Bassel, George W. and Glaab, Enrico and Marquez, Julietta and Holdsworth, Michael J. and Barcardit, Jaume (2011) Functional network construction in Arabidopsis using rule-based machine learning on large-scale data sets. Plant Cell, 23 (9). pp. 3101-3116. ISSN 1040-4651

Access from the University of Nottingham repository:
http://eprints.nottingham.ac.uk/2434/1/line_15_Functional_Network_Construction.pdf

Copyright and reuse:

The Nottingham ePrints service makes this work by researchers of the University of Nottingham available open access under the following conditions.

This article is made available under the University of Nottingham End User licence and may be reused according to the conditions of the licence. For more details see: http://eprints.nottingham.ac.uk/end_user_agreement.pdf

A note on versions:

The version presented here may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the repository url above for details on accessing the published version and note that access may require a subscription.

For more information, please contact eprints@nottingham.ac.uk
Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets

George W. Bassel, Enrico Glaab, Julietta Marquez, Michael J. Holdsworth, and Jaume Bacardit

INTRODUCTION

Advances in postgenomic technologies and their use by the scientific community are generating increasing quantities of high-quality genome-wide transcriptomic data sets. Deposition of these data sets into publicly accessible online databases (Zimmermann et al., 2004; Toufighi et al., 2005) enables researchers to analyze the collated data and uncover novel information. The current rate at which data sets are being deposited is not matched by the generation of analytical tools capable of fully exploiting relevant information within these data. Therefore, there is a great need for additional analytical approaches to maximize the return on the large collective investment made in data generation. This demand for analytical tools is particularly pertinent for the investigation of complex traits, where a greater number of genes comprise the regulatory networks.

A range of statistical and computational methodologies of increasing complexity have been used to extract novel meaning from large data sets. Common approaches to uncovering gene function from transcriptome data include both differential expression across conditions and the calculation of correlations between gene expression levels across a large number of samples (Zimmermann et al., 2004; Toufighi et al., 2005; Brady and Provart, 2009; Usadel et al., 2009). This latter approach, termed coexpression analysis, is based on the guilt-by-association paradigm, where genes under the control of a common transcriptional regulatory mechanism have a greater probability of being involved in the same biochemical or developmental pathway and for their corresponding proteins to interact (Hughes et al., 2000; Usadel et al., 2009; Lee et al., 2010; Bassel et al., 2011; Mutwil et al., 2011). This correlative approach has successfully elucidated gene function in Arabidopsis thaliana through the use of genome-wide coexpression networks, leading to the identification of genes essential in the life cycle of Arabidopsis (Mutwil et al., 2010) and in the regulation of seed germination (Bassel et al., 2011).
Coexpression analyses consider all samples together and establish connections between genes based on all the collective information available. This approach can limit the interpretation of the data in processes such as developmental transitions, where discrete intermediate biological states are present. If biologically relevant gene associations are present transiently within subsets of samples representing these transitions, they will not be captured by coexpression where all samples are simultaneously considered equal.

Alternative methods to coexpression employing other measures to establish associations between genes may also be used. An example of such an alternative is machine learning (ML). ML can broadly be described as computer algorithms that automatically learn from experience (Mitchell, 1997). The data sets provided to these algorithms, termed training sets, are used to learn a predictive model based on the observations within the data. Two broad categories of ML approaches can be distinguished based on whether or not the training sets are annotated by the user according to the variables under investigation (in the case of biological data, the variable could be developmental state). If the data are annotated, then the process is termed supervised ML. If the data are not annotated, then this is termed unsupervised learning. Only supervised ML is able to take advantage of the additional information provided to the algorithm by the user when generating the predictive model. The model generated by supervised ML can then in turn be used to predict the annotation for samples of an undetermined state and identify processes controlling a specific developmental outcome.

Recently, ML techniques for microarray analysis that are able to identify patterns in subgroups of samples have been described. Biclustering (also known as two-way clustering or coclustering) is an unsupervised ML approach that establishes associations between groups of genes with a significantly similar expression profile across a subset of samples within a data set (Kluger et al., 2003; Sheng et al., 2003). Given that this approach uses nonannotated data, it does not use or consider information related to the biological status of the samples within the training set and as a result is unable to generate predictions.

Two types of annotation of training sets may be used with supervised ML. The first is using a categorical output describing the sample as belonging to a given category. This is referred to as a classification problem. Alternatively, a so-called regression program uses numerical values, such as those describing a gradient response. Here, we present a classification problem describing the binary fate of seeds, which may germinate or remain dormant. There are many different types of methods to perform classification. Some generate probabilistic models for each class of the problem based on the input variables, such as the Naïve Bayes algorithm (John and Langley, 1995). Other methods try to find a mathematical formulation that can split the variable space into two, so that samples of each class lie at one side or the other. This can occur either in the original space of variables, as in a linear classifier, or in a higher dimensionality space, such as in nonlinear support vector machines (SVMs) (Vapnik, 1995). An alternative approach is to decompose the space of variables into an arbitrary number of hyper-rectangular subsets (see Supplemental Figure 1 online). Rule-based learning methods (Furnkranz, 1999) follow this last approach, where a model is composed by an arbitrary set of decision rules, where each rule specifies a subset of the input space.

The partitioning process used by rule-based learning methods focuses on identifying subgroups of samples contained within the training set. In the context of the analysis of large-scale biological data sets, discrete developmental states can be identified within the training set given that samples belonging to a given state are likely to have similar characteristics. This partitioning process highlights a key difference to unsupervised coexpression analysis where all samples are considered equally and the analytical computational power presented by rule-based ML. An additional benefit of rule-based learning methods is that they produce human-readable rules (Figure 1A), in contrast with other methods such as nonlinear SVM and artificial neural networks. It is not trivial to interpret the predictions using this latter method as complex mathematical model output can be difficult to comprehend.

ML has been employed previously to analyze transcriptomic data in diverse fields ranging from cancer (Hampton and Frierson, 2003; Quackenbush, 2006; Glaab et al., 2009; van der Vegt et al., 2009) to plant science research (Kell et al., 2001; Li et al., 2006). These approaches have identified genes preferentially expressed in discrete developmental states, such as healthy tissue versus tumorous tissue (Dagliyan et al., 2011). Despite the ability to identify differentially regulated transcripts between developmental states, relationships between these genes cannot be inferred using these ML approaches. This is because these previous approaches do not use an associative measure between genes within the models generated. The resulting models are therefore only capable of examining genes on an individual basis, leading to the identification of their differential expression, but not the associations between them. The use of model trees has been proposed as a means to functionally associate genes (Nepomuceno-Chamorro et al., 2010). This ML approach can be applied to regression problems and has yet to be validated experimentally.

