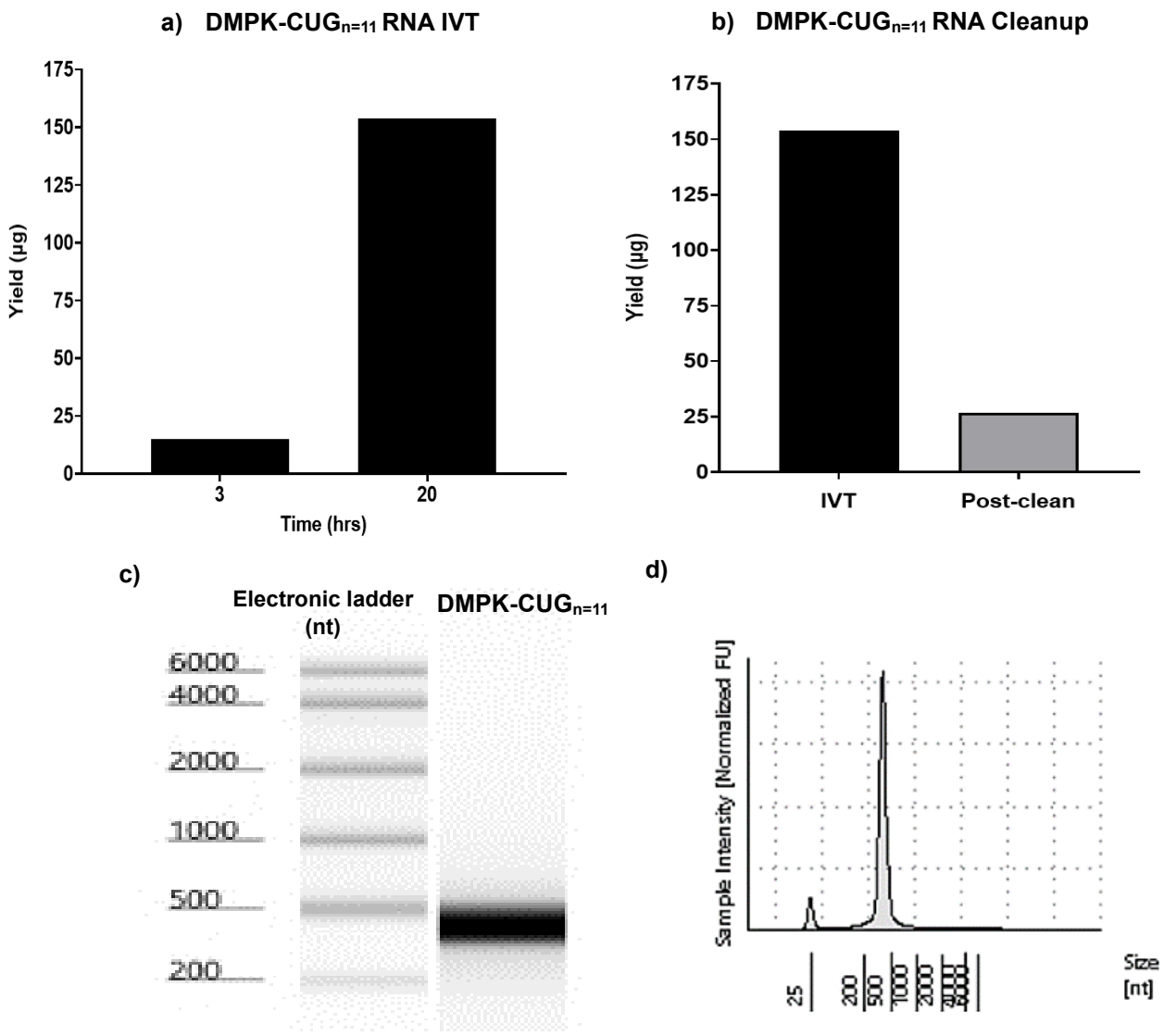


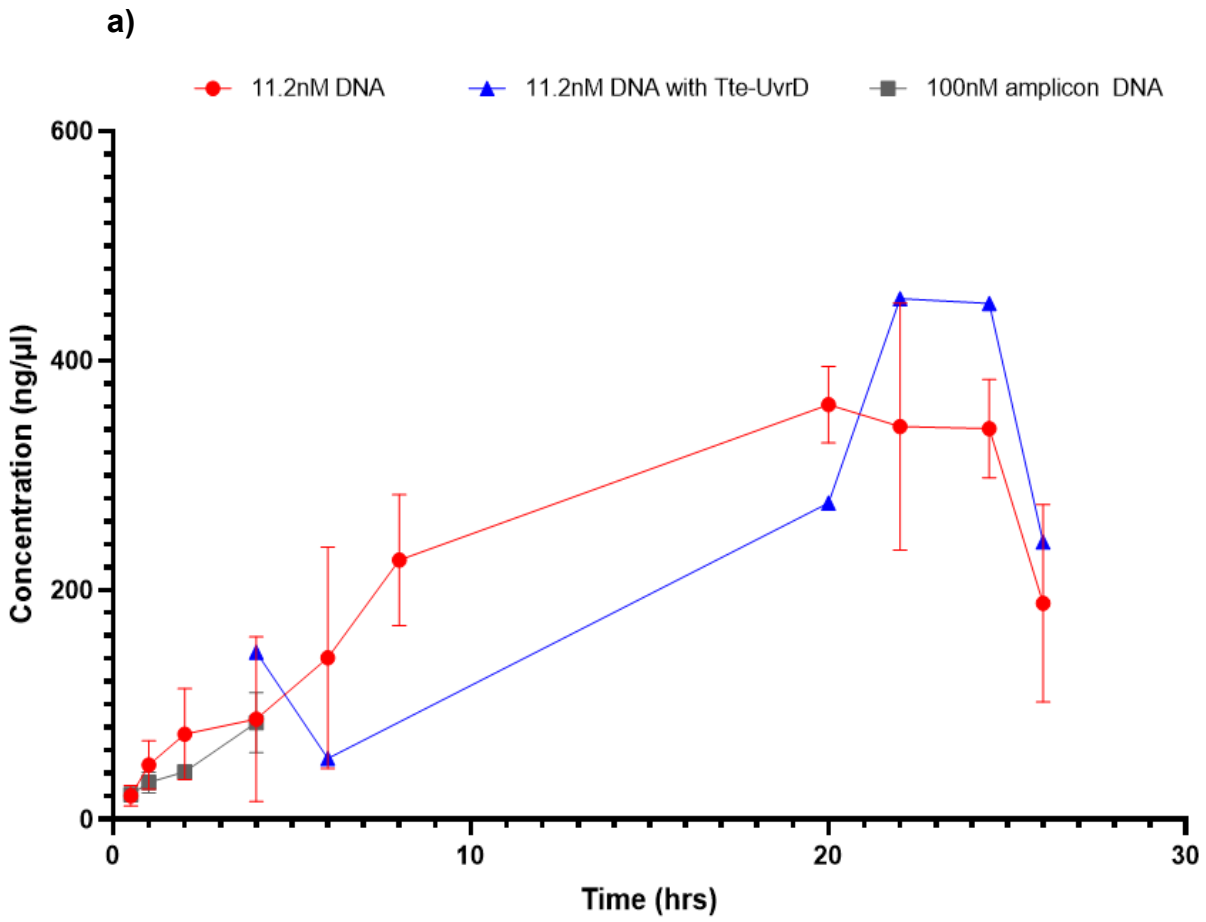
Figure 14) a) DMPK-CUG$_{n=11}$ RNA concentration values (n=2) during IVT time points (3 hrs/20 hrs) of linearised pBSKII-DMPK-CTG$_{n=11}$ with HiScribe® T7 High Yield RNA Synthesis Kit, measured by Qubit RNA BR assay. **b)** IVT total ng's yield comparison after Monarch® RNA Cleanup. **c)** Agilent 4200 TapeStation system banding pattern for clean DMPK-CUG$_{n=11}$ RNA compared to electronic ladder band assignment, and **d)** Relative banding intensity quantification via sample intensity normalised fluorescent units (FU).

Figure 14 shows a successful DMPK-CUG$_{n=11}$ RNA creation, at the expected product size of ~437 bp, with a lack of non-specific banding shown in Agilent 4200 TapeStation system assay analysis (c-d). The technique analyses RNA by migrating transcripts through a pre-cast gel matrix in a microfluidics system, intercalating fluorescent dye, and detects RNA fragments optically using a laser-induced fluorescence detector. The above DMPK-CUG$_{n=11}$ RNA was stored for further validation and usage. Validation

would require sequencing with designed primers, which has not been performed here yet. pUC19_CTG$_{n=11}$ PCR amplicon products were restriction digested and tried in IVT reactions at optimal molarity template values (100 nM), however, these reactions yielded low concentrations of RNA transcripts, and efforts here were discontinued. The next stage of DM1 RNA creation was focused on producing CUG$_{n=54}$, and CUG$_{n=166}$ repeat RNA due to time and resource constraints, to complete a representative DM1 CUG^{exp} repeat set.

