# Genome Regulation and Functional Genomics Analyses

## I. Introduction

Organisms respond to their environment by changing the expression of their genes. These genes encode proteins with specific functions that modify cellular function in ways appropriate for the environment. One example is when a seed germinates. Prior to germination all the cells in a seed are metabolically inactive, but upon exposure to the right environmental conditions the cells very rapidly become active and the plant begins to grow. This requires that thousands of genes are expressed at the right time. The fundamental aspects of gene expression are conserved across all higher organisms. Briefly, transcription factors recognize and bind sequences in the promoters of genes, then promote or repress expression of those genes. This process of gene expression regulation is complex, because plants and animals typically encode hundreds or more transcription factors. Each transcription factor may regulate hundreds of genes and any individual gene may be targeted by more than one transcription factor.

This seminar and workshop will explain how we analyse gene regulation at genome scale, both at the lab bench and once we have our data. It will highlight some key data types and how these are practically generated. You will work with a software package called Dynamic Regulatory Events Miner (DREM) that allows us to integrate large-scale time-series gene expression data with transcription factor-DNA binding data. DREM uses a Hidden Markov Model-based approach to produce models that can determine when transcription factors (TFs) activate genes and what genes they regulate [1]. DREM outputs an annotated dynamic regulatory map that highlights bifurcation events in the time-series data. In other words, it identifies places in the time series where sets of genes, which previously had roughly similar expression levels, diverge. Often these bifurcation events can be explained by TFs selectively regulating a certain subset of genes. DREM annotates these events with TFs potentially responsible for them [2].

We recommend you read [1-3] before the seminar, which detail the software and give a full example of where we have applied it. Given the time restrictions, we will use the pre-computed DREM model (**SD3_file4_outmodel.txt**) provided in the supplementary materials of [3] (https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-017-1302-3/MediaObjects/13059_2017_1302_MOESM1_ESM.zip) to navigate through the results and interpret them. The input files required to produce this model (**SD3_file1_expression.txt** and **SD3_file2_TFbinding.txt**) can also be found in the link provided.

## II. Prerequisites

### 1. Software requirements (PC and MAC)

Please ensure that these are downloaded and installed prior to the workshop:

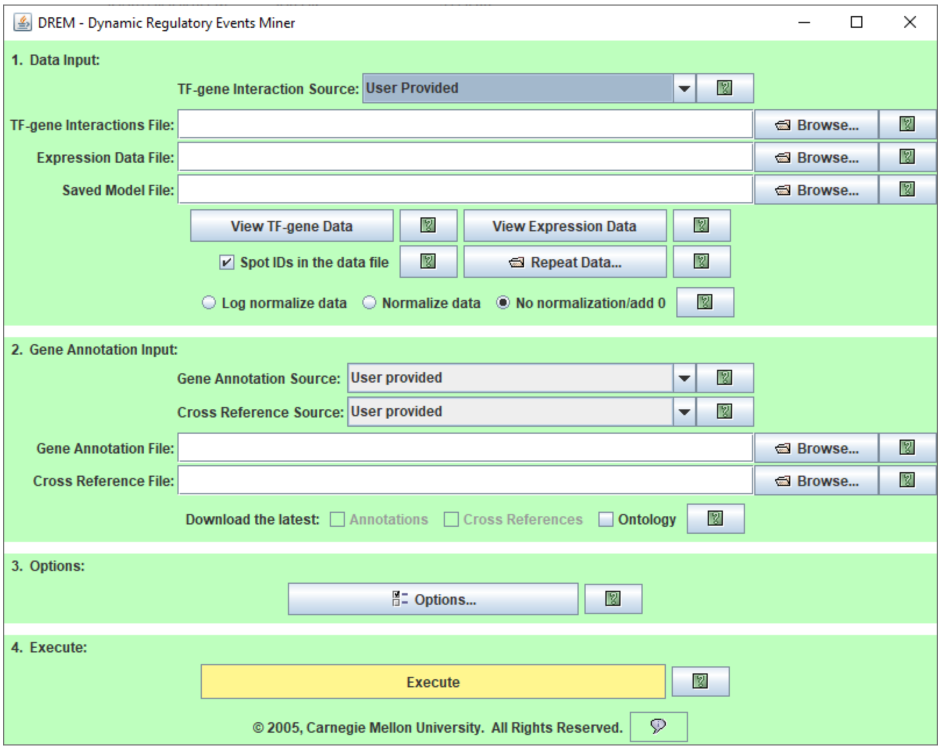
a. Java 1.5 or later from http://www.java.com
b. Download DREM v. 2.0.5 (latest version) from https://github.com/jernst98/STEM_DREM/archive/master.zip

### 2. Installation

a. Unzip **STEM_DREM-master.zip** and go to the **STEM_DREM-master** folder.
b. Unzip **drem2.zip** and go to the **drem2** folder.
c. Start DREM:
i. PC: Double click on **drem.cmd**

ii. MAC: type **java -mx1024M -jar drem.jar** on the Terminal app

d. An input interface will appear as shown in Figure 1.