Here, we propose a novel approach in the construction of functional networks based on gene expression data, through the use of rule-based ML. The premise of this approach is that genes present within the same rule that predicts a developmental outcome will have an increased likelihood of being functionally related in the developmental process in question given their collective ability to act together in generating the prediction. We term this associative measure “coprediction.” The use of coprediction enables functional gene associations to be inferred that cannot be detected using coexpression analysis, for example. Coprediction is not restricted by similarities in expression pattern that are the only measure used by coexpression. Another difference between these two methods stems from rule-based methods focusing on identifying interactions within subsets of samples, such as those belonging to the discrete developmental states or transitions between them, rather than correlations across all samples, as is the case with coexpression. In this way, state-dependant data can be considered independently from one another and novel knowledge extracted from the data.

As an experimental system, seed germination is well suited to assess coprediction for network construction. The decision to complete germination is a binary and irreversible developmental
A Example Germination Rule

If At1g27595>100.87 and At3g49000>68.13 and At2g40475>55.96
→ Predict germination

Example Germination Rule Set

If At1g27595>100.87 and At3g49000>68.13 and At2g40475>55.96
→ Predict germination
If At4g34710>349.67 and At4g37760>150.75 and At1g30135>17.66
→ Predict germination
If At3g03050>37.90 and At2g20630>96.01 and At3g02865>9.66
→ Predict germination
If At1g54910>45.03 and At4g18975>16.74 and At3g28910>52.76
and At1g48320>56.80
→ Predict germination
Everything else
→ Predict non-germination

Example Non-Germination Rule Set

If At2g31350>241.18 and At1g76460>162.81
→ Predict non-germination
If At4g12600>147.90 and At1g79210>520.68
→ Predict non-germination
If At3g02080>1433.57 and At1g73500>57.19
→ Predict non-germination
If At2g28700>88.90 and At3g14930>102.58 and At5g64390>68.64
→ Predict non-germination
Everything else
→ Predict germination

B

Figure 1. Generation of a Rule-Based ML Coprediction Network Based on Arabidopsis Seed Microarray Data.

(A) An example rule and two example rule sets predicting the germination and non-germination developmental outcomes in Arabidopsis seeds. The example rule represents the first rule within the example germination rule set. Within each rule is an Arabidopsis gene identifier followed by the > operator followed by a number, representing a gene expression level. (B) Pipeline used to generate the coprediction functional gene network based on rules produced through rule-based ML. The associated software can be downloaded at www.vseed.nottingham.ac.uk.

RESULTS AND DISCUSSION

Predicting Seed Germination Using ML

We used as the input data, or training set, for this study 122 publicly available microarray data sets generated from imbibed Arabidopsis seeds (Ogawa et al., 2003; Yamauchi et al., 2004; Nakabayashi et al., 2005; Cadman et al., 2006; Penfield et al., 2006; Finch-Savage et al., 2007; Bassel et al., 2008; Carrera et al., 2008) and genetic resources in the model plant Arabidopsis (Alonso et al., 2003) make it possible to produce computational predictions in silico that can be validated in vivo. A seed is said to have completed germination when the embryonic root emerges through the surrounding structures of the seed, while a seed is dormant when it fails to germinate under otherwise favorable conditions (Bewley, 1997). The decision for a seed to maintain or terminate dormancy and commence germination is a complex trait that determines where and when plants enter into ecosystems (Bewley, 1997; Holdsworth et al., 2008). Numerous key regulators of seed germination have been revealed using forward genetic screens (Holdsworth et al., 2008), and the means by which they collectively act to control this developmental transition is being uncovered (Holdsworth et al., 2008; Bassel et al., 2011).

In this study we aimed to (1) investigate whether rule-based ML methods can be used to predict the developmental output of a biological system, (2) identify novel regulators using these predictions, and (3) uncover functional associations between genes controlling a developmental phase transition using this approach. We used the large number of publicly available transcriptomic data sets representing each the dormant and germinating states of Arabidopsis seeds (Bassel et al., 2008) to generate a functional network using rule-based ML. Data presented show this network to represent an accurate model of seed germination capable of predicting both novel regulators and functional associations between genes independent of their respective expression patterns.
according to the developmental status of the seed as either a nongerminating or germinating sample (Bassel et al., 2008) and represented the binary annotation used as the class label for this training set. Additionally, genes were filtered to remove those with low expression such that a total of 13,942 genes were included in further analyses (Bassel et al., 2011).

Four different ML methods representing different types of prediction techniques were compared to determine the accuracy with which they can predict developmental fate in seeds. The methods tested were Naïve Bayes (John and Langley, 1995), C4.5 (Quinlan, 1993), SVMs (Vapnik, 1995), and the rule-based method bioinformatics-oriented hierarchical evolutionary learning (BioHEL) (Bacardit et al., 2009a). These represent diverse and robust ML algorithms. Predictive accuracy was evaluated using stratified 10-fold cross-validation that previously was shown to work well for transcriptome data (Molinaro et al., 2005). All methods examined were capable of predicting developmental outcome accurately, ranging from 79.8 to 93.5% (Table 1). The BioHEL algorithm produced predictions with the highest average accuracy of 93.5%, demonstrating this rule-based method to be robust with respect to other ML methods in the prediction of seed developmental fate. Thus, we can be confident that the rules from which we will construct functional networks are reliable.

BioHEL was developed to analyze large-scale biological data sets through the generation of predictive models using rules (Bacardit et al., 2009a). In this study, using the annotated seed germination gene expression training set, a rule consists of two or more genes and a condition (gene expression level) associated with each gene (Figure 1A). All conditions within a given rule must be satisfied in order for the rule to predict the developmental outcome. In the example rule given in Figure 1A, all three genes must be above the respective expression level determined by the algorithm to predict a seed will germinate (At1g27595>100.87, At3g49000>68.13, and At2g40475>55.96).

The BioHEL algorithm generates rules one by one using a mechanism known as separate-and-conquer (Furnkranz, 1999). Each rule is generated with the aim of predicting the outcome with the greatest accuracy possible, while at the same time using as many of the samples within the training set as possible. A genetic algorithm (Goldberg, 1989) is employed to generate each rule. Once a rule is produced, all of the samples that were used are removed and no longer available for the generation of subsequent rules. Afterwards, the process starts again to learn the next rule. This iterative learning and rule generation process ends when all samples in the data set have been classified. The end product of this process is a series of rules collectively termed a “rule set” (Figure 1A). BioHEL’s rule learning process is represented in Supplemental Figure 1 online. The example germination rule set presented in Figure 1A consists of four different rules.