CUG$_{n=54}$ RNA Transcription

This section describes results obtained for CUG$_{n=54}$ RNA creation.



b)

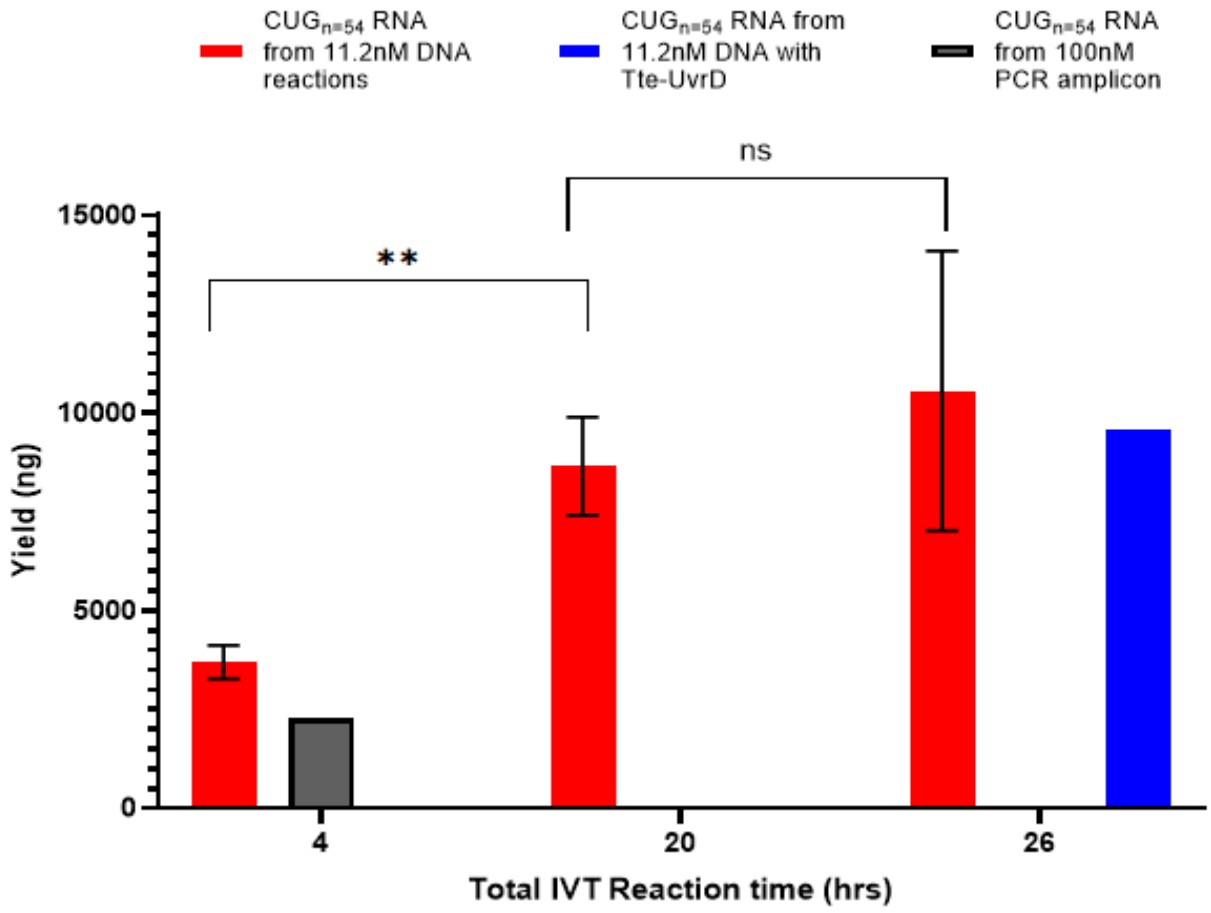


Figure 15) a) CUG$_{n=54}$ IVT reaction quantification by Qubit 4 BR RNA Assay, with T7 RiboMAX™ Express kit for 11.21 nM template DNA plasmid pUC19_CTG$_{n=54}$ alone (red), with helicase Tte-UvrD addition (blue), and 100 nM PCR Amplicon DNA Template (grey). Reaction results compiled from reactions run until 4 hrs, 20 hrs, 26 hrs for pUC19_CTG$_{n=54}$ showing mean ± SD error bars (n=3), 26 hrs (n=2) with helicase Tte-UvrD, and 4 hrs (n=2) PCR Amplicon Template. All quantification time points taken at 0.5, 1, 2, 4, 6, 8, 20, 22, 24.5, and 26 hrs. **b)** CUG$_{n=54}$ Pure RNA yields quantified following Monarch® RNA clean-up kit procedure for each reaction end timepoint in '**a**', post DNase treatment and elution.

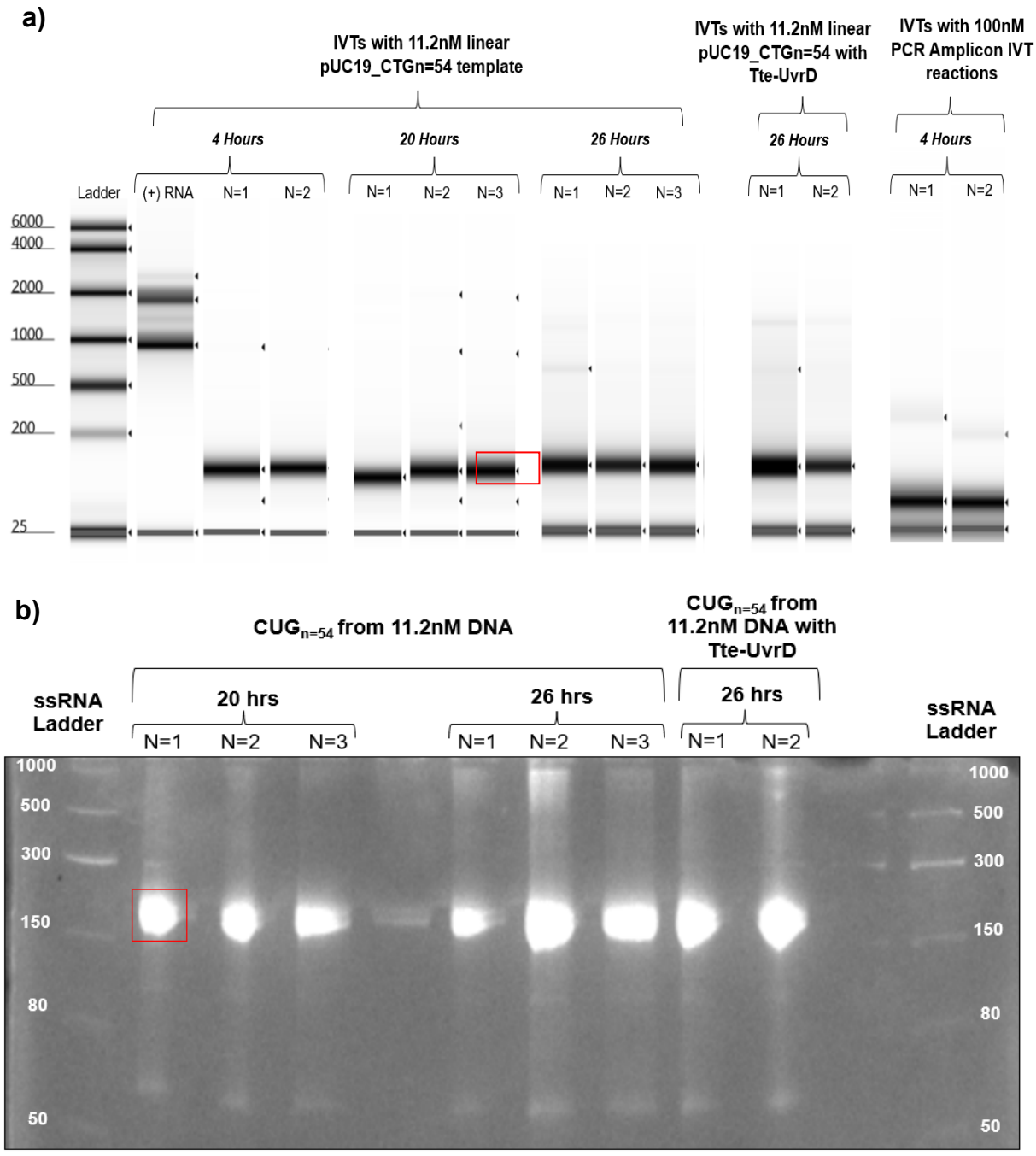


Figure 16) a) Agilent 4200 Tapestation trace for purified RNA from reactions run until 4 hrs (n=3), 20 hrs (n=3), 26 hrs (n=3) for pUC19_CTG$_{n=54}$ DNA, 26 hrs (n=2) with helicase Tte-UvrD, and 4 hrs (n=2) PCR Amplicon Template. Banding sizing was relative to peak assignment against the electronic ladder (nt). One (+) RNA sample banding pattern is shown; however, each set of results also included this control. **b)** 15% TBE-Urea Gel image of CUG$_{n=54}$ samples mentioned above, without 100nM PCR amplicon RNA transcript. The red box in a/b refers to the desired band size for CUG$_{n=54}$ at 162nt.

Figure 15 (a) shows $CUG_{n=54}$ RNA created from 11.2nM pUC19_$CTG_{n=54}$ linearised template without Tte-UvrD addition using the T7 RiboMAX™ Express kit, produced the highest yields at 20 hrs, and was not significantly different compared to the 26 hrs mark (paired t-test: $t = 2.682$, $df = 2$, $p = 0.1154$), and significantly higher than the 4-hour mark values (paired t-test: $t = 24.52$, $df = 2$, $p = 0.0017$) (Figure 15b). This is likely a result of reaction conditions plateauing due to reaction substrate depletion. Inclusion of the thermostable helicase Tte-UvrD into IVT reactions did not show an obvious improvement in banding pattern, including non-specificity, as observed in figure 16 Tapestation banding reports, verified against TBE-urea gel results, suggesting this addition is not necessary here. As the reaction with helicase Tte-UvrD inclusion was run in n=2, concentration values could not be determined as significantly better when including the enzyme.

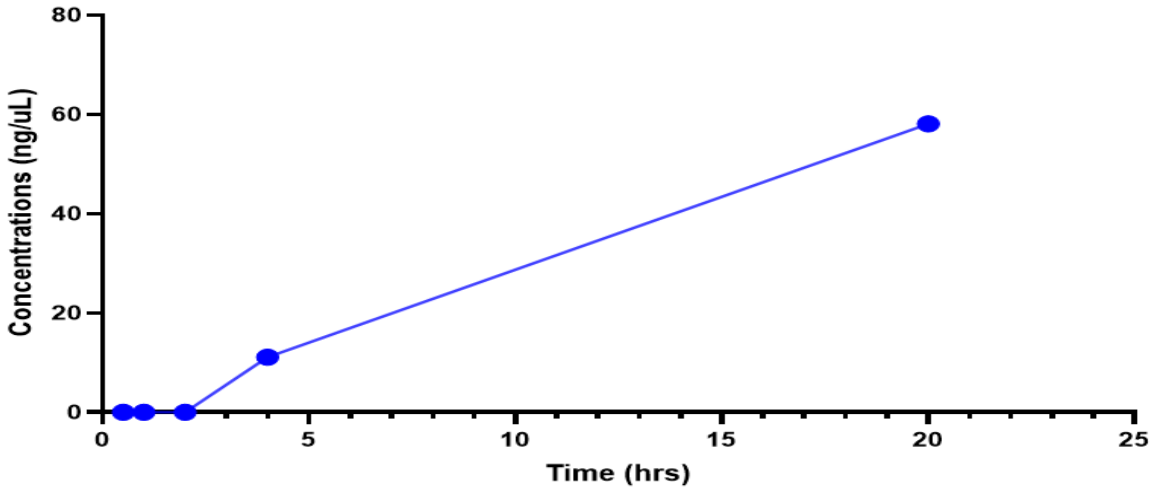
IVT products of PCR amplicons showed a lower band of interest size and considerably more background banding seen as the darker degradation pattern and are therefore not ideal for this RNA creation process. Agilent Tapestation nucleotide band size assignment consistently returned band of interest values ranging from 129-133nt long. To test if this is a result of the native conditions under which the Tapestation analysis is performed and considering that n=54 repeat with likely have a stable secondary structure at these conditions, we performed a denaturing TBE-urea gel electrophoresis (figure 16 (b)). Here, the band sat just above the 150nt ladder mark, as would be expected for a 162nt CUG RNA sequence. Thus, the incorporation of strongly denaturing TBE-urea gel electrophoresis to separate triplicate RNA repeat band fragments, negating secondary structure effects, complemented the Agilent TapeStation method for visualising banding patterns in IVT reactions.