DREM - Dynamic Regulatory Events Miner

1. Data Input:

TF-gene Interaction Source: User Provided

TF-gene Interactions File:

Expression Data File:

Saved Model File:

☑ Spot IDs in the data file

○ Log normalize data ○ Normalize data ● No normalization/add 0

2. Gene Annotation Input:

Gene Annotation Source: User provided

Cross Reference Source: User provided

Gene Annotation File:

Cross Reference File:

Download the latest: □ Annotations □ Cross References □ Ontology

3. Options:

4. Execute:

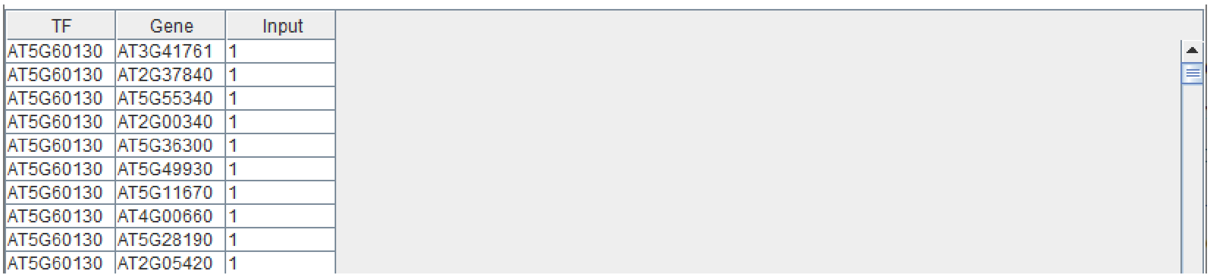
© 2005, Carnegie Mellon University. All Rights Reserved.

Figure 1. The main input interface of DREM. This is the first screen that appears when DREM is launched. From this screen a user specifies the input data, gene annotation information, and various execution options. Pressing the execute button at the bottom of the interface causes the DREM algorithm to execute.

## 3. Input files

The following files are needed to run DREM:

a. Transcription factor-gene interactions file (**SD3_file2_TFbinding.txt**). The file looks like this:



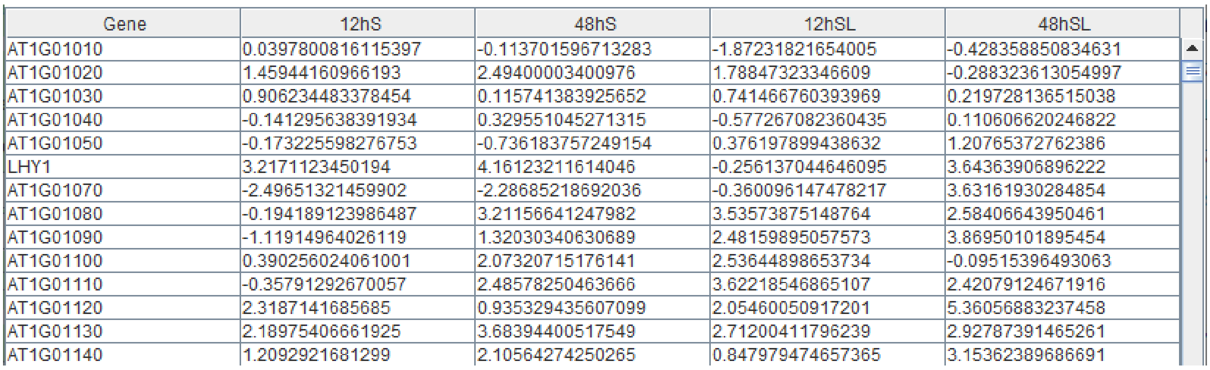
| TF | Gene | Input |
|---|---|---|
| AT5G60130 | AT3G41761 | 1 |
| AT5G60130 | AT2G37840 | 1 |
| AT5G60130 | AT5G55340 | 1 |
| AT5G60130 | AT2G00340 | 1 |
| AT5G60130 | AT5G36300 | 1 |
| AT5G60130 | AT5G49930 | 1 |
| AT5G60130 | AT5G11670 | 1 |
| AT5G60130 | AT4G00660 | 1 |
| AT5G60130 | AT5G28190 | 1 |
| AT5G60130 | AT2G05420 | 1 |

Figure 2. Three-column format to represent TF-gene interaction information data.