A final default rule is included within each rule set, such that the outcome of all samples in the data set not covered by the other rules are predicted (everything else → predict nongermination) (Figure 1A). We will refer to this type of rule set as a germination rule set in this manuscript. Conversely, a rule set consisting of rules predicting nongermination and a default rule predicting germination is referred to as a nongermination rule set (Figure 1A).

We tested whether the prediction accuracy using Arabidopsis seed transcriptomics training sets would be affected by changing the default rule from nongermination to germination (and generating rules predicting nongermination). Using this setting, BioHEL obtained an accuracy of 92.4% ± 1.5%, which is only slightly lower than using germination rule sets at 93.5% ± 1.0% (Table 1). BioHEL therefore predicts either germination or nongermination with approximately equal accuracy.

Next we investigated whether the prediction accuracy of developmental outcome using BioHEL was robust with respect to the sample annotation or due to random chance. This was achieved by randomizing the assignment of labels to the samples in the training set such that the distribution of labels remained equal and then repeating the learning process. With the randomly labeled samples, a much lower prediction accuracy of 49.2% ± 7.2% was obtained for germination rule sets and 54.4% ± 5.0% for nongermination rule sets. This shows that the ability of BioHEL to accurately predict the developmental outcome of seeds is robust and dependent on the accurate annotation of samples as belonging to either the germinating or nongerminating state.

**Table 1. Predictability of Germination Outcome Based on Gene Expression for Various ML Methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioHEL-germination</td>
<td>93.5 ± 1.0</td>
</tr>
<tr>
<td>BioHEL-nongermination</td>
<td>92.4 ± 1.5</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>88.0 ± 2.4</td>
</tr>
<tr>
<td>C4.5</td>
<td>79.8 ± 3.6</td>
</tr>
<tr>
<td>SVM</td>
<td>82.4 ± 0.4</td>
</tr>
</tbody>
</table>

**Extraction of Knowledge from Rule Sets Generated Using the BioHEL Algorithm**

Despite a given rule set generated using BioHEL being very accurate, it does not represent the only possible model predictor of developmental outcome that can be extracted from the data set. Each time the BioHEL algorithm goes through the learning process and generates a rule set, variations are expected within the predictive model generated due to its stochastic nature. The examination of multiple rule sets generated following repeated independent learning processes reveals that some genes appear with a greater frequency than others. Collating the results of multiple repetitions of rule set generation allows for the identification of the genes that appear more frequently and represent the best predictors of developmental fate and the highest confidence candidate regulators of the biological process. The interpretation of rules generated by BioHEL in this way considers individual genes as regulators, representing a simple way of extracting knowledge from the rules.

**Functional Association Network Generation Using Rule-Based ML**

BioHEL was applied to the seed transcriptomic training set to generate 10,000 germination rule sets and an additional 10,000 nongermination rule sets. These rule sets were used to associate genes functionally and produce a network (Figure 1B).
Two node scores were assigned to each gene based on the frequency of its appearance within each rule set (germination and nongermination; see Supplemental Table 2 online). Many of the genes with the highest node scores, representing the highest confidence candidate regulators of seed germination and nongermination, previously have been demonstrated to be involved in the control of this developmental transition (Table 2). Of these previously described regulators with high node scores, more were characterized by high nongermination node scores than high germination node scores.

The accurate prediction of developmental fate using rule-based ML depends on the condition associated with each gene within a rule being satisfied. It is therefore not the critical expression level of an individual gene within a rule that makes it accurate, but the collective expression levels of all genes present within the rule. In this way, genes that come together in a rule to predict a chosen developmental outcome can be associated with one another and edges established between them given their collective prediction capacity. The establishment of connections between genes coappearing within a rule represents the basis of generating a coprediction functional association network.

The strength of the connections between genes, termed the edge weight, was calculated based on the frequency with which gene pairs coappear within each the nongermination and germination rule sets and the frequency with which individual genes appeared within the rules sets. Point-wise mutual information was used to quantify these associations (Tsuruoka et al., 2008), as it normalizes co-occurrences of gene pairs with the frequency the individual genes appear within the rules. The point-wise mutual information scores can be used to rank the edges and prioritize them for the investigation of putative functional associations between their corresponding genes. The resulting coprediction network generated using BioHEL with the Arabidopsis seed gene expression training set was termed “Seed Co-Prediction Network” (SCoPNet) and consisted of 13,532 nodes and 146,933 edges (Figure 2A).

Transcripts that are developmentally upregulated by the nongerminating and germinating states in Arabidopsis seeds previously have been identified through the use of the SAM algorithm (Tusher et al., 2001; Bassel et al., 2011). These lists were termed SAM Nongermination (SAM NG) and SAM Germination (SAM G), respectively, and are a statistically robust list of genes that are significantly associated with each of these developmental states (see Supplemental Data Set 1 online). The developmental status of the SAM-defined genes represented by nodes in SCoPNet was indicated by color, based on the classification by each of these SAM gene lists and examining their distribution. SAM G and SAM NG genes were concentrated in different domains of SCoPNet (Figure 2A); this ML-based network therefore captures the state-dependent gene interactions associated with the developmental states of nongermination and germination in Arabidopsis seeds. The state-dependent distribution of SAM genes is lost in the network when the annotation labels of the samples are randomly assigned (see Supplemental Figure 2 online). This demonstrates that developmental association of these domains in SCoPNet is not due to random chance.