Attempts to produce $CUG_{n=54}$ RNA with the MEGAshortscript™ T7 Transcription Kit were also tested at optimal DNA template concentrations per reaction, yielding poor results with considerable non-specific banding presence in comparison to the results presented above.

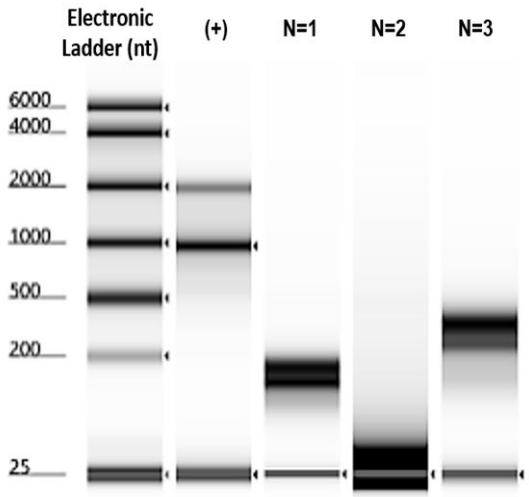
$CUG_{n=166}$ RNA transcription

The arrival of pUC19_$CTG_{n=166}$ towards the end of the project allowed for a triplicate repeat IVT reaction with this plasmid construct with the T7 RiboMAX™ Express Large Scale RNA Production System due to its success with $CUG_{n=54}$ RNA transcription.

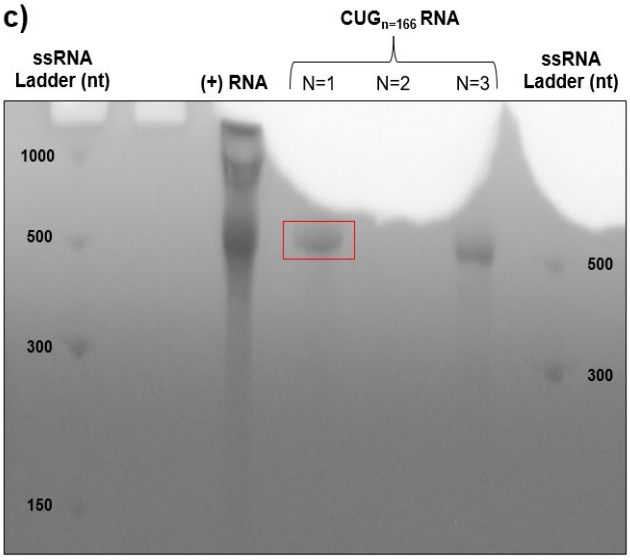
a)


Concentrations (ng/uL)
80
60
40
20
0
0
5
10
15
20
25
Time (hrs)

b)


Electronic Ladder (nt)
(+)
N=1
N=2
N=3
6000
4000
2000
1000
500
200
25

c)


ssRNA Ladder (nt)
(+) RNA
CUG n=166 RNA
N=1
N=2
N=3
ssRNA Ladder (nt)
1000
500
300
150
500
300

d)

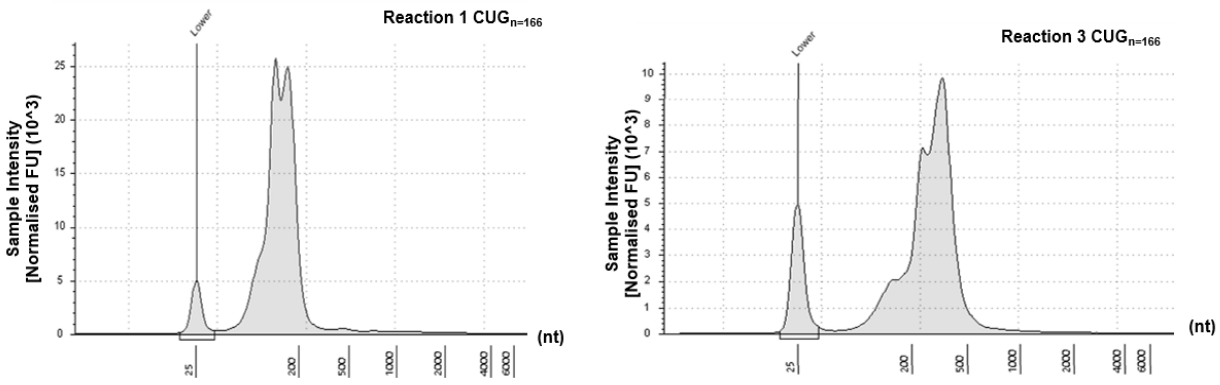


Figure 17) a) IVT results (n=3) for linear pUC19_CTG$_{n=166}$ (61.42 nM) with T7 RiboMAX™ Express Large Scale RNA Production System, producing CUG$_{n=166}$ at 30 mins, 1 hr, 2 hrs, 4 hrs, 20 hrs reaction timepoints. **b)** Agilent 4200 Tapestation System automated electrophoresis trace of submitted CUG$_{n=166}$ repeat reactions 1-3. Red box shown indicates location of expected 498nt band for CUG$_{n=166}$. **c)** 15% TBE-Urea Gel image of replicate CUG$_{n=166}$ RNA against Low-range ssRNA and (+) control transcript RNA. The red boxes indicate the desired band size of CUG$_{n=166}$ at ~498nt. **d)** Agilent 4200 Tapestation report showing sample intensity FU (fluorescent units - normalised) for each CUG$_{n=166}$ replicate, separated by transcript length against electronic ladder band location.

Figure 17 shows CUG$_{n=166}$ transcription success in replicates 1 and 3, although at least two distinct species peaks are forming according to the Tapestation trace. The yield is considerably lower than the expected result from (+) control plasmid, a common result observed throughout all IVT kit reactions tested during this project, likely due to the larger size of positive control reaction product. Sampling concentrations were also considerably lower than those observed in figure 15a, potentially from CUG$_{n=160}$ secondary structure formation impacting ribosomal speed and efficiency when compared to shorter CUG$_{n=54}$ transcript products and explaining split peaks from premature termination driven by structure constraints. Template dsDNA structure could potentially hinder transcription efficiency here. Cleaned RNA yields were 825 ng and 2180 ng total for replicates 1 and 3, highlighting the variability in expected RNA yield with this reaction format. Band sizing expected of 498nt can be observed best under the strong denaturing conditions of the 15% TBE-Urea Gel image, when aligned to the ssRNA ladder.

The process followed for DM1 RNA creation, applicable to all constructs described, relied on plasmid design, correct template preparation, successful IVT RNA transcript creation, RNA clean-up, and quality control validation. From the results presented, this aim has been achieved for DMPK-CUG$_{n=11}$, CUG$_{n=54}$, and CUG$_{n=166}$ RNA transcripts.

3.3. Discussion

DM1 RNA transcript creation*:*

The results shown in this study show a capacity to make high amounts of DM1-representative RNA molecules for future structural biology studies. The best yields were achieved at longer reaction incubation times of ~20 hrs, statistically confirmed with $CUG_{n=54}$ creation, but also apparent with $CUG_{n=166}$ (figures 15-17). Between 20-26 hrs, there was no significant difference in terms of an increase in yield for producing $CUG_{n=54}$ (figure 15), and a general decrease in trend of RNA amounts, possibly due to RNA shearing effects and rNTP and cofactor substrate depletion[122].

Increasing yield to increase RNA obtained for studies requiring considerable biological material, e.g. for crystal formation in X-ray Crystallography[123], was investigated by aiming for the 1 µg "sweet spot" for kit performance as directed by manufacturers. The difficulty in obtaining projected yields in every IVT reaction was mainly due to the dilution of template plasmid DNA during restriction digestion reactions. This, however, was circumvented by using MW separator spin columns presented in figure 9(b-c) with linear pUC19_$CTG_{n=54}$ plasmid, with minimal loss of linear DNA template material. When attempting to produce short RNA transcripts, such as $CUG_{n=11}$ and $CUG_{n=54}$ with the MEGAShortscript T7 Transcription kit, this problem is amplified due to the template requirement to be generally between 100-125 nM per reaction, which for a 3 kb dsDNA linear plasmid in a 20 µL volume reaction equates to 4.87 µg, a high template requirement concentration considering each reaction generally allows volume addition of <6 µL of DNA. This was the reason for testing the PCR attempts designed to amplify regions containing the 5'-T7 promoter, CTG repeats, and 86 bases past the 3' restriction digestion cut-site only, which means restriction digestion would still need to be performed, as designed PCR oligos could not be made to cover only until 3' EcoRI cut-site. The PCR amplicon product created required considerably less DNA ug's per IVT reaction than intact linear plasmid template to hit the 100-125 nM "sweet spot". This PCR approach for template preparation worked very well with pUC19_$CTG_{n=11}$ as seen in figure 11a, however with pUC19_$CTG_{n=54}$ plasmids the PCR amplicons generated a consistent double-band pattern shown in figure 11b, possibly indicating problems relating to dsDNA secondary structure with repeat regions. This is reinforced when comparing UNAfold tool[124] predictions for nucleic acid folding of $CTG_{n=54}$ and $CTG_{n=11}$ regions, where the former has a much stronger propensity to form spontaneous secondary structures (-ΔG). When the $CTG_{n=54}$ PCR amplicon product was run in IVT reactions at 100 nM template concentration, banding pattern was of poor quality and incorrect sizing in comparison to the linear intact template DNA plasmid seen in figure 16a, likely partly due to heterogeneity in PCR amplicon species in the template DNA added to the IVT reaction, and this approach was therefore inadequate following multiple rounds of control kit failures.

$CUG_{n=54}$ and $CUG_{n=166}$ RNA transcription gave much better results with the T7 RiboMAX™ Express Large Scale RNA Production System kit, however at same time-points and similar molarity within each IVT reaction, concentration values at 20 hr timepoints were ~6-8 fold less with $CUG_{n=166}$ RNA transcripts (Figure 17). There could

be issues of rNTP usage depletion, and it would be good to test increased rNTP additions for bases rC, rU and rG. Addition of thermostable helicase Tte-UvrD to IVT reaction to try to negate any possible effects of DNA secondary structure did not provide better specificity or yield improvement banding patterns, and the TapeStation traces and RNA gel results were identical in profile. With $CUG_{n=54}$ RNA, this addition seems to have shown an extension of RNA concentration value plateau past the 20-hr mark, but more identical IVT repeats would be needed to statistically assess reactions with and without this addition.