In this tab-delimited file, the first column corresponds to transcription factors (TFs), the second column corresponds to genes, and the third column denotes the presence or absence of interaction between a TF and a gene; an entry of **1** corresponds to the prediction that the TF regulates the gene while an entry of **0** corresponds to the prediction that there is no regulatory interaction between the TF and the gene.

The format used here is the three-column format. For other acceptable formats please see the DREM manual.

b. Expression Data file (**SD3_file1_expression.txt**). The file looks like this:



| Gene | 12hS | 48hS | 12hSL | 48hSL |
|---|---|---|---|---|
| AT1G01010 | 0.0397800816115397 | -0.113701596713283 | -1.87231821654005 | -0.428358850834631 |
| AT1G01020 | 1.45944160966193 | 2.49400003400976 | 1.78847323346609 | -0.288323613054997 |
| AT1G01030 | 0.906234483378454 | 0.115741383925652 | 0.741466760393969 | 0.219728136515038 |
| AT1G01040 | -0.141295638391934 | 0.329551045271315 | -0.577267082360435 | 0.110606620246822 |
| AT1G01050 | -0.173225598276753 | -0.736183757249154 | 0.376197899438632 | 1.20765372762386 |
| LHY1 | 3.2171123450194 | 4.16123211614046 | -0.256137044646095 | 3.64363906896222 |
| AT1G01070 | -2.49651321459902 | -2.28685218692036 | -0.360096147478217 | 3.63161930284854 |
| AT1G01080 | -0.194189123986487 | 3.21156641247982 | 3.53573875148764 | 2.58406643950461 |
| AT1G01090 | -1.11914964026119 | 1.32030340630689 | 2.48159895057573 | 3.86950101895454 |
| AT1G01100 | 0.390256024061001 | 2.07320715176141 | 2.53644898653734 | -0.09515396493063 |
| AT1G01110 | -0.35791292670057 | 2.48578250463666 | 3.62218546865107 | 2.42079124671916 |
| AT1G01120 | 2.3187141685685 | 0.935329435607099 | 2.05460050917201 | 5.36056883237458 |
| AT1G01130 | 2.18975406661925 | 3.68394400517549 | 2.71200411796239 | 2.92787391465261 |
| AT1G01140 | 1.2092921681299 | 2.10564274250265 | 0.847979474657365 | 3.15362389686691 |

Figure 3. Format to represent gene expression data. The expression values shown here have been log normalised.

In this study, Arabidopsis plants were grown to maturity and seeds were harvested from the plants. The seeds were kept under dry conditions for two weeks to ripen and they were collected as follows:

a. 0 h: immediately after ripening (control)
b. 12h S: 12 hours after ripening and exposure to cold dark stratification (S)
c. 24 h S: 24 hours after ripening and exposure to cold dark stratification (S)
d. 12 h SL: 12 hours after ripening and exposure to continuous light (SL)
e. 24 h SL: 24 hours after ripening and exposure to continuous light (SL)

In the tab-delimited file shown above, the first column represents the genes and the remaining columns represent the log2-fold changes of differentially expressed genes (DEGs) relative to 0 h. Three biological replicates were collected for each condition (a. to e.). An example of how the log2-fold changes were calculated for each gene at 12 h S compared to 0 h is shown below:

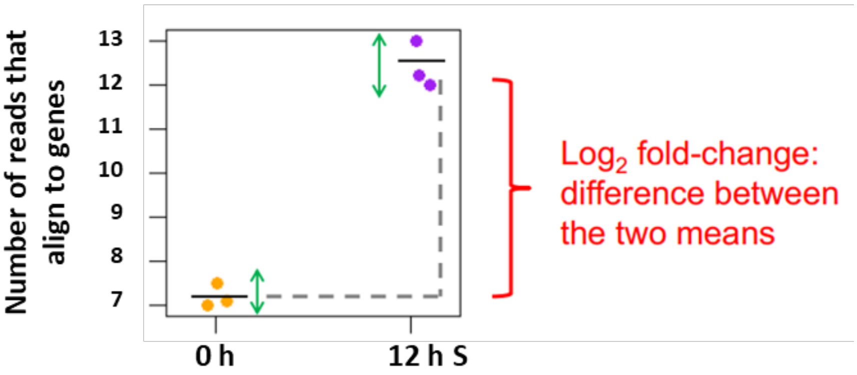
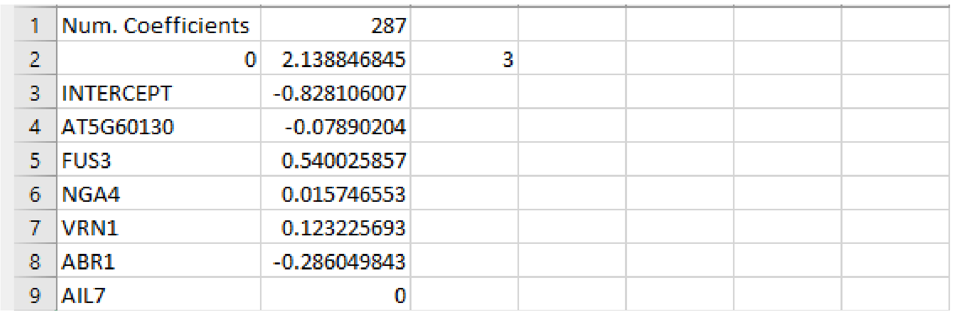


Figure 4. Example of how log2-fold changes are calculated between treatment (12 h S) and control (0 h).

c. Saved model file (**SD3_file4_outmodel.txt**). For the purpose of this workshop, we will use this pre-computed DREM model. The file looks like this:



| | | | |
|---|---|---|---|
| 1 | Num. Coefficients | 287 | |
| 2 | 0 | 2.138846845 | 3 |
| 3 | INTERCEPT | -0.828106007 | |
| 4 | AT5G60130 | -0.07890204 | |
| 5 | FUS3 | 0.540025857 | |
| 6 | NGA4 | 0.015746553 | |
| 7 | VRN1 | 0.123225693 | |
| 8 | ABR1 | -0.286049843 | |
| 9 | AIL7 | 0 | |

Figure 5. Format of the saved DREM model provided by Narsai et al. (2017).

In the above, **Num. Coefficients** represents 287 known TF-gene interactions that have been derived experimentally using DNA Affinity purification (DAP-Seq) [4]. Additionally, the first column contains the names of the TFs, and the second column contains a coefficient for each TF. This coefficient is derived from a logistic regression classifier that uses the TF-gene interaction data as supervised input, and it is used to classify genes into diverging paths at a split node in the model. We will have a closer look at this when we interpret the results later.

The above model can be re-created using **SD3_file2_TFbinding.txt, SD3_file1_expression.txt** and the parameters saved in **SD3_file3_parameters.txt.**

# III. Running DREM

## 1. Data input

We need to load the following files:

a. Transcription factor-gene interactions file: **SD3_file2_TFbinding.txt**
b. Expression data file: **SD3_file1_expression.txt**
c. Saved Model File: **SD3_file4_outmodel.txt**

## 2. Gene annotation input

a. We will select **Arabidopsis thaliana (TAIR)** in the **Gene Annotation Source** box. When doing so, we will see that **tair.gaf.gz** is automatically loaded in the **Gene Annotation File** box. This file carries important information about the genes and TFs involved in this study such as the gene/TF names, a description of the gene/TF function, protein domain information and gene ontology (GO) terms that have been assigned to the gene/TF. GO terms allow the genes/TFs to be grouped according to the biological process, molecular function and cellular component they are involved in; one gene/TF can have multiple GO term. For more information regarding the content of the gene annotation file, please refer to: http://geneontology.org/docs/go-annotation-file-gaf-format-2.0/ and http://geneontology.org/docs/ontology-documentation/
b. We will also use the latest versions of **Annotations** and **Ontology** by selecting the respective boxes.

## 3. Options

**Options** contains a list of parameters that be specified by the user**.** These include: **Gene annotations, GO Analysis, DECOD Options, Expression Scaling Options, DREMmir, Filtering Options, Search options,** and **Model Selection Options.** Please consult the DREM manual for more information regarding these options.