<table>
<thead>
<tr>
<th>AGI</th>
<th>Annotation</th>
<th>Node Score</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>At2g28350</td>
<td>ARF10</td>
<td>206</td>
<td>22</td>
</tr>
<tr>
<td>At3g24220</td>
<td>NCED6</td>
<td>159</td>
<td>1</td>
</tr>
<tr>
<td>At2g04240</td>
<td>XERICO</td>
<td>112</td>
<td>8</td>
</tr>
<tr>
<td>At3g62090</td>
<td>PIL2</td>
<td>106</td>
<td>6</td>
</tr>
<tr>
<td>At5g07200</td>
<td>Gibberellin 20-oxidase3</td>
<td>104</td>
<td>12</td>
</tr>
<tr>
<td>At1g33060</td>
<td>ANAC014</td>
<td>100</td>
<td>19</td>
</tr>
<tr>
<td>At1g03790</td>
<td>SOMNUS</td>
<td>81</td>
<td>13</td>
</tr>
<tr>
<td>At2g26300</td>
<td>G Protein Alpha Subunit1</td>
<td>80</td>
<td>14</td>
</tr>
<tr>
<td>At1g04950</td>
<td>AtGA2ox2</td>
<td>80</td>
<td>3</td>
</tr>
<tr>
<td>At3g45640</td>
<td>AtMPK3</td>
<td>76</td>
<td>7</td>
</tr>
<tr>
<td>At3g24650</td>
<td>ABI3</td>
<td>68</td>
<td>14</td>
</tr>
<tr>
<td>At1g09570</td>
<td>PHY A</td>
<td>67</td>
<td>11</td>
</tr>
<tr>
<td>At1g55255</td>
<td>HUB2</td>
<td>53</td>
<td>16</td>
</tr>
<tr>
<td>At5g25900</td>
<td>GA3</td>
<td>53</td>
<td>14</td>
</tr>
<tr>
<td>At4g25420</td>
<td>GA5</td>
<td>50</td>
<td>36</td>
</tr>
<tr>
<td>At2g18790</td>
<td>PHYB</td>
<td>48</td>
<td>76</td>
</tr>
<tr>
<td>At1g50420</td>
<td>SCL3</td>
<td>47</td>
<td>72</td>
</tr>
<tr>
<td>At1g01360</td>
<td>PYL9</td>
<td>46</td>
<td>67</td>
</tr>
<tr>
<td>Known Regulators in Nongermination Rules</td>
<td>Nongermination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At1g09570</td>
<td>PHY A</td>
<td>67</td>
<td>11</td>
</tr>
<tr>
<td>At1g55255</td>
<td>HUB2</td>
<td>53</td>
<td>16</td>
</tr>
<tr>
<td>At5g25900</td>
<td>GA3</td>
<td>53</td>
<td>14</td>
</tr>
<tr>
<td>At4g25420</td>
<td>GA5</td>
<td>50</td>
<td>36</td>
</tr>
<tr>
<td>At2g18790</td>
<td>PHYB</td>
<td>48</td>
<td>76</td>
</tr>
<tr>
<td>At1g50420</td>
<td>SCL3</td>
<td>47</td>
<td>72</td>
</tr>
<tr>
<td>At1g01360</td>
<td>PYL9</td>
<td>46</td>
<td>67</td>
</tr>
<tr>
<td>Known Regulators in Germination Rules</td>
<td>Germination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At1g09570</td>
<td>PHY A</td>
<td>67</td>
<td>11</td>
</tr>
<tr>
<td>At1g55255</td>
<td>HUB2</td>
<td>53</td>
<td>16</td>
</tr>
<tr>
<td>At5g25900</td>
<td>GA3</td>
<td>53</td>
<td>14</td>
</tr>
<tr>
<td>At4g25420</td>
<td>GA5</td>
<td>50</td>
<td>36</td>
</tr>
<tr>
<td>At2g18790</td>
<td>PHYB</td>
<td>48</td>
<td>76</td>
</tr>
<tr>
<td>At1g50420</td>
<td>SCL3</td>
<td>47</td>
<td>72</td>
</tr>
<tr>
<td>At1g01360</td>
<td>PYL9</td>
<td>46</td>
<td>67</td>
</tr>
</tbody>
</table>

Regulatory genes displayed are present within the top 2.5% of node scores for each nongermination and germination. AGI, Arabidopsis Genome Initiative.
The distribution of the genes with the greatest germination and nongermination node scores were examined within SCoPNet. The nodes with the greatest nongermination node strength were concentrated within the domain associated with SAM NG genes (Figure 2B). Conversely, the domain of the network associated with the developmental outcome of seed germination and SAM G genes contained the nodes with the highest germination node scores (Figure 2C). The distribution of gene co-occurrence frequencies in each of the germination and nongermination predicting rule sets within SCoPNet was examined. The frequency of gene pairs within each of these classes of rule set was determined, and edges within SCoPNet colored with...
increasingly darker shades of blue with increasing co-occurrence frequency (Figures 2B and 2C). The domain of SCoPNet associated with SAM NG genes and high nongermination node scores also contained gene pairs occurring within the nongermination-predicting rule sets (Figure 2B). Conversely, the germination-associated domain of SCoPNet contained an abundance of SAM G genes, genes with a high germination node score, and a high frequency of genes co-occurring within the rule sets predicting an outcome of germination (Figure 2C).

The two domains of SCoPNet that are associated with developmentally regulated genes as demonstrated using the SAM gene lists are also associated with the same coprediction node scores and co-occurrence frequencies predicting nongermination and germination, respectively. The generation of a network demonstrating developmentally regulated domains of genes and associations between the genes in Arabidopsis seeds reflects the accurate reconstruction of functional gene associations using coprediction. This provides support for the robust nature of this network model as a means to probe predicted uncharacterized putative regulatory genes and functional links between genes associated with each of these two developmental fates.

Given that an equal number of germination and nongermination rule sets were generated, we compared the absolute node strengths for each of these predicted developmental outcomes. The highest scoring nongermination nodes had stronger node scores than the highest scoring germination nodes (Figure 2D). This indicates that individual genes in Arabidopsis seeds have a greater capacity to predict a nongerminating developmental fate than germination.

The distribution of nodes with the greatest degree (number of connections) was examined within SCoPNet and found to be concentrated within the germination domain of the network (Figure 2E). The greater number of connections between the germination-associated nodes within SCoPNet indicates there are a greater number of different combinations of genes within 10,000 rule sets predicting a developmental outcome of germination and fewer different combinations of genes predicting nongermination.

Examination of the coexpression network SeedNet revealed there to be a greater state of transcriptional coordination in the nongerminating state that during germination in Arabidopsis seeds (Bassel et al., 2011). This conclusion was based on the nongermination domain of the graph containing the highest order nodes (hubs) of the network, indicating the greatest coordination of cohorts of transcripts during this developmental state. Analysis of SCoPNet leads to a similar conclusion albeit for different reasons. Within the nongermination domain of SCoPNet, individual genes act as better predictors of germination based on their greater nongermination node scores than the equivalent germination node scores within the germination domain (Figure 2D). Additionally, the lower connectivity of the nongermination domain (Figure 2E) indicated that a lower diversity of genes was used by BioHEL to accurately predict the nongermination developmental fate. Therefore, individual genes more accurately predict a developmental fate of nongermination in Arabidopsis seeds demonstrating a greater coordination of the predicted functional network controlling this state.

Comparison of the Coprediction Network SCoPNet with the Coexpression Network SeedNet

The gene expression data from Arabidopsis seeds used as the training set to compute SCoPNet has also been used previously to calculate the genome-wide gene coexpression network SeedNet (Bassel et al., 2011). We compared these two networks generated by two different approaches using the same data to evaluate the outputs of these methods.