Optimisation strategies to improve IVT reactions:

Optimisation strategies to improve yield and quality, through reduction of erroneous non-specific RNA product formation, could be tested by reducing the reaction temperature from 37 °C to 21 °C, 16 °C, and 4 °C, and testing yields over time with qualitative RNA product assessment as has been performed here. This approach could benefit the reaction due to the intrinsic kinetics of polymerase action in stages of transcription initiation, elongation, and detachment[125]. The enzyme binds the promoter, opens the DNA strand, and initiates RNA-DNA hybrid band formation[126]. After 3-7 promoter nucleotides, the pressure exerted on the polymerase rotates it to 40°, and after 9-12 nucleotides, a larger 220° rotation occurs[127] which releases the enzyme from promoter contacts and continues transcribing in a stable elongation complex[128, 129]. Previous studies have shown that abortive RNA byproducts resulting from the dissociation of the enzyme from the RNA-DNA hybrid strand, and abortive initiation events before the elongation phase, are commonplace, with only ~56% transcription reactions proceeding to elongation and full-length RNA product formation[125, 130]. Therefore, reducing the temperature to reduce enzyme error in elongation stage initiation by allowing more time for this step transition to occur may be particularly useful with kits like the MEGAShortScript™ kit, PCR amplicon template IVT reactions, and IVT reactions generating shorter RNA products e.g. $CUG_{n=11}$ RNA. Another way to increase the yield of RNA per IVT reaction could be to redesign the initial plasmid template to include 5'-GGG ATAAT-3' between T7 promoter +1 base position and CTG repeats, as evidenced by Conrad et al., (2020)[131] to drastically increase transcription yields by T7 RNA polymerase action through decreased premature dissociation from potential abortive transcripts. This would come at the cost of introducing those extra bases that are not DM1-specific.

Ideally, generating transcripts with an in-house assembled IVT reaction, such as the one constructed in this work, would be the best option to increase yields in the laboratory cost-efficiently, as this reaction should be scalable from 100 µL and applicable to any template with a T7 Promoter site. A typical IVT reaction should have the following components: DNA template, RNA Polymerase, rNTP's, RNAse Inhibitors, and reaction buffer containing salts, Mg^{2+} [132]. Our reaction included 40 mM HEPES-KOH as a buffering agent, 2 mM Spermidine which has been shown to aid transcription by aiding disassociation of T7 RNA Polymerase from DNA template strand [133] and 20 mM DTT as a protection agent against oxidation for T7 RNA polymerase [134]. Our

reaction did not produce any measurable RNA and was only repeated once. This section should be further investigated to benchmark an in-house assembled, scalable IVT reaction. The reaction could be improved with the addition of inorganic pyrophosphatase (iPPase), which is recommended to cleave any pyrophosphate that precipitates in the presence of Mg^{2+} ions, lowering ion bioavailability for T7 RNA Polymerase [135]. Future reaction experiment design should also include gradients of $MgCl_2$, template concentration series, and T7 polymerase concentration series testing, rNTP concentration series (with $MgCl_2$ adjusted to exceed NTP concentration by 5 mM, as Mg^{2+} binds NTPs [132] across various time-points and reaction temperatures at pH 8.0. This was not possible in this piece of work due to the high amounts of expensive T7 RNA Polymerase enzyme needed, and time constraints, therefore producing and validating the activity of this protein in-house, for example from plasmid pT7-911q [136] would precede in-house IVT optimisation.

Other DNA-dependent-RNA polymerases could be trialled as well, specifically SP6 RNA Polymerase also known for its T7-comparable, high specificity to bind its relevant promoter region[137]. A *E.coli* poly(A) RNA Polymerase [138]reaction could be performed in tandem with RNA reactions, for example, as part of the QC workflow to generate PolyA tails for cDNA library synthesis for sequencing product transcripts.

Quality control improvements for RNA transcript assessment:

To ensure correct RNA transcript creation, the Quality Control element of this work has been challenging when trying to determine the correct sequence of repeat CUG sequences. Initially, analysing clean RNA transcripts by Agilent 4200 TapeStation System, based on automated electrophoresis technology, can provide a valuable assessment of RNA integrity through a RNA Integrity Number (RIN) value, this score, however, is commonly based on 28s:18s Ribosomal RNA presence[139], and does not apply to this study due to the origin of RNA transcripts made from IVT reactions only. Therefore, the assessment focus was on observing non-specific banding, multiple species creation, and RNA degradation patterns. Peak RNA assignment to determine the size of the transcript created compares to the electronic ladder provided, and this peak assignment can be flawed as the conditions for sample preparation and during the run are non-denaturing conditions, which may fail to negate the effects of RNA secondary structure [140]. This is a large factor to consider with highly structured pathogenic RNA, therefore, an approach to run the equivalent RNA sample under denaturing gel electrophoresis conditions was taken, in this case via 15% TBE-urea gels, with sample preparation in formamide-based buffer and heating to 95˚C to denature secondary structure before gel loading. These results with $CUG_{n=54}$ have shown in figure 16b a difference in band size presence on Agilent 4200 Tapestation system analysis, averaging ~125-130 nts (peak number assignment not shown in results here), with variation between samples apparent. In denaturing gels, the bands show a consistent height at the ~162 nt mark, suggesting correct transcript identity. With $CUG_{n=166}$ RNA transcripts, this difference observed was much larger, with repeats in figure 17b n=1 and n=3 showing a double-band presence at ~150/200nt mark and a shouldered peak at ~300-400nt mark respectively. When run on denaturing RNA

gels, these samples show single banding corresponding to the 500bp mark on the ssRNA ladder, suggesting they should be correct. Ideally, a lower percentage acrylamide gel (e.g., 10%) should be chosen to best differentiate bands <500bp to allow any larger fragments present to separate, as higher MW ladder bands struggle to migrate down 15% gels[141]. It would be ideal to perform sequencing on the RNA transcripts to confirm triplicate base uniformity, however, due to the presence of only CUG repeats in IVT transcripts, it would make traditional cost-effective methods difficult, such as Sanger sequencing[142], or PCR cDNA library preparation steps in typical Next Generation Sequencing (NGS)[143], and even with adapter sequence ligation during RNA-seq methods sample preparation[144]. NGS could be performed for DMPK-$CUG_{n=11}$ transcripts following a cDNA library creation step using DMPK and 3' regions outside of $CUG_{n=11}$ repeats for primer compatibility. For the other transcripts, an approach such as FRT-seq could be taken[145], by providing ligation of an adaptor sequence to 3' end of RNA transcripts and an on-run reverse transcription step to read outputs. Direct RNA Sequencing (DRS)[146] could also work, with the addition of a Poly-A 3' tail to our RNA transcripts. With new emerging technologies for direct RNA measurement, sequencing could be performed via Oxford Nanopore Technologies[143, 147], by Poly-Adenylating RNA transcripts, ligating relevant adapter sequences to this region, and measuring current changes in base-by-base reads as the constructs are pulled through the flow-cell channels. Due to the current cost of this sequencing technology, it was not feasible for this study at the time being. A feasible, cost-efficient approach to try would be generating cDNA libraries via SMART technology and using Template Switch Oligos coupled with Moloney Murine Leukemia Virus (MMLV) reverse transcriptase. After Polyadenylation of IVT RNA transcripts, a first complementary strand is generated with an Oligo dT primer, where MMLV activity adds a few bases (deoxycytidine) at the complementary 3' end. The next oligo with an end sequence complementary to these new C's is added, and the MMVL reverse transcriptase switches from RNA strand to TS Oligo strand, creating cDNA with end-specific sequences[148, 149]. This can be amplified for library prep. Once this is complete, this cDNA library can be read by more traditional, cheaper methods such as Sanger Sequencing.

Conclusions:

Future development work for establishing predictable, reproducible outcomes of the RNA creation process explored in this work is necessary as the process optimisation steps taken here, and improvements made were often below n=3. This is partly due to resource and project time constraints, and due to the reliability of the reactions themselves. Utmost care should be taken when working under RNAse-free conditions to not contaminate the work area, reactions, or reaction component stocks. Many reaction components are sensitive to temperature changes and handling procedures; therefore, careful record-keeping of stock material should be followed. With these considerations in mind, however, DM1 representative RNA has been produced in individual IVT reactions to a satisfactory level for progressing to large-scale synthesis. DMPK-$CUG_{n=11}$ RNA transcript results showed correct banding patterns with low non-specificity when analysed, as well as $CUG_{n=54}$ and $CUG_{n=166}$ RNA when produced with

the T7 RiboMAX™ Express Large Scale RNA Production System kit. Optimising the MEGAshortscript™ T7 Kit would be ideal for maximising short RNA transcript creation; however, more method development would be required. Producing PCR template amplicons to overcome these issues achieves this endpoint in terms of starting molarity, but the quality of transcripts deteriorated with this type of DNA template and would not be recommended as performed here. Yields of $CUG_{n=166}$ were lower when compared to $CUG_{n=54}$ transcript creation, therefore attempts described to optimise the scale-up of this construct should be considered. Although quality control steps developed were appropriate for transcript identity determination and unique species creation checks, a final sequencing quality control step with potential methods discussed would be necessary for base-to-base RNA confirmation to take to further structural study.

Appendix I

Primer design:

All primers were used in this study were designed based on feature predictions from Primer3 technology platform[150].

Table 6) Primer details for Sanger sequencing of pBSKII-DMPK-$CTG_{n=11}$, and pTRE2 Luc-$CTG_{n=160}$.

Primer Name	5'-3' Primer Sequence	Tm	GC %	Nt
pBSKII-DMPK-CTGn=11				
M13fwd(-20), pBSKII CTG11	GTAAAACGACGGCCAGTG	53.3	55.56	18
M13rev(-27), pBSKII CTG11	GGAAACAGCTATGACCATG	50.1	47.37	19
pTRE2 Luc-$CTG_{n=160}$				
CMVmin	CGC CAT CCA CGC TGT TTT G	57.1	57.89	19
Primer 1 fwd, pTRE2LucCTG	AGTCGATGTACACGTTCGTC	54.3	50	20
Primer 2 fwd, pTRE2LucCTG	GGCTCACTGAGACTACATCAG	53.9	52.38	21
Primer 3 fwd, pTRE2LucCTG	TAGAACTGTCTTCGACTCCG	52.9	50	20

Primer 4 fwd, pTRE2LucCTG	TCGAAGGGTCCTTGTAGC	52.6	55.56	18
Primer 5 rev, pTRE2LucCTG	ATGCTGCAGAGATCTGGATC	53.6	50	20

Table 7) Primer details for PCR amplification of IVT linear template DNA of pUC19_CTG$_{n=11,}$ pUC19_CTG$_{n=54}$.