In today's workshop we will use the same parameters that were employed by Narsai et al. (2017) to create their DREM model. The majority of parameters used in their study were set according to the default DREM except for the following:

a. Search Options
   - Allow_Path_Merges: true
   - Convergence_Likelihood_%: 0.1

b. Expression Scaling Options
   - Regulator_Types_Used_For_Activity_Scoring: TF

### 4. Execute

We press the **Execute** button to load the pre-saved DREM model.

**IMPORTANT NOTE FOR PC USERS ONLY (to be performed prior to the workshop):**

After pressing the Execute button, you may notice that the program takes a long time to load or does not load the DREM map at all. If this happens, please run DREM without the pre-computed model (**SD3_file4_outmodel.txt**), i.e. run DREM with the TF-gene interactions file (**SD3_file2_TFbinding.txt**), expression data file (**SD3_file1_expression.txt**) and gene annotation file (**tair.gaf.gz**) ONLY and have the DREM map ready on the day of the workshop (this should take about 15-20 mins to run). Please also ensure that you change the Options (Search Options and Expression Scaling Options) as described above.

## IV. Results

Below is a screenshot of the DREM output where the x-axis represents the time points and the y-axis shows the log2-fold changes of the DEGs relative to 0h. A genes/TF with a positive log2-fold change is said to be up-regulated, while a negative log2-fold change indicates down-regulation.

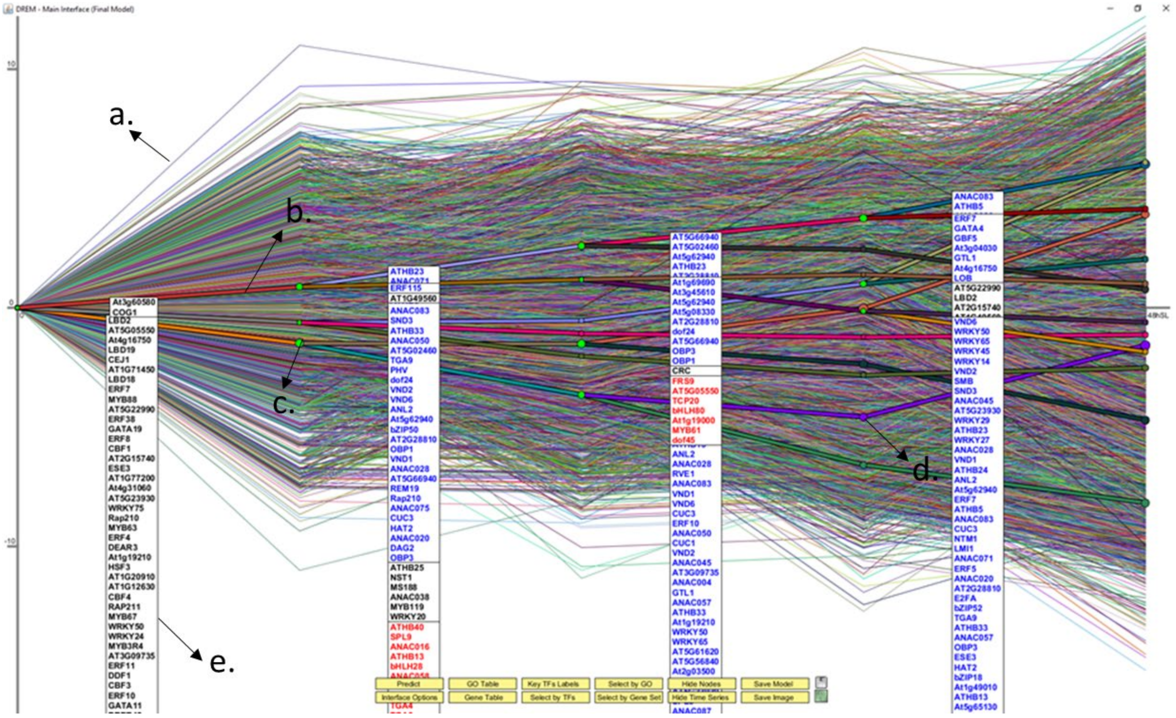


Figure 6. The main output interface window of DREM showing a DREM map.