A total of 580 nodes had at least one shared connection between SCoPNet and SeedNet (4.3% of the SCoPNET), connected by 356 shared connections (0.01% of SCoPNet) (Figure 2F). The intersection network captured by both of these approaches is represented primarily by genes whose expression is associated with the nongerminating state and the SAM NG gene list (red nodes, Figure 2F). The previously characterized regulators of germination HUB2 (Liu et al., 2007), HAB1 (Saez et al., 2004), and FLC (Chiang et al., 2009) share common edges within the largest portion of this shared network. Other known regulators, including AGL67, AGD2, and ANAC014 (Bassel et al., 2011), are also present within smaller orphan networks. The low overlap between the networks generated using these two different approaches indicates that few gene associations established using coprediction represent coexpressed gene pairs. This reflects the differences in methodology used for network generation.

Genes with the 100 highest nongermination and 100 highest germination node scores from SCoPNet were plotted within the coexpression SeedNet to examine the distribution of genes identified using ML. Domains of the SeedNet graph previously have been shown to be strongly associated with discrete developmental states in seeds (Bassel et al., 2011). Genes with the highest germination node scores were strongly associated with the domain of SeedNet associated with SAM G genes and seed germination (Figure 2G). Genes with the greatest nongermination node scores were primarily associated with the domain of SeedNet associated with SAM NG and nongermination, yet also were sparsely present within the germination domain.

Similarities between SCoPNet and SeedNet are observed through the significantly represented Gene Ontology (GO) categories that are present within the two developmentally distinct domains of this coprediction network (Maere et al., 2005). The nongermination domain of SCoPNet included genes associated with seed dormancy, chromatin modification, and response to abiotic stress (Figure 3A), consistent with the significant GO terms identified in the nongerminating domain of SeedNet (Bassel et al., 2011). Similar to the germination domain of SeedNet, GO categories relating to cellular expansion are present within the germination domain of SCoPNet, including cell wall biogenesis and modification. Metabolic processes associated with seedling establishment are also present in this domain (Figure 3B). Additionally, genes involved in cellular differentiation and phase transitions are also present within the germination domain of SCoPNet.

These distinct GO categories are consistent with the processes occurring during each of these developmental states in seeds and further support the finding that the two domains of SCoPNet capture discrete developmental states.

In SCoPNet, a greater number of previously published regulatory genes were characterized by high nongermination node
Figure 3. Significantly Represented GO Biological Process Categories within the Nongermination and Germination Domains of SCoPNet.

(A) Significant GO categories within the nongermination domain of SCoPNet.
(B) Significant GO categories within the germination domain of SCoPNet.

A greater node size indicates more genes within a given GO category. Node color indicates the P value significance using the scale from yellow to orange in the bottom left of (A) and (B). A threshold of $P < 0.05$ was used to identify significant GO categories.
scores and associated with the nongermination domain than by high germination node scores and associated with the germination domain (Table 2) (Bassel et al., 2011). This is also the case with SeedNet, where this distribution was a consequence of the transcripts of these known regulatory genes having a greater abundance in the nongerminating state. In the case of the coprediction network SCoPNet, this distribution is due to these previously characterized regulatory genes collectively predicting the developmental fate of nongermination.

Despite the lack of common gene associations established using each coprediction in SCoPNet and coexpression in SeedNet (Figure 2F), common topological properties of these two different networks capturing seed germination remain. One common element is the greater abundance of known regulatory genes within the nongermination domains of these networks (Table 2). Consistent with the greater abundance of key regulatory genes associated with the nongermination domain is the greater absolute nongermination node scores in SCoPNet compared with the equivalent germination node scores (Figure 2D). Nodes of the highest degree were found in the nongermination region of SeedNet, indicating there to be greater transcriptional coordination in this region. The highest degree nodes in SCoPNet were within the germination domain. This indicates a greater number of different gene combinations were needed within the rules to accurately predict this outcome and that more coordinated processes are capable of predicting nongermination than germination.

Finally, similar GO ontology categories are overrepresented by the germination and nongermination domains of SCoPNet and SeedNet, respectively, highlighting the capture of common processes by these different gene association approaches.

**SCoPNet Predicts Novel Regulators of Germination**

To evaluate the predictive capacity of SCoPNet in uncovering novel regulators of seed germination, we examined the phenotypes of seeds carrying mutations in genes that carry both high node scores and a high degree within the network (see Supplemental Table 3 online). A total of 24 homozygous insertion lines representing 17 high confidence candidate regulators of germination based on SCoPNet were identified and examined for germination-related phenotypes. Insertion lines corresponding to four of these genes showed altered germination responses, representing a 24% accuracy (i.e., 4 out of 17) in the identification of novel germination regulators using this approach. These newly characterized germination-regulating genes were termed ALTERED SEED GERMINATIONS (ASG5, ASG6, ASG7, and ASG8 (see Supplemental Table 4 online).

Insertional mutants of these genes showed altered germination behavior in the presence of the germination inhibiting hormone abscisic acid (ABA) (Figures 4A, 4C, 4E, and 4G), and the gibberellin (GA) and ethylene synthesis inhibiting compound paclobutrazol (PAC) (Rademacher, 2000) (Figures 4B, 4D, 4F, and 4H).

ASG5, ASG6, and ASG7 were selected based on their high degree within the network and high nongermination node scores. All three of these genes are located within the nongermination associated region of SeedNet and are present on the SAM NG gene list (see Supplemental Figure 3 online) (Bassel et al., 2011).
Conversely, ASG8 had a high degree and high germination node score and is present with the germination-associated region of SeedNet and on the SAM G gene list.

ASG5 (At1g20650) encodes an uncharacterized Ser/Thr kinase that acts to inhibit seed germination. ASG6 (At1g70520) is a Cys-rich receptor-like kinase that also acts to inhibit germination. Both ASG7 (At5g47580) and ASG8 (At2g40475) encode proteins of unknown function that repress and promote seed germination, respectively.

SCoPNet Predicts Functional Associations between Genes Controlling Arabidopsis Seed Germination

We investigated the relationship between previously characterized and newly uncovered regulators of germination identified in this study within SCoPNet. None of the previously published regulatory factors (see Supplemental Data Set 1 online) that are connected in SeedNet share these same predicted associations in SCoPNet (Figure 2F).

Central to the connections between known regulatory genes within the nongermination domain of SCoPNet are the germination regulatory genes ASG5, ASG6, and ASG7 identified in this study (Figure 5A). All three of these genes are connected to each other and to the key dormancy-regulating locus ABI3. The relationship between ASG5, ASG6, and ASG7 and ABI3 was investigated by examining the transcript abundance of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulated in abi3-4 mutant seeds (Figure 5B). The ABI3 gene is therefore involved either directly or indirectly in the regulation of the expression of these newly characterized germination regulators within previously published microarray data performed using abi3-4 mutant seeds (Carrera et al., 2008). All three of these ASG transcripts are downregulat.
ASG6 is strongly associated with ABA-related genes, including the ABA receptor PYL9 and ABA homeostasis regulator XERICO in addition to the key GA synthesis gene GA3 (Figure 5A). Expression of ASG7 is inhibited by GA in Arabidopsis seeds (Figure 5C), suggesting a functional association between this hormone synthesis gene and newly characterized regulator of germination.