Primer Name	5'-3' Primer Sequence	Tm	GC %	Nt
T7.CTGn Forward Primer	ACGCGGCCTTTTTACGGTTCCTGGC	65.9	60	25
T7.CTGn Reverse Primer	TGGCGAAAGGGGGATGTGCTGCAAG	65.4	60	25

Table 8) Primer details for T7 promoter sequence incorporation into pTRE2-Luc-CTG$_{n=160}$ by PCR and tHDA reactions.

Primer Name	5'-3' Primer Sequence	PCR or tHDA	Tm	GC %	Nt
Forward Primer (T7)	TAATACGACTCACTATAGGG CTTGTAGCCGGGAATG	PCR	64.4	47.22	36
Reverse Primer (T7)	GCCGAAAGAAAGAAAT GGTCTGTGATCCCC	PCR	63.0	50	30
tHDA Primer Fwd T7 CTG160	TAATACGACTCACTATAG GGTCCTTGTAGCCGGG	tHDA	64.1	50	34
tHDA Primer Rev 1 T7 CTG160	CTGGCCGAAAGAAAGA AATGGTCTGTGATCCCC	tHDA	65.4	51.52	33
tHDA Primer Rev 2 T7 CTG160	TTGCTGCCTTCCCA GGCCTGCAGTTTG	tHDA	67.0	59.26	27

Chapter 4: Future perspectives

Potential strategies for RNP high-resolution structure determination:

For MBNL1, X-ray crystallography structures have been obtained for MBNL1 ZnF domains 1-2 at 2.70Å (PDB no. 3D2N) and MBNL1 ZnF3-4 at 1.50Å resolution (PDB no. 3D2Q) alone, and at 1.7Å when combined with CGCUGU RNA (PDB no. 3D2S)[75]. However, full-length MBNL1 structures with DM1 disease-state RNA have not been reported to date, or by alternative structure-based methods. In the context of this technique, various caveats may contribute to this lack of publication. The main setback here is likely the difficulty in crystallising MBNL1. Analysis of MBNL1 full-length structure predictions by disorder probability residue plot of Uniprot entry Q9NR56 from PrDOS structure disorder prediction web server[151], shows considerable disorder, particularly in the C-term domain of MBNL1. Disordered proteins tend to struggle to form structured, packed crystal conformations when compared to their more globular counterparts. Crystallisation trials tend to need covering of a wide range of conditions for crystal formation [152, 153], which in turn requires high concentrations of protein for saturation points and crystal formation. With the current full-length MBNL1 expression and purification process, much refinement is needed for the gram quantities required. However, once pure MBNL1 constructs are obtained, crystal trials can be accelerated through modern automation approaches, such as with the Mosquito® crystal Instrument [154]. Trials would be performed with MBNL1 constructs alone, and in combination with $CUG_{n=11}$, $CUG_{n=54}$, and $CUG_{n=166}$. These crystals would then be exposed to an X-ray beam source, and diffraction patterns developed into high-resolution 3D structures[155].

Cryo-Electron Microscopy (Cryo-EM) near-atomic structure resolutions have been obtained rivaling those from traditional approaches such as X-ray Crystallography or Nuclear Magnetic Resonance (NMR)[156, 157]. Cryo-EM can also bypass caveats associated with conventional techniques, such as the high failure rate of crystallisation trials with difficult targets. For this study to be taken to Cryo-EM, DM1 disease-representative transcripts should be re-constituted biochemically with purified MBNL1 protein, in a stoichiometry corresponding to that described in DM1 pathology, such as previously found MBNL1:$CUG_{n=12}$ in ~3:1[59] stoichiometry conformation, and in increasing ratios of MBNL1:CUG^{exp}, until re-constituted under native cell-like conditions. Due to the complex, not fully understood composition of disease-state RNP complexes, predicting MBNL1 binding saturation for each type of RNA transcript produced could be performed computationally through generating PDB files for RNA 3D predicted structure[158], binding these to full-length MBNL1 structure predicted files in a Docking prediction server such as HDOCK[159], while also specifying protein residue binding sites existing in current published structural datafiles, such as, specified within NMR data with MBNL1 ZnF1-2, 3-4 binding Cardiac Troponin T RNA[76]. The results could be re-docked to MBNL1 only PDBs until saturation is achieved (initial stages represented in figure 18 below).

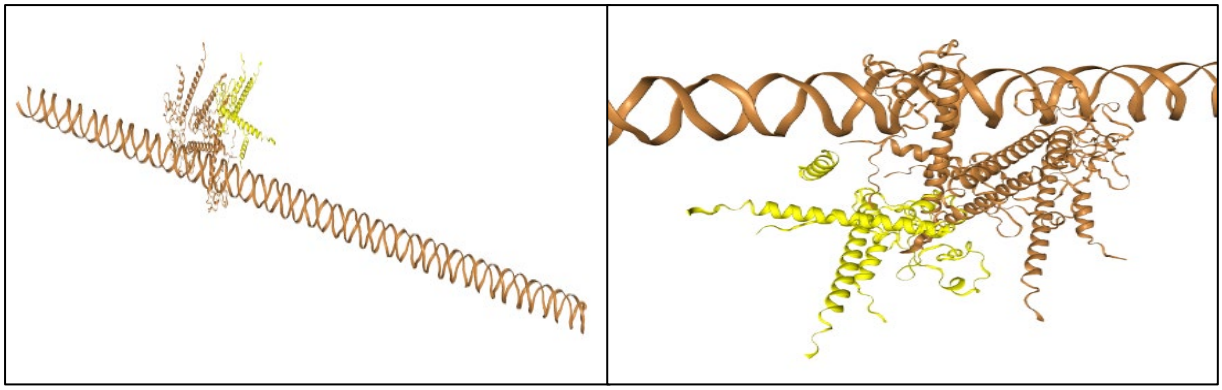


Figure 18) Visual representation of second iterations of MBNL1 full length (Yellow, Alphafold predicted structure) prediction binding to $CUG_{n=166}$ RNA (3D RNAcomposer predicted structure) plus 1x MBNL1-FL bound (brown), using parameters explained above, and relevant shortlisting, to generate binding predictions via HDOCK.

Once samples are deposited on grids, negative-stain TEM could be performed to streamline grid screening, with the best conditions applied to a Cryo-EM pipeline process for sample preparation and image collection. Here, cryo-stage preparation of samples should be performed with a vitrification apparatus e.g. Vitrobot[160]. Once well-dispersed particle distribution images in multiple orientations are collected by cryo-TEM, automated 3D structure reconstructions can be performed computationally[161, 162]. In addition to RNP complex DM1 grids, positive controls for each separate molecular component and negative control grids should be prepared for comparison and confidence when screening for structures.

At the core of DM1 pathogenesis, and consequent clinical disease manifestations, is the formation of large nuclear RNP foci formed by the sequestering of MBNL1 proteins with toxic mRNA formed of CUG^{exp} repeats (as well as other factors previously described). Here, a technique which could investigate these complexes in detail is Cryo-Electron Tomography (Cryo-ET). The popularity of Cryo-ET as a technique for determining within-cell structures at sub-atomic resolution has resulted in 3D structure determination results at <4Å resolution[163]. When applying this to DM1 foci, their approximate size (<200nm single foci, 200->1250nm) is above published averages for structure analysis via cryo-ET capabilities. Model cell lines exist to represent severe DM1 disease, for example, immortalised DM myoblast cell line from patient biopsy tissue containing ~2600 CTG repeats[164], many established primary-derived DM1 fibroblast cell lines with varying CTG repeats, confirmed to form nuclear RNP foci (≥80) [19, 70, 165, 166], or commercially available non disease/disease DM1 skin fibroblasts [GM04033 (CTG_{1000}) and GM03132 (CTG_{2000})] [167]. These longer repeats would yield higher incidence/ size of foci to be studied, especially with documented cytoplasmic-present RNP foci where CUG repeats were ≥2000[168]. This idea is focused on ultrastructure determination in a specific cell line; however, it could be adapted as a way of visualising RNP foci changes under the addition of future therapeutic compounds for the formation or disruption of foci, compared to control cell lines treated.

Additional biophysical investigations into DM1 RNA-protein interactions, complex conformation, and supplementary experiments:

In addition to structure determination, it would be interesting to directly observe MBNL1 binding profiles to the different created DM1 representative RNA transcripts, against the non-disease $CUG_{n=11}$ repeats as a negative control, in real time. This has previously been performed with RNP complexes of dsRNA nanostructures, and has visualised direct changes to RNA conformations and binding sites in real-time[169], through an approach known as High-Speed Atomic Force Microscopy (HS-AFM). Binding kinetics could be studied with MBNL1: RNA at different stoichiometries via Bio-layer Interferometry (BLI) or Surface Plasmon Resonance (SPR) approaches. These methods would be a good addition to the predicted docking of MBNL1 on DM1 RNA, as described above. In addition to high-resolution structure data, access to a Small Angle X-ray Scattering (SAXS) beamline would be beneficial to provide low-resolution (~10-50Å *d*-spacing) in solution data around MBNL1-RNA complex folding dynamics, flexible domain structuring, or even allosteric effects when introducing additional identified macromolecular complex binding partners to the MBNL1-RNA complex, or at varying stoichiometries when reconstituted[170].

Chapter 5: Conclusions

The project has aimed to produce DM1 representative molecules, at a high quality and in a quantity sufficient to be taken towards further structural biology study, to eventually produce a high-resolution representative structure of CUG^{exp}-MBNL1 RNP complex to take forwards for structure-based drug design of novel compounds to develop an effective cure for DM1 disease.

This project has led to development of good molecular production processes for DM1 repeat CUG sequence RNA and has shed light on the difficulty of obtaining and validating repeat triplicate sequence RNA. Applied process development for expressing and purifying MBNL1 protein constructs within *E. coli* has been shown potential for obtaining high purity yields from constructs described and would need re-designing if applying to techniques requiring considerable starting material amounts. To take CUG^{exp}-MBNL1 RNP complexes through future structural studies, expression of MBNL1-FL, ZnF1-4 domains, and C-term should be the next step for optimisation. Then, once the protein molecules have been expressed at a high yield and purity level, crystallisation trials of these could commence, coupled with TEM and cryo-TEM grid preparation for screening and in-solution high-resolution structure determination.

