# Identification of Cis-Regulatory Sequences Controlling Pollen-Specific Expression of Hydroxyproline-Rich Glycoprotein Genes in *Arabidopsis thaliana*

# Yichao Li, Maxwell Mullin, Yingnan Zhang, Frank Drews, Lonnie Welch, Allan Showalter

# Abstract

# Introduction

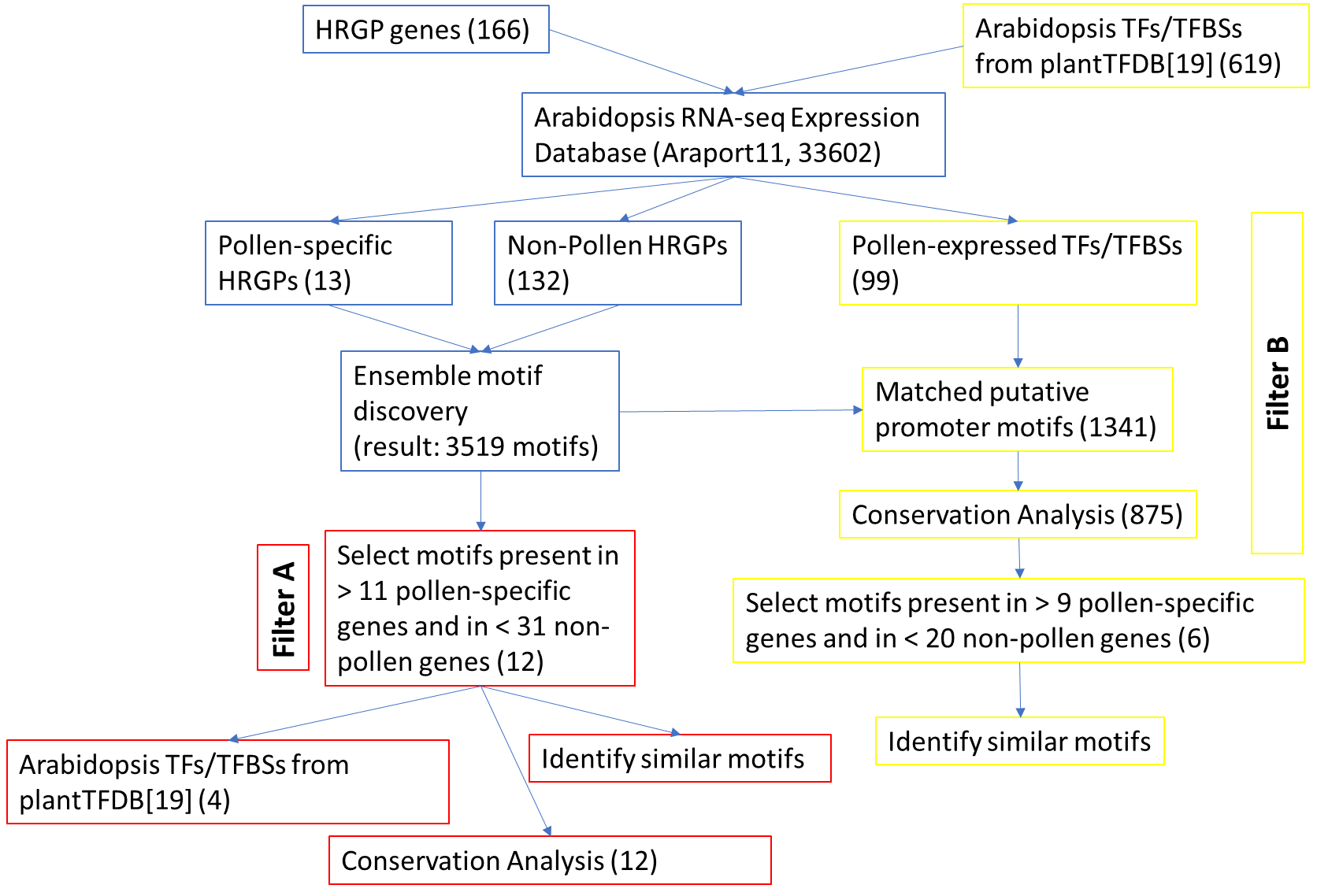
Hydroxyproline-rich glycoproteins (HRGPs) are a superfamily of plant cell wall proteins involved in various aspects of plant growth and development. The HRGP superfamily consists of three family members, the extensins (EXTs), arabinogalactan-proteins (AGPs), and proline-rich proteins (PRPs). Although all HRGPs contain hydroxyproline (Hyp or O), a rare amino acid produced by the post-translational modification of proline residues by the enzyme prolyl hydroxylase, the three family members are distinguished by their unique amino acid compositions, repeating amino acid motifs, and degree and type of glycosylation (i.e., sugar modifications). For example, AGPs can be identified by their biased amino acid compositions of Pro (P)/Hyp (O), Ala (A), Ser (S), and Thr (T), their frequent occurrence of AP and PA dipeptide repeats, and their numerous, large arabinose and galactose-rich polysaccharide chains attached to O residues. In contrast, EXTs tend to be rich in S, P/O, Val (V), Tyr (Y), and Lys (K), have numerous SOOOO pentapeptide repeats, and have numerous, short arabinose oligosaccharide side chains attached to their O residues. Finally, PRPs are rich in P, V, K, Cys (C), and T, often have various Pro/Hyp-rich amino acid repeat motifs in which not all the P residues are modified to form O, and are the least glycosylated members of the HRGP superfamily.