The newly characterized ASG8 gene is connected to several known regulatory genes within the germination domain of SCoPNet (Figure 5E). The ABA response regulator XLG2 (Ding et al., 2008) is connected to ASG8 along with the ABA receptor PYL2, to which XLG1 is also connected. ASG8 represents an endosperm-enriched transcript during germination with expression commencing in the seedling following radicle protrusion (Figure 5F).

These observations demonstrate that coprediction can identify both novel regulators and functional associations between genes controlling developmental fate in Arabidopsis seeds.

Associations Established Using Coprediction Are Not Restricted to Genes Sharing Common Expression Patterns

Connections established using coprediction are not restricted to genes sharing common expression patterns, the measure used with coexpression analysis. On the contrary, functionally related genes that are connected within SCoPNet often have divergent expression patterns. This can be observed by looking at the expression patterns of each ABI4 and ABI3, which are linked to each other within this network (Figure 5E). The ABI4 transcription factor was identified as a response regulator to the hormone ABA, which is synthesized by the protein product of the ABI3 gene (Nambara and Marion-Poll, 2003). Over a time course of seed germination, the ABI3 transcript sharply declines, while the ABI4 transcript is induced (Figure 6A). These genes therefore have opposite transcriptional regulation during the developmental transition of seed germination, yet affect the same developmental outcome. Coprediction captures this functional association.

Within the nongermination domain of SCoPNet (Figure 5A), the ethylene response regulator EIN3 is connected to the DELLA gene RGL3. The transcripts of these two genes also show diverging expression patterns over a time course of seed germination (Figure 6B). The two newly identified regulators of germination ASG5 and ASG7 are connected in SCoPNet and show variation in their expression pattern over a time course of seed germination (Figure 6C). The ABA response and synthesis genes SAD1 and ABA regulatory factor SOMNUS, ABA synthesis gene β-Hydroxylase1 and ABA receptor PYL4, and the ABA response regulator ABI3 and ABA receptor PYL9 are also connected to each other yet show different transcriptional regulation (Figures 6D to 6F). These examples highlight the ability to establish functional associations between genes using coprediction that are not manifest through common expression patterns.

Statistically Significant cis-Elements Are Enriched within Modules of SCoPNet

Clustering of SCoPNet using MCODE (Bader and Hogue, 2003) led to the identification of 44 modules, representing significantly interconnected groups of genes (see Supplemental Data Set 2 online). Previously characterized regulatory genes (see Supplemental Data Set 1 online) were present within modules 2, 4, 5, 6, 7, 8, and 44 (see Supplemental Table 5 online). Module 2 contains an abundance of ABA-related genes and the GA synthesis genes GA4 and GA5. Module 4 contains both of the functionally redundant MYB33 and MYB101 transcription factors that together act to modulate germination in response to ABA.
(Reyes and Chua, 2007). This association is established between these two transcription factors despite their divergent expression pattern in seeds (Figure 6G). Similarly, the ABA response transcription factors ABI3 and ABI4, present in module 6, both function to regulate the response of seeds to the germination inhibiting hormone ABA yet exhibit divergent expression patterns (Figure 6H). These modules highlight the associative power of coprediction and its ability to establish connections between genes in the same developmental pathway independently of their gene expression profiles.

The modules identified by MCODE were examined to establish whether enriched cis-elements exist within the promoters of their constituent genes (O’Connor et al., 2005). Modules 2, 4, and 5 were found to contain significantly overrepresented cis-elements (see Supplemental Table 6 online).

Transcriptionally coordinated genes often contain overrepresented cis-elements within their promoters as shared transcription factors act to coordinate their common expression pattern. The finding that the promoters of genes within modules in SCoPNet also contain enriched cis-elements is intriguing as these genes were not associated based on common expression profiles, nor do they share common expression profiles. This suggests that developmentally coordinated processes that are mediated through common cis-elements and transcription factors, yet not observable at the level of common transcript abundance, have been captured through coprediction. This is supported by the finding that homologous transcription factors with the same DNA binding site can act as both activators and repressors of gene expression. The bZIP transcription factors ENHANCED EM LEVEL (EEL/bZIP12) and ABIS repress and

**Figure 7.** Screenshot of the Online Network Query Tool Generated in This Study to Query SCoPNet.

The seed germination regulatory gene RGL2 was queried using the gene name in the query box and is highlighted within the network view window. SCoPNet is available at http://www.vseed.nottingham.ac.uk/.
activate the expression of At-Em1, respectively (Bensmihen et al., 2002). Both of these proteins bind the same ABRE cis-element within the At-Em1 promoter and antagonistically fine-tune the expression of this transcript. The ABRE cis-element was also identified as statistically overrepresented within the promoters of the genes within modules 2, 4, and 5 of SCoPNet. Coprediction may therefore capture biological processes under the control of common regulatory elements using gene expression data that are not manifest at the level of coordinated transcription.

Development of an Online Web Tool to Query SCoPNet

A Web-based community resource has been developed enabling users to query SCoPNet at www.veed.net/nottingham.ac.uk/ (Figure 7). This tool is based on the WIGis visualization framework (www.wigis.net) (Gretarsson et al., 2009) and enables the position of either individual genes or lists of genes to be identified within the network. The first neighbors of a selected gene may in turn be highlighted. This network query tool has been integrated with other online Web resources, including the Seed eFP Browser at the BAR website (Winter et al., 2007; Bassel et al., 2008) and the cis-element discovery program ATHENA (O’Connor et al., 2005), to maximize the utility of this tool within the context of other Web-based seed resources.

The “How to Generate SCoPNet” link on this site provides complete instructions on how to install and implement BioHEL and links to download the associated software.

Perspectives and General Utility of Rule-Based ML with Diverse Data Types

The use of rule-based ML in the prediction of functional associations between variables is not restricted to microarray data. This approach can be used with any type of biological data, so long as the labels annotating the samples are finite. BioHEL has previously been applied to elucidate protein structure prediction problems (Stout et al., 2008, 2009; Bacardit et al., 2009b) and can also be used with the quantitative data generated using either next-generation sequencing or proteomics methodologies.

The BioHEL rule-based learning approach is as well not restricted to the binary class labels used in this study, and multiple classes can be used to annotate training sets. This enables the associations between variables controlling multiple developmental fates or biological outputs to be elucidated.