References

1. Thornton, C.A., *Myotonic dystrophy.* Neurologic clinics, 2014. **32**(3): p. 705-719.
2. Suominen, T., et al., *Population frequency of myotonic dystrophy: higher than expected frequency of myotonic dystrophy type 2 (DM2) mutation in Finland.* European Journal of Human Genetics, 2011. **19**(7): p. 776-782.
3. Johnson, N.E., et al., *Population-based prevalence of myotonic dystrophy type 1 using genetic analysis of statewide blood screening program.* Neurology, 2021. **96**(7): p. e1045-e1053.
4. Norwood, F.L., et al., *Prevalence of genetic muscle disease in Northern England: in-depth analysis of a muscle clinic population.* Brain, 2009. **132**(11): p. 3175-3186.
5. Siciliano, G., et al., *Epidemiology of myotonic dystrophy in Italy: re-apprisal after genetic diagnosis.* Clinical genetics, 2001. **59**(5): p. 344-349.
6. Liao, Q., et al., *Global prevalence of myotonic dystrophy: An updated systematic review and meta-analysis.* Neuroepidemiology, 2022. **56**(3): p. 163-173.
7. Brook, J.D., et al., *Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member.* Cell, 1992. **68**(4): p. 799-808.
8. Fu, Y., et al., *An unstable triplet repeat in a gene related to myotonic muscular dystrophy.* Science, 1992. **255**(5049): p. 1256-1258.
9. Mahadevan, M., et al., *Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene.* Science, 1992. **255**(5049): p. 1253-1255.
10. Liquori, C.L., et al., *Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9.* Science, 2001. **293**(5531): p. 864-867.
11. Turner, C. and D. Hilton-Jones, *The myotonic dystrophies: diagnosis and management.* Journal of Neurology, Neurosurgery & Psychiatry, 2010. **81**(4): p. 358-367.
12. Heatwole, C., et al., *Patient-reported impact of symptoms in myotonic dystrophy type 1 (PRISM-1).* Neurology, 2012. **79**(4): p. 348-357.
13. Ashizawa, T., et al., *Consensus-based care recommendations for adults with myotonic dystrophy type 1.* Neurology: Clinical Practice, 2018. **8**(6): p. 507-520.
14. Bioethics, C.o., *Committee on Genetics, and, The American College of Medical Genetics and, Genomics Social, Ethical, and Legal Issues Committee: Ethical and Policy Issues in Genetic Testing and Screening of Children.* Pediatrics, 2013. **131**(3): p. 620-622.
15. Kamsteeg, E.-J., et al., *Best practice guidelines and recommendations on the molecular diagnosis of myotonic dystrophy types 1 and 2.* European Journal of Human Genetics, 2012. **20**(12): p. 1203-1208.
16. Prior, T.W., *Technical standards and guidelines for myotonic dystrophy type 1 testing.* Genetics in Medicine, 2009. **11**(7): p. 552-555.
17. Michalowski, S., et al., *Visualization of double-stranded RNAs from the myotonic dystrophy protein kinase gene and interactions with CUG-binding protein.* Nucleic Acids Research, 1999. **27**(17): p. 3534-3542.
18. Pettersson, O.J., et al., *Molecular mechanisms in DM1—a focus on foci.* Nucleic acids research, 2015. **43**(4): p. 2433-2441.

19. Ketley, A., et al., *High-content screening identifies small molecules that remove nuclear foci, affect MBNL distribution and CELF1 protein levels via a PKC-independent pathway in myotonic dystrophy cell lines.* Human molecular genetics, 2014. **23**(6): p. 1551-1562.
20. Dansithong, W., et al., *Cytoplasmic CUG RNA Foci Are Insufficient to Elicit Key DM1 Features.* PLOS ONE, 2008. **3**(12): p. e3968.
21. Davis, B.M., et al., *Expansion of a CUG trinucleotide repeat in the 3' untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts.* Proceedings of the National Academy of Sciences, 1997. **94**(14): p. 7388-7393.
22. Taneja, K.L., et al., *Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues.* The Journal of cell biology, 1995. **128**(6): p. 995-1002.
23. Mankodi, A., et al., *Muscleblind localizes to nuclear foci of aberrant RNA in myotonic dystrophy types 1 and 2.* Human molecular genetics, 2001. **10**(19): p. 2165-2170.
24. Xia, G. and T. Ashizawa, *Dynamic changes of nuclear RNA foci in proliferating DM1 cells.* Histochemistry and cell biology, 2015. **143**: p. 557-564.
25. Miller, J.W., et al., *Recruitment of human muscleblind proteins to (CUG) n expansions associated with myotonic dystrophy.* The EMBO journal, 2000. **19**(17): p. 4439-4448.
26. Fardaei, M., et al., *Three proteins, MBNL, MBLL and MBXL, co-localize in vivo with nuclear foci of expanded-repeat transcripts in DM1 and DM2 cells.* Human molecular genetics, 2002. **11**(7): p. 805-814.
27. Ho, T.H., et al., *Transgenic mice expressing CUG-BP1 reproduce splicing mis-regulation observed in myotonic dystrophy.* Human molecular genetics, 2005. **14**(11): p. 1539-1547.
28. Paul, S., et al., *Expanded CUG repeats dysregulate RNA splicing by altering the stoichiometry of the muscleblind 1 complex.* Journal of Biological Chemistry, 2011. **286**(44): p. 38427-38438.
29. Querido, E., et al., *Stochastic and reversible aggregation of mRNA with expanded CUG-triplet repeats.* Journal of cell science, 2011. **124**(10): p. 1703-1714.
30. Smith, K.P., et al., *Defining early steps in mRNA transport: mutant mRNA in myotonic dystrophy type I is blocked at entry into SC-35 domains.* The Journal of cell biology, 2007. **178**(6): p. 951-964.
31. Kalsotra, A., et al., *The Mef2 transcription network is disrupted in myotonic dystrophy heart tissue, dramatically altering miRNA and mRNA expression.* Cell reports, 2014. **6**(2): p. 336-345.
32. Batra, R., et al., *Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease.* Molecular cell, 2014. **56**(2): p. 311-322.
33. Timchenko, N.A., et al., *RNA CUG repeats sequester CUGBP1 and alter protein levels and activity of CUGBP1.* Journal of Biological Chemistry, 2001. **276**(11): p. 7820-7826.
34. Konieczny, P., et al., *Autoregulation of MBNL1 function by exon 1 exclusion from MBNL1 transcript.* Nucleic Acids Research, 2016. **45**(4): p. 1760-1775.
35. López-Martínez, A., et al., *An overview of alternative splicing defects implicated in myotonic dystrophy type I.* Genes, 2020. **11**(9): p. 1109.

36. Dixon, D.M., et al., *Loss of muscleblind-like 1 results in cardiac pathology and persistence of embryonic splice isoforms.* Scientific reports, 2015. **5**(1): p. 9042.
37. Kino, Y., et al., *MBNL and CELF proteins regulate alternative splicing of the skeletal muscle chloride channel CLCN1.* Nucleic acids research, 2009. **37**(19): p. 6477-6490.
38. Bannister, R.A. and K.G. Beam, *CaV1. 1: The atypical prototypical voltage-gated Ca2+ channel.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2013. **1828**(7): p. 1587-1597.
39. Fugier, C., et al., *Misregulated alternative splicing of BIN1 is associated with T tubule alterations and muscle weakness in myotonic dystrophy.* Nature medicine, 2011. **17**(6): p. 720-725.
40. Nitschke, L., et al., *Alternative splicing mediates the compensatory upregulation of MBNL2 upon MBNL1 loss-of-function.* Nucleic Acids Research, 2023. **51**(3): p. 1245-1259.
41. Konieczny, P., et al., *Autoregulation of MBNL1 function by exon 1 exclusion from MBNL1 transcript.* Nucleic acids research, 2017. **45**(4): p. 1760-1775.
42. Cass, D., et al., *The four Zn fingers of MBNL1 provide a flexible platform for recognition of its RNA binding elements.* BMC Molecular Biology, 2011. **12**(1): p. 1-7.
43. Fernandez-Costa, J.M. and R. Artero, *A conserved motif controls nuclear localization of Drosophila Muscleblind.* Molecules and cells, 2010. **30**: p. 65-70.
44. Kino, Y., et al., *Nuclear localization of MBNL1: splicing-mediated autoregulation and repression of repeat-derived aberrant proteins.* Human molecular genetics, 2015. **24**(3): p. 740-756.
45. Yuan, Y., et al., *Muscleblind-like 1 interacts with RNA hairpins in splicing target and pathogenic RNAs.* Nucleic Acids Research, 2007. **35**(16): p. 5474-5486.
46. Fernandez-Costa, J.M., et al., *Alternative splicing regulation by Muscleblind proteins: from development to disease.* Biological Reviews, 2011. **86**(4): p. 947-958.
47. Onishi, H., et al., *MBNL1 associates with YB-1 in cytoplasmic stress granules.* Journal of neuroscience research, 2008. **86**(9): p. 1994-2002.
48. Consortium, I.M.D., *New nomenclature and DNA testing guidelines for myotonic dystrophy type 1 (DM1).* Neurology, 2000. **54**(6): p. 1218-1221.
49. Howeler, C., et al., *Anticipation in Myotonic Dystrophy II Anticipation in Myotonic Dystrophy: Fact or Fiction?(1989).* OXFORD MONOGRAPHS ON MEDICAL GENETICS, 2004. **51**(1): p. 228-243.
50. Martorell, L., et al., *Somatic instability of the myotonic dystrophy (CTG) n repeat during human fetal development.* Human Molecular Genetics, 1997. **6**(6): p. 877-880.
51. Morales, F., et al., *Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity.* Human Molecular Genetics, 2012. **21**(16): p. 3558-3567.
52. Lavedan, C., et al., *Myotonic dystrophy: size-and sex-dependent dynamics of CTG meiotic instability, and somatic mosaicism.* American journal of human genetics, 1993. **52**(5): p. 875.
53. Pearson, C.E., K.N. Edamura, and J.D. Cleary, *Repeat instability: mechanisms of dynamic mutations.* Nature Reviews Genetics, 2005. **6**(10): p. 729-742.
54. Yang, Z., et al., *Replication inhibitors modulate instability of an expanded trinucleotide repeat at the myotonic dystrophy type 1 disease locus in human cells.* The American Journal of Human Genetics, 2003. **73**(5): p. 1092-1105.