Bioinformatic programs analyzing genomic/proteomic data from the model genetic plant, *Arabidopsis thaliana*, have identified 162 HRGPs consisting of 85 AGPs, 59 EXTs, and 18 PRPs. This information has provided new insight to the HRGP superfamily and is being used by researchers to facilitate and guide further research in the field. Two of most interesting questions regarding HRGPs are the following. 1. What are specific functions and mechanisms of action of the various HRGPs? And 2. How are the various HRGP genes regulated at the transcriptional level? With regard to the first question, HRGP researchers are increasing utilizing a reverse genetic approach. In other words, they look for or produce genetic mutants in the specific HRGP gene that they are interested in and examine any abnormal phenotypes in such mutants to provide insight to the gene’s function. Once such a function is identified, biochemical approaches need to be utilized to discover the mechanism by which the specific HRGP accomplishes its function. With regard to the second question, microarray data, and more recently even more robust RNA-Sequencing (RNA-Seq) data, has revealed organ- and tissue-specific patterns of expression for the HRGP genes. HRGP researchers can now utilize such expression data to address how the HRGP genes are expressed in tissue-specific patterns. In this paper, we have utilized RNA-Seq data to identify 13 HRGP genes that are specifically expressed in pollen, the male microgametophytes of seed plants responsible for the production of male gametes (sperm cells) and have applied a novel bioinformatics approach to identify cis-regulatory sequences (motifs) that likely control the pollen-specific expression of HRGP genes in *Arabidopsis thaliana*. Motifs are specific short DNA sequences, often 8-20 nucleotides in length [1], which are statistically overrepresented in pollen-specific genes. De novo motif discovery is a traditional computational method to identify regulatory motifs in a set of phenotypic related sequences. Numerous motif discovery methods have been developed. The bioinformatics approach we propose is based on GimmeMotifs [2]. GimmeMotifs is a generative motif discovery tool [citation]; the term “generative” means that no background sequence set is required. It is also an ensemble motif learning tool; the term “ensemble” means that a suite of different motif discovery tools is being used to identify regulatory motifs. We also used discriminative motif discovery tools. These tools require a set of background sequences to contrast the foreground sequences of interest; the term “background” means that the discovered motifs are not expected to be highly enriched in that set. Our approach is based on DECOD [3] and DME [4] are two well-known discriminative motif discovery tools.