Conclusion

Here, we present coprediction as a powerful novel associative means for the investigation of gene function and prediction of functional networks using gene expression data from Arabidopsis. This computational methodology represents a useful alternative approach for the extraction of biological knowledge from existing data that other approaches are not capable of inferring. This technique will serve in diverse areas of plant biology in the elucidation of functional networks and increase the return on the collective investment made by the research community in the generation of large-scale data sets.

METHODS

Microarray Data Compilation and Normalization

Gene expression data from Arabidopsis thaliana seeds generated using the Affymetrix ATH1 microarray platform were collated as previously described (Bassel et al., 2008), normalized using GCOS/MASS with a TGT value set to 100. Only data from imbibed seeds were used, representing a total of 122 arrays with 53 capturing the germinating state and 69 the nongerminating state (see Supplemental Table 1 online). Genes not expressed at least once above the level of 100 expression units (representing 5 times greater expression than the background level of 20 units) were removed, leaving 15,942 genes used in the analysis.

Accuracy Estimation Using Stratified 10-Fold Cross-Validation

Many methodologies exist in ML to estimate the prediction capacity of a method on a particular data set. The most widespread of them, which is known to be suitable for microarray data, is stratified 10-fold cross-validation (Molina et al., 2005). This methodology randomly partitions the set of samples into 10 strata, making sure that each strata presents the same distribution of germination and nongermination samples as the whole set. Afterwards, it generates 10 pairs of training and test sets. For each pair, one of the 10 strata is used as test set and the other nine as training set. Thus, each stratum is used once as test set and nine times as part of a training set. Afterwards, each ML method is trained using the 10 training sets, and the learn models are evaluated using the corresponding test sets. The prediction capacity of each ML method is estimated as the average accuracy (number of correctly classified samples/total number of samples) over the 10 test sets. Given the low number of samples in the data set, it is recommended to repeat the cross-validation process several times with different strata. In our case, we performed 10 repetitions.

ML

BioHEL is a ML system that follows the separate-and-conquer rule learning paradigm employing a genetic algorithm to learn each individual rule, as described in Results. BioHEL has been designed specifically to cope with large-scale data sets, incorporating mechanisms to deal with large numbers of variables, such as the attribute list knowledge representation (Bacardit et al., 2009a), as well as with a high number of samples, such as the incremental learning with alternating strata (ILAS) (Bacardit et al., 2004; Stout et al., 2009). Due to the stochastic nature of BioHEL, each run of the system generates different results, a fact that is exploited by determining the consensus prediction (by a simple majority vote) from rule sets generated by independent runs of the system. This mechanism has been shown to improve BioHEL’s performance on most data sets. For all the BioHEL experiments reported in this article, we employed 500 iterations of the genetic algorithm, a coverage breakpoint of 0.1 and two strata for ILAS. An ensemble of 100 rule sets was employed for the cross-validation experiments. All other parameters were set to their default values (Bacardit et al., 2009a). BioHEL can be downloaded at http://www.infobiots.org/.

For comparison purposes, we employed three other well-known classification methods: Naïve Bayes (John and Langley, 1995), C4.5 (Quinlan, 1993), and SVMs (Vapnik, 1995). For C4.5 and Naïve Bayes, we used the WEKA implementation (Hall et al., 2009). For SVM, we used the LIBSVM implementation (http://www.csie.ntu.edu.tw/~cjlin/libsvm). Default parameters were employed for all three methods.

Point-wise mutual information for edge strength was calculated by dividing the number of co-occurrences of the single genes in the rules, p(x,y), by the product of the total number of occurrences of the single genes, p(x)p(y), and taking the logarithm of this relation (Tsuruoka et al., 2008): $\text{pmi}(x,y) = \log(p(x,y)/p(x)p(y))$. 
Cluster Identification Using MCODE

The MCODE clustering algorithm plugin for Cytoscape (Bader and Hogue, 2003) was used to identify modules within SCOPNet. Clustering was performed with a degree cutoff of 2, node score cutoff of 0.2, k core equal to 2, and the maximum depth of 100.

Promoter Motif Identification

Significant enrichment of previously characterized cis-element regulatory motifs within the promoters of genes present in clusters identified by MCODE was performed using the ATHENA Web tool (O’Connor et al., 2005). The analysis suite option of this Web tool was used to examine 1000 upstream bases for each gene promoter unless an adjacent gene was present within this sequence, at which point the sequence was cut off. A hypergeometric background model of known cis-element frequencies in the genome was used with a P value cutoff threshold of 10\(^{-4}\) following a Bonferroni correction.

GO Term Enrichment Analysis

Significantly enriched GO categories were identified using the BiNGO GO Term Enrichment Analysis following a Bonferoni correction.

GO Term Enrichment Analysis

Significantly enriched GO categories were identified using the BINGO plugin for cytoscape (Maere et al., 2005). A P value significance was calculated using a hypergeometric test with a Benjamin and Hochberg false discovery rate correction. A threshold of P < 0.05 was used as the threshold for significantly enriched GO categories.

Plant Materials

All Arabidopsis seed lots were obtained from the Nottingham Arabidopsis Stock Centre (University of Nottingham, UK). Plants were grown to maturity in controlled environment rooms using 16 h light (light intensity 8 \(\mu\)mol m\(^{-2}\) s\(^{-1}\)) at 23°C and 70% relative humidity/8 h dark at 18°C and 80% relative humidity. When plants had ceased flowering and siliques began to brown, seeds were harvested, cleaned through a 500-mesh, and stored at 24°C in glassine bags in the dark for 1 month to remove primary dormancy.

Identification of Homozygous T-DNA Insertion Lines

Identification of homozygous insertion lines (Alonso et al., 2003) was performed using 100 ng of genomic DNA as template in a three primer PCR reaction using a 57°C annealing temperature and 35 cycles. A list of primers used in this study can be found in Supplemental Table 7 online.

Germination and Seedling Establishment Conditions

All germination analyses were performed with seeds obtained from plants grown at the same time within the same tray within the same controlled environment chambers to minimize differences in postharvest history. Prior to germination, seeds were surface-sterilized in 5% (v/v) bleach for 5 min and then washed three times in sterile water (Holman et al., 2009). Seeds were pipetted onto Petri plates containing 0.7% (w/v) agarose (type PGP; Park Scientific) and the appropriate hormone, stratified for 48 h at 4°C, and then incubated at 22°C under continuous light (150 \(\mu\)mol m\(^{-2}\) s\(^{-1}\)) for 7 d. At this point, the final percentage germination was scored. Germination was recorded as radicle emergence. In all cases, experiments were performed in quadruplicate, using 50 to 80 seeds per replicate. All germination data are expressed as the mean with standard error of the mean.