55. Fortune, M.T., et al., *Dramatic, expansion-biased, age-dependent, tissue-specific somatic mosaicism in a transgenic mouse model of triplet repeat instability.* Human Molecular Genetics, 2000. **9**(3): p. 439-445.
56. Sobczak, K., et al., *RNA structure of trinucleotide repeats associated with human neurological diseases.* Nucleic acids research, 2003. **31**(19): p. 5469-5482.
57. Mooers, B.H.M., J.S. Logue, and J.A. Berglund, *The structural basis of myotonic dystrophy from the crystal structure of CUG repeats.* Proceedings of the National Academy of Sciences, 2005. **102**(46): p. 16626-16631.
58. Masuda, A., et al., *CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay.* Scientific reports, 2012. **2**(1): p. 209.
59. Haghighat Jahromi, A., et al., *Single-molecule study of the CUG repeat–MBNL1 interaction and its inhibition by small molecules.* Nucleic acids research, 2013. **41**(13): p. 6687-6697.
60. Pascual-Gilabert, M., R. Artero, and A. López-Castel, *The myotonic dystrophy type 1 drug development pipeline: 2022 edition.* Drug Discovery Today, 2023: p. 103489.
61. Pascual-Gilabert, M., A. Lopez-Castel, and R. Artero, *Myotonic dystrophy type 1 drug development: A pipeline toward the market.* Drug Discovery Today, 2021. **26**(7): p. 1765-1772.
62. Nakamori, M., et al., *Oral administration of erythromycin decreases RNA toxicity in myotonic dystrophy.* Annals of clinical and translational neurology, 2016. **3**(1): p. 42-54.
63. Johnson, N., et al., *Study Design of AOC 1001-CS1, a Phase 1/2 Clinical Trial Evaluating the Safety, Tolerability, Pharmacokinetics and Pharmacodynamics of AOC 1001 Administered Intravenously to Adult Patients with Myotonic Dystrophy Type 1 (DM1)(MARINA)(S23. 006)*. 2022, AAN Enterprises.
64. Zanotti, S., et al. *Repeat Dosing with DYNE-101 is Well Tolerated and Leads to a Sustained Reduction of DMPK RNA Expression in Key Muscles for DM1 Pathology in hTfR1/DMSXL Mice and NHPs*. in *MOLECULAR THERAPY*. 2022. CELL PRESS 50 HAMPSHIRE ST, FLOOR 5, CAMBRIDGE, MA 02139 USA.
65. Wolf, D., et al., *P50 A phase 1/2 randomized, placebo-controlled, multiple ascending dose study (ACHIEVE) of DYNE-101 in individuals with myotonic dystrophy type 1 (DM1).* Neuromuscular Disorders, 2023. **33**: p. S71.
66. Arandel, L., et al., *Reversal of RNA toxicity in myotonic dystrophy via a decoy RNA-binding protein with high affinity for expanded CUG repeats.* Nature Biomedical Engineering, 2022. **6**(2): p. 207-220.
67. Li, H., et al., *Inhibition of HBV expression in HBV transgenic mice using AAV-delivered CRISPR-SaCas9.* Frontiers in immunology, 2018. **9**: p. 2080.
68. Batra, R., et al., *The sustained expression of Cas9 targeting toxic RNAs reverses disease phenotypes in mouse models of myotonic dystrophy type 1.* Nature biomedical engineering, 2021. **5**(2): p. 157-168.
69. Iftikhar, M., et al., *Current and emerging therapies for Duchenne muscular dystrophy and spinal muscular atrophy.* Pharmacology & Therapeutics, 2021. **220**: p. 107719.
70. Ketley, A., et al., *CDK12 inhibition reduces abnormalities in cells from patients with myotonic dystrophy and in a mouse model.* Science Translational Medicine, 2020. **12**(541): p. eaaz2415.

71. Xing, X., et al., *Disrupting the Molecular Pathway in Myotonic Dystrophy.* International Journal of Molecular Sciences, 2021. **22**(24): p. 13225.
72. Kalyaanamoorthy, S. and Y.-P.P. Chen, *Structure-based drug design to augment hit discovery.* Drug discovery today, 2011. **16**(17-18): p. 831-839.
73. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, *How many drug targets are there?* Nature reviews Drug discovery, 2006. **5**(12): p. 993-996.
74. Santos, R., et al., *A comprehensive map of molecular drug targets.* Nature reviews Drug discovery, 2017. **16**(1): p. 19-34.
75. Teplova, M. and D.J. Patel, *Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1.* Nature Structural & Molecular Biology, 2008. **15**(12): p. 1343-1351.
76. Park, S., et al., *Structural basis for interaction of the tandem zinc finger domains of human muscleblind with cognate RNA from human cardiac troponin T.* Biochemistry, 2017. **56**(32): p. 4154-4168.
77. Consortium, U. *Q9NR56 (SND1_HUMAN)*. 2025.
78. Gasteiger, E., et al., *Protein identification and analysis tools on the ExPASy server*. 2005: Springer.
79. Bjellqvist, B., et al., *Reference points for comparisons of two‐dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions.* Electrophoresis, 1994. **15**(1): p. 529-539.
80. Kino, Y., et al., *Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats.* Human Molecular Genetics, 2004. **13**(5): p. 495-507.
81. Fu, Y., et al., *MBNL1–RNA recognition: contributions of MBNL1 sequence and RNA conformation.* Chembiochem, 2012. **13**(1): p. 112-119.
82. Botta, A., et al., *MBNL142 and MBNL143 gene isoforms, overexpressed in DM1-patient muscle, encode for nuclear proteins interacting with Src family kinases.* Cell Death & Disease, 2013. **4**(8): p. e770-e770.
83. Harper, S. and D.W. Speicher, *Purification of proteins fused to glutathione S-transferase*, in *Protein chromatography: Methods and protocols*. 2010, Springer. p. 259-280.
84. Smyth, D.R., et al., *Crystal structures of fusion proteins with large‐affinity tags.* Protein science, 2003. **12**(7): p. 1313-1322.
85. Zhan, Y., X. Song, and G.W. Zhou, *Structural analysis of regulatory protein domains using GST-fusion proteins.* Gene, 2001. **281**(1-2): p. 1-9.
86. Tegel, H., et al., *Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta (DE3).* Protein expression and purification, 2010. **69**(2): p. 159-167.
87. Esposito, D. and D.K. Chatterjee, *Enhancement of soluble protein expression through the use of fusion tags.* Current opinion in biotechnology, 2006. **17**(4): p. 353-358.
88. Hewitt, S.N., et al., *Expression of proteins in Escherichia coli as fusions with maltose-binding protein to rescue non-expressed targets in a high-throughput protein-expression and purification pipeline.* Acta Crystallographica Section F: Structural Biology and Crystallization Communications, 2011. **67**(9): p. 1006-1009.
89. Raran-Kurussi, S., K. Keefe, and D.S. Waugh, *Positional effects of fusion partners on the yield and solubility of MBP fusion proteins.* Protein expression and purification, 2015. **110**: p. 159-164.

90. Chen, X., J.L. Zaro, and W.-C. Shen, *Fusion protein linkers: property, design and functionality.* Advanced drug delivery reviews, 2013. **65**(10): p. 1357-1369.
91. Bai, Y. and W.-C. Shen, *Improving the oral efficacy of recombinant granulocyte colony-stimulating factor and transferrin fusion protein by spacer optimization.* Pharmaceutical research, 2006. **23**: p. 2116-2121.
92. Takamatsu, N., et al., *Production of enkephalin in tobacco protoplasts using tobacco mosaic virus RNA vector.* FEBS letters, 1990. **269**(1): p. 73-76.
93. Hartl, F.U. and M. Hayer-Hartl, *Molecular chaperones in the cytosol: from nascent chain to folded protein.* Science, 2002. **295**(5561): p. 1852-1858.
94. Nishihara, K., et al., *Overexpression of trigger factor prevents aggregation of recombinant proteins in Escherichia coli.* Applied and environmental microbiology, 2000. **66**(3): p. 884-889.
95. Choi, J.H. and S.Y. Lee, *Secretory and extracellular production of recombinant proteins using Escherichia coli.* Applied Microbiology and Biotechnology, 2004. **64**(5): p. 625-635.
96. Gopal, G.J. and A. Kumar, *Strategies for the Production of Recombinant Protein in Escherichia coli.* The Protein Journal, 2013. **32**(6): p. 419-425.
97. Francis, D.M. and R. Page, *Strategies to optimize protein expression in E. coli.* Current protocols in protein science, 2010. **61**(1): p. 5.24. 1-5.24. 29.
98. Tripathi, N.K., *Production and purification of recombinant proteins from Escherichia coli.* ChemBioEng Reviews, 2016. **3**(3): p. 116-133.
99. Rosano, G.L. and E.A. Ceccarelli, *Recombinant protein expression in Escherichia coli: advances and challenges.* Frontiers in microbiology, 2014. **5**: p. 172.
100. Ghaemmaghami, S., et al., *Global analysis of protein expression in yeast.* Nature, 2003. **425**(6959): p. 737-741.
101. Kost, T.A., J.P. Condreay, and D.L. Jarvis, *Baculovirus as versatile vectors for protein expression in insect and mammalian cells.* Nature biotechnology, 2005. **23**(5): p. 567-575.
102. Zhu, J., *Mammalian cell protein expression for biopharmaceutical production.* Biotechnology Advances, 2012. **30**(5): p. 1158-1170.
103. Giddings, G., et al., *Transgenic plants as factories for biopharmaceuticals.* Nature biotechnology, 2000. **18**(11): p. 1151-1155.
104. Tripathi, N.K. and A. Shrivastava, *Recent developments in bioprocessing of recombinant proteins: expression hosts and process development.* Frontiers in bioengineering and biotechnology, 2019. **7**: p. 420.
105. Brookwell, A., J.P. Oza, and F. Caschera, *Biotechnology applications of cell-free expression systems.* Life, 2021. **11**(12): p. 1367.
106. Demain, A.L. and P. Vaishnav, *Production of recombinant proteins by microbes and higher organisms.* Biotechnology advances, 2009. **27**(3): p. 297-306.
107. Baghban, R., et al., *Yeast expression systems: overview and recent advances.* Molecular biotechnology, 2019. **61**: p. 365-384.
108. Cereghino, J.L. and J.M. Cregg, *Heterologous protein expression in the methylotrophic yeast Pichia pastoris.* FEMS microbiology reviews, 2000. **24**(1): p. 45-66.
109. Good, N.E., et al., *Hydrogen ion buffers for biological research.* biochemistry, 1966. **5**(2): p. 467-477.
110. Tsumoto, K., et al., *Effects of salts on protein–surface interactions: applications for column chromatography.* Journal of pharmaceutical sciences, 2007. **96**(7): p. 1677-1690.