The DREM map shows the following features:

- Thin coloured lines (a.): these show the pattern of gene expression of individual genes. In other words, these lines show how the expression levels of a specific gene/TF changes over time.
- Solid coloured lines (b.): these represent the paths to which the DEGs have been assigned.
- Nodes: each node is associated with a Gaussian distribution, and the size of the node is proportional to Gaussian's standard deviation. A relatively small node implies the expression of the genes going through that node will be tightly centered around the node. A relatively large node indicates genes assigned to the path through that node will not necessarily pass closely through the node. A green node (c.) indicates a bifurcation/split and represent sets of genes that change their expression between consecutive time points. TFs are assigned to split nodes allowing DREM to infer their time of activation. The non-green nodes show that there are no splits in the paths.
- Tables/modules (e.): these contain the TFs that have been assigned to their respective paths; these TFs are predicted to regulate genes that change their expression patterns, i.e. genes that diverge at the green split nodes. The TFs are colourcoded; blue TFs are upregulated, red TFs are downregulated and black TFs are those that do not exhibit significant transcriptional changes.

A simplified version of this map can be produced by clicking the **Hide Time Series** option and manually moving the TF boxes/modules around so that they do not overlap each other:

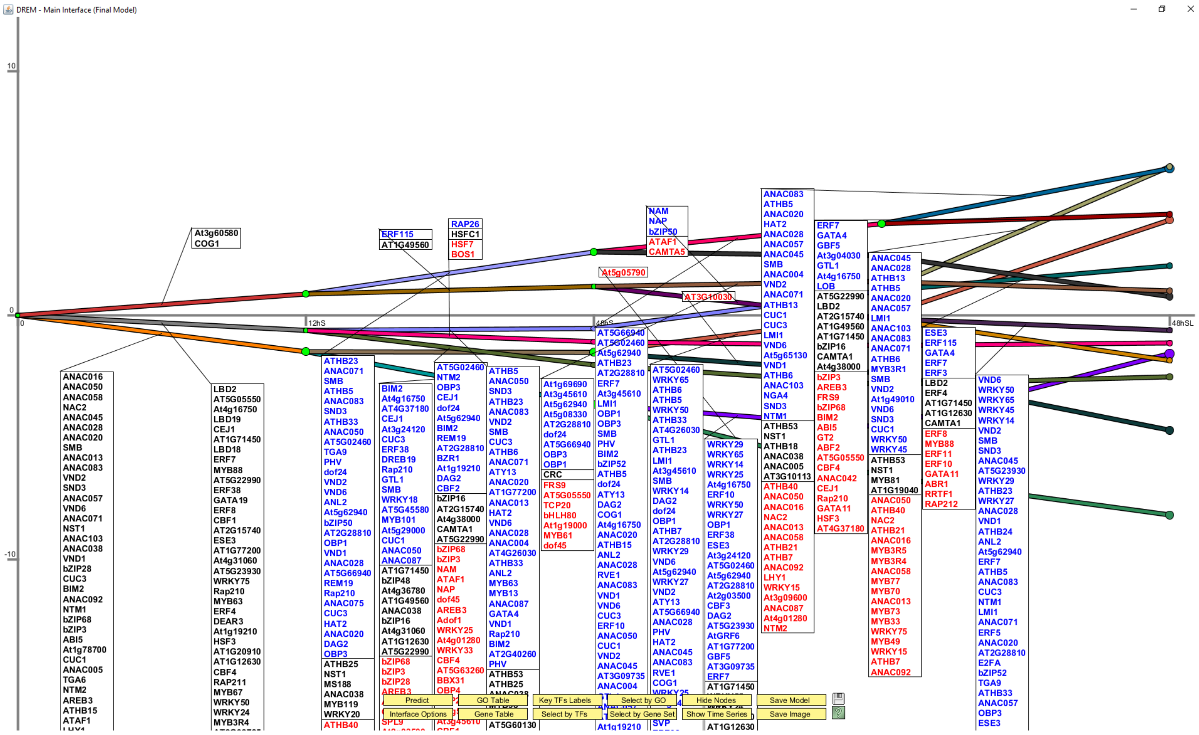


Figure 7. Simplified version of the DREM map from Figure 6.

# References

1. Schulz, M.H., et al., *DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data.* BMC systems biology, 2012. **6**(1): p. 104.
2. Ernst, J., et al., *Dynamic Regulatory Events Miner (v2.0.5) User Manual.*
3. Narsai, R., et al., *Extensive transcriptomic and epigenomic remodelling occurs during Arabidopsis thaliana germination.* Genome biology, 2017. **18**(1): p. 172.
4. O'Malley, R.C., et al., *Cistrome and epicistrome features shape the regulatory DNA landscape.* Cell, 2016. **165**(5): p. 1280-1292.