Accession Numbers

Sequence data from this article can be found in the GenBank/EMBL data libraries under the following accession numbers: XERICO (At2g04240), AUXIN RESPONSE FACTOR10 (ARF10; At2g28350), GA REQUIRING3 (GA3; At5g25900), PLANT U-BOX9 (PUB9; At3g07360), NAC 014 (ANAC014; At1g33060), GA 20-OXIDASE1 (GA20ox1; At4g25420), GIBBERELLIN 3-\(\beta\)-HYDROXYLASE1 (GA4/GA3ox1; At1g15550), GIBBERELLIN 3-\(\beta\)-HYDROXYLASE2 (GA3ox2; At1g08340), PYR1-like9 (PYL9; At1g01360), ASG4 (myb family; At1g01520), ASG5 (Protein kinase; At1g20650), ASG6 (CRK2 Protein kinase; At1g07520), ASG7 (unknown protein; At5g47580), ASG8 (unknown protein; At2g04075), SCARECROW-LIKE3 (SCL3; At1g05420), G-PROTEIN ALPHA SUBUNIT1 (GPA1; At2g26300), PHYTOCHROME INTERACTION FACTOR3-LIKES (PI5; At2g20180), SCARECROW-LIKE14 (SCL14; At1g07530), ABABINVSENSITIVE3 (ABIS; At3g24650), HISTONE MONOUBIQUITINATION2 (HUB2; At1g52550), HOMOLOGY TO AB11 (HAB1; At1g72770), FLOWERING LOCUS C (FLC; At5g10140), AGAMOUS-LIKE67 (AGL67; At1g77950), ARF-GAP DOMAIN2 (AGD2; At1g60860), LATE ELOGATED HYPOCOTYL (LHY; At1g01060), PHYTOCHROME A (PHYA; At1g09570), ETHYLENE INSENSITIVE5 (EIN5; At3g20770), RGA-LIKE3 (RGL3; At5g17490), SENSITIVE TO ABA AND DROUGHT1 (SAD1; At5g48870), SOMNUS (At1g03790), HIGH MOBILITY GROUPS (HMGs; At1g48620), EXTRA-LARGE GTP BINDING PROTEIN1 (XLG1; At2g23460), XLG2 (At4g34390), PYR1-like2 (PYL2; At2g26500), GIGANTEA (GI; At1g22770), REPRESSOR OF GA1-3 (RG1; At2g01570), ABABINVSENSITIVE4 (ABIV; At2g04220), ABA DEFICIENT3 (ABAD3; At1g16540), \(\beta\)-Hydroxylase1 (At4g25700), PYL4 (At2g38310), SPA1-RELATED2 (SPA2; At4g11110), \(\xi\)-CAROTENE DESATURASE (ZDS; At3g04870), PYRABACTIN1 (PYR1; At1g17870), PYL8 (At5g53160), SIZ1 (At SIZ1; At5g60410), SPATULA (SPT; At4g36930), GATA NITRATE CARBON (GNC; At5g56860), ABI3 INTERACTING PROTEIN2 (AIP2; At5g20910), MYB33 (At5g06100), MYB101 (At2g32460), SPINDLY (SPY; At3g11540), SnRK2.6 (At4g33950), ABI5 BINDING PROTEIN2 (AFP2; At1g15370), GIBBERELLIN INSENSITIVE DWARF1 (GID1A; At3g05120), and ALDEHYDE OXIDASE1 (AAO1; At1g20960).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. A Simplified Example of How BioHEL Generates a Rule Set.

Supplemental Figure 2. The Seed Coprediction Network Generated Using Randomly Assigned Class Labels to the Seed Microarray Samples.

Supplemental Figure 3. Location of the Newly Characterized ASG Genes within SeedNet.

Supplemental Table 1. Summary of Nongerminating and Germinating Arabidopsis Seed Microarray Samples Used as the Training Set with BioHEL to Generate SCOPNet.

Supplemental Table 2. Top 20 Most Frequent Genes in Rules for Germination and Nongermination.

Supplemental Table 3. Non-germination-Associated Candidate Genes Investigated for Germination Phenotypes in This Study.

Supplemental Table 4. Alleles of the Germination Regulatory Genes Identified in This Study and the Properties of Their Corresponding Nodes in SCOPNet.

Supplemental Table 5. Presence of All Previously Characterized Regulators of Seed Germination within the Modules Identified by MCODE within SCOPNet.

Supplemental Table 6. Presence of All Statistically Overrepresented cis-Elements within the Modules Identified by MCODE within SCOPNet.

Supplemental Table 7. List of Primers Used in this Study.
Supplemental Data Set 1. Gene Lists Used in This Study.
Supplemental Data Set 2. Genes Present with the Modules Identified by MCODE in SCoPNet.

ACKNOWLEDGMENTS

G.W.B. was funded by a Marie Curie International Incoming Fellowship. J.M. is funded by the University of Nottingham, M.J.H. was supported by Biotechnology and Biological Sciences Research Council Grant BBG02488X1 (ERANET-PG virtual seed, vSEED). J.B. was supported by the UK Engineering and Physical Sciences Research Council Grant EP/H016597/1. We thank Daniel J. Gibbs and Ute Voss (University of Nottingham) for their comments on the manuscript prior to submission.

AUTHOR CONTRIBUTIONS

G.W.B., M.J.H., and J.B. designed the research, analyzed the data, and wrote the article. G.W.B., E.G., and J.M. performed the research.

Received June 14, 2011; revised August 1, 2011; accepted August 25, 2011; published September 6, 2011.

REFERENCES


Large-Scale Data Sets

This information is current as of March 17, 2014

<table>
<thead>
<tr>
<th>Supplemental Data</th>
<th><a href="http://www.plantcell.org/content/suppl/2011/08/25/tpc.111.088153.DC1.html">http://www.plantcell.org/content/suppl/2011/08/25/tpc.111.088153.DC1.html</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>eTOCs</td>
<td>Sign up for eTOCs at: <a href="http://www.plantcell.org/cgi/alerts/ctmain">http://www.plantcell.org/cgi/alerts/ctmain</a></td>
</tr>
<tr>
<td>CiteTrack Alerts</td>
<td>Sign up for CiteTrack Alerts at: <a href="http://www.plantcell.org/cgi/alerts/ctmain">http://www.plantcell.org/cgi/alerts/ctmain</a></td>
</tr>
<tr>
<td>Subscription Information</td>
<td>Subscription Information for <em>The Plant Cell</em> and <em>Plant Physiology</em> is available at: <a href="http://www.aspb.org/publications/subscriptions.cfm">http://www.aspb.org/publications/subscriptions.cfm</a></td>
</tr>
</tbody>
</table>