111. Vagenende, V., M.G. Yap, and B.L. Trout, *Mechanisms of protein stabilization and prevention of protein aggregation by glycerol.* Biochemistry, 2009. **48**(46): p. 11084-11096.
112. Trivedi, M.V., J.S. Laurence, and T.J. Siahaan, *The role of thiols and disulfides on protein stability.* Current Protein and Peptide Science, 2009. **10**(6): p. 614-625.
113. Seddon, A.M., P. Curnow, and P.J. Booth, *Membrane proteins, lipids and detergents: not just a soap opera.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2004. **1666**(1-2): p. 105-117.
114. Cummins, P.M., K.D. Rochfort, and B.F. O'Connor, *Ion-exchange chromatography: basic principles and application.* Protein chromatography: methods and protocols, 2017: p. 209-223.
115. Lambert, N., et al., *RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins.* Molecular cell, 2014. **54**(5): p. 887-900.
116. Taylor, K., et al., *MBNL splicing activity depends on RNA binding site structural context.* Nucleic acids research, 2018. **46**(17): p. 9119-9133.
117. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466-1467.
118. Schäfer, F., et al., *Purification of GST-tagged proteins*, in *Methods in enzymology*. 2015, Elsevier. p. 127-139.
119. Kapust, R.B., et al., *The P1' specificity of tobacco etch virus protease.* Biochemical and biophysical research communications, 2002. **294**(5): p. 949-955.
120. Keil, B., *Specificity of proteolysis*. 2012: Springer Science & Business Media.
121. An, L., et al., *Characterization of a thermostable UvrD helicase and its participation in helicase-dependent amplification.* Journal of Biological Chemistry, 2005. **280**(32): p. 28952-28958.
122. Triana-Alonso, F.J., et al., *Self-coded 3'-Extension of Run-off Transcripts Produces Aberrant Products during in Vitro Transcription with T7 RNA Polymerase (∗).* Journal of Biological Chemistry, 1995. **270**(11): p. 6298-6307.
123. Ke, A. and J.A. Doudna, *Crystallization of RNA and RNA–protein complexes.* Methods, 2004. **34**(3): p. 408-414.
124. Markham, N.R. and M. Zuker, *UNAFold: software for nucleic acid folding and hybridization.* Bioinformatics: structure, function and applications, 2008: p. 3-31.
125. Lenk, R., et al., *Understanding the impact of in vitro transcription byproducts and contaminants.* Frontiers in Molecular Biosciences, 2024. **11**: p. 1426129.
126. Tang, G.-Q. and S.S. Patel, *Rapid binding of T7 RNA polymerase is followed by simultaneous bending and opening of the promoter DNA.* Biochemistry, 2006. **45**(15): p. 4947-4956.
127. Tang, G.-Q., et al., *Real-time observation of the transition from transcription initiation to elongation of the RNA polymerase.* Proceedings of the National Academy of Sciences, 2009. **106**(52): p. 22175-22180.
128. Ramírez-Tapia, L.E. and C.T. Martin, *New insights into the mechanism of initial transcription: the T7 RNA polymerase mutant P266L transitions to elongation at longer RNA lengths than wild type.* Journal of Biological Chemistry, 2012. **287**(44): p. 37352-37361.

129. Ma, K., et al., *Probing conformational changes in T7 RNA polymerase during initiation and termination by using engineered disulfide linkages.* Proceedings of the National Academy of Sciences, 2005. **102**(49): p. 17612-17617.
130. Koh, H.R., et al., *Correlating transcription initiation and conformational changes by a single-subunit RNA polymerase with near base-pair resolution.* Molecular cell, 2018. **70**(4): p. 695-706. e5.
131. Conrad, T., et al., *Maximizing transcription of nucleic acids with efficient T7 promoters.* Communications Biology, 2020. **3**(1): p. 439.
132. Beckert, B. and B. Masquida, *Synthesis of RNA by in vitro transcription.* RNA: methods and protocols, 2011: p. 29-41.
133. Gumport, R.I., *Effects of spermidine on the RNA polymerase reaction.* Annals of the New York Academy of Sciences, 1970. **171**(3): p. 915-938.
134. Kartje, Z.J., et al., *Revisiting T7 RNA polymerase transcription in vitro with the Broccoli RNA aptamer as a simplified real-time fluorescent reporter.* Journal of Biological Chemistry, 2021. **296**.
135. Tersteeg, S., et al., *Purification and characterization of inorganic pyrophosphatase for in vitro RNA transcription.* Biochemistry and Cell Biology, 2022. **100**(5): p. 425-436.
136. Ichetovkin, I.E., G. Abramochkin, and T.E. Shrader, *Substrate recognition by the leucyl/phenylalanyl-tRNA-protein transferase: conservation within the enzyme family and localization to the trypsin-resistant domain.* Journal of Biological Chemistry, 1997. **272**(52): p. 33009-33014.
137. Krieg, P.A. and D. Melton, *[25] In vitro RNA synthesis with SP6 RNA polymerase*, in *Methods in enzymology*. 1987, Elsevier. p. 397-415.
138. Yehudai-Resheff, S. and G. Schuster, *Characterization of the E. coli poly (A) polymerase: nucleotide specificity, RNA-binding affinities and RNA structure dependence.* Nucleic acids research, 2000. **28**(5): p. 1139-1144.
139. Schroeder, A., et al., *The RIN: an RNA integrity number for assigning integrity values to RNA measurements.* BMC Molecular Biology, 2006. **7**(1): p. 3.
140. Lilley, D.M., A. Bhattacharyya, and S. Mcateer, *Gel electrophoresis and the structure of RNA molecules.* Biotechnology and Genetic Engineering Reviews, 1992. **10**(1): p. 379-402.
141. Au - Summer, H., R. Au - Grämer, and P. Au - Dröge, *Denaturing Urea Polyacrylamide Gel Electrophoresis (Urea PAGE).* JoVE, 2009(32): p. e1485.
142. Edwards, A., et al., *Automated DNA sequencing of the human HPRT locus.* Genomics, 1990. **6**(4): p. 593-608.
143. Hu, T., et al., *Next-generation sequencing technologies: An overview.* Human Immunology, 2021. **82**(11): p. 801-811.
144. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews genetics, 2009. **10**(1): p. 57-63.
145. Mamanova, L., et al., *FRT-seq: amplification-free, strand-specific transcriptome sequencing.* Nature Methods, 2010. **7**(2): p. 130-132.
146. Ozsolak, F., et al., *Direct RNA sequencing.* Nature, 2009. **461**(7265): p. 814-818.
147. Garalde, D.R., et al., *Highly parallel direct RNA sequencing on an array of nanopores.* Nature Methods, 2018. **15**(3): p. 201-206.
148. Wellenreuther, R., et al., *SMART amplification combined with cDNA size fractionation in order to obtain large full-length clones.* BMC genomics, 2004. **5**: p. 1-8.

149. Zhu, Y., et al., *Reverse transcriptase template switching: A SMART™ approach for full-length cDNA library construction.* Biotechniques, 2001. **30**(4): p. 892-897.
150. Untergasser, A., et al., *Primer3—new capabilities and interfaces.* Nucleic acids research, 2012. **40**(15): p. e115-e115.
151. Ishida, T. and K. Kinoshita, *PrDOS: prediction of disordered protein regions from amino acid sequence.* Nucleic acids research, 2007. **35**(suppl_2): p. W460-W464.
152. Dubach, V.R.A. and A. Guskov, *The Resolution in X-ray Crystallography and Single-Particle Cryogenic Electron Microscopy.* Crystals, 2020. **10**(7): p. 580.
153. Maveyraud, L. and L. Mourey, *Protein X-ray Crystallography and Drug Discovery.* Molecules, 2020. **25**(5): p. 1030.
154. Li, D., et al., *Use of a robot for high-throughput crystallization of membrane proteins in lipidic mesophases.* JoVE (Journal of Visualized Experiments), 2012(67): p. e4000.
155. Yamamoto, M., et al., *Protein microcrystallography using synchrotron radiation.* IUCrJ, 2017. **4**(5): p. 529-539.
156. Nakane, T., et al., *Single-particle cryo-EM at atomic resolution.* Nature, 2020. **587**(7832): p. 152-156.
157. Yip, K.M., et al., *Atomic-resolution protein structure determination by cryo-EM.* Nature, 2020. **587**(7832): p. 157-161.
158. Antczak, M., et al., *New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure.* Acta Biochimica Polonica, 2016. **63**(4): p. 737-744.
159. Yan, Y., et al., *The HDOCK server for integrated protein–protein docking.* Nature protocols, 2020. **15**(5): p. 1829-1852.
160. Iancu, C.V., et al., *Electron cryotomography sample preparation using the Vitrobot.* Nature protocols, 2006. **1**(6): p. 2813-2819.
161. Scheres, S.H., *RELION: implementation of a Bayesian approach to cryo-EM structure determination.* Journal of structural biology, 2012. **180**(3): p. 519-530.
162. Zivanov, J., et al., *New tools for automated high-resolution cryo-EM structure determination in RELION-3.* elife, 2018. **7**: p. e42166.
163. Schur, F.K., et al., *An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation.* Science, 2016. **353**(6298): p. 506-508.
164. Arandel, L., et al., *Immortalized human myotonic dystrophy muscle cell lines to assess therapeutic compounds.* Disease models & mechanisms, 2017. **10**(4): p. 487-497.
165. Jenquin, J.R., et al., *Molecular characterization of myotonic dystrophy fibroblast cell lines for use in small molecule screening.* Iscience, 2022. **25**(5).
166. Hamshere, M.G., et al., *Transcriptional abnormality in myotonic dystrophy affects DMPK but not neighboring genes.* Proceedings of the National Academy of Sciences, 1997. **94**(14): p. 7394-7399.
167. Rodriguez, R., et al., *Altered nuclear structure in myotonic dystrophy type 1-derived fibroblasts.* Molecular biology reports, 2015. **42**: p. 479-488.
168. Pettersson, O.J., et al., *DDX6 regulates sequestered nuclear CUG-expanded DMPK-mRNA in dystrophia myotonica type 1.* Nucleic acids research, 2014. **42**(11): p. 7186-7200.
169. Osada, E., et al., *Engineering RNA–protein complexes with different shapes for imaging and therapeutic applications.* ACS nano, 2014. **8**(8): p. 8130-8140.

170. Skou, S., R.E. Gillilan, and N. Ando, *Synchrotron-based small-angle X-ray scattering of proteins in solution.* Nature Protocols, 2014. **9**(7): p. 1727-1739.