# Results

In this section we present the identified motifs that control the pollen-specific expression of HRGP genes in *Arabidopsis thaliana.*

The overall bioinformatics workflow of our approach is shown is Figure 1. The strategy used allows us to find a large set of possible motifs from all HRGP genes and eliminate any that prove to occur in non-pollen tissues, leaving pollen specific tissues.



**Fig 1. Bioinformatics workflow used to identify of cis-regulatory sequences (motifs) controlling pollen-specific expression of hydroxyproline-rich glycoprotein genes in Arabidopsis thaliana. The red and yellow boxes represent the two different strategies (Filter A (red) ad Filter B (yellow).**

We started with a set of all (166) known HRGP genes in Arabidopsis thaliana. We then split the set into three categories based on their pollen specificity index; those that were only expressed in pollen tissue, non-pollen tissue (i.e., genes which are not expressed in pollen, but instead in one or more other plant tissues), and both. We discarded genes expressed in both pollen and the non-pollen tissue, as they would not help find pollen specific motifs. This left 145 genes in total, 13 pollen-specific genes and 132 non-pollen specific genes. We then used an Ensemble of motif discovery tools to find 3519 motifs. Subsequently, we applied to different filter strategies (Filter A = red, Filter B = yellow). The red filter simply selects motifs present in more than 11 pollen-specific genes and in at most 30 non-pollen genes. This resulted in 12 motifs. The yellow filter strategy first matches the set of all motifs to known TFBSs known to be expressed in pollen. Subsequently, conservation analysis and motif filtering were performed which resulted in overall 3 putative motifs.

**FILTER A RESULTS:**

Filter A selects motifs present in 11(84.5%) or more pollen specific genes and present in at most 30(22.7%) non-pollen genes (see Figure 2; the thresholds are marked in red)

**A close up of a device

Description generated with high confidenceFig 2. Pollen Specific vs. Non-pollen Motif Occurrences. Each point on this scatter plot represents a single motif. The x-axis and y-axis represents the number of times the motif is present (with a p-value of 1e-4 or less) in the pollen specific and non-pollen specific genes, respectively. To select the most promising motifs, we chose thresholds (marked in red) of motifs present in 11(84.5%) or more pollen specific genes and present in at most 30(22.7%) non-pollen genes.**

Table 1 shows the resulting selected list of 12 motifs. The rightmost column of table 1 shows the related known transcription factor binding sites we found with pollen related functions and their match p-value.

**Table 1. Motifs in 11 or more pollen specific and 30 or fewer non-pollen HRGP sequences. The foreground coverage indicates the percentage (and number) of pollen-specific HRGP genes having that motif in its promoter region. The background coverage is the percentage (and number) of the non-pollen HRGP genes having that motif in its promoter region. The relative frequency is the foreground coverage divided by the background coverage. The matched TFBS p-value is a known transcription factor binding site that relates to pollen along with the accompanying p-value.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Motif Name | Motif Logo | Foreground Coverage | Background Coverage | Relative Frequency | Best Matched TFBS  (p-value) |
| DME\_GADGAYKAS |  | 85% (11) | 19% (25) | 4.47 | AT3G11280, MYB-LIKE PROTEIN  (0. 000356486) |
| DME\_GATYTKRHG | C:\Users\Owner\Desktop\logoDME_GATYTKRHG.png | 85% (11) | 20% (27) | 4.47 |  |
| DME\_GRHTGDTGA |  | 85% (11) | 20% (27) | 4.47 | AT5G58620, TZF9(1.32812e-05) |
| DME\_MARKGDSRGA |  | 85% (11) | 22% (29) | 3.85 |  |
| gimme\_102\_Improbizer\_GCGTTATACCCGAGGATCAG |  | 92% (12) | 15% (20) | 6.09 |  |
| gimme\_104\_Improbizer\_GTGCAACGGAGAGT |  | 92% (12) | 14% (18) | 6.77 |  |
| gimme\_105\_Improbizer\_AACACACGTTTATTAGATGTTT |  | 100% (13) | 18% (24) | 5.50 | AT1G75080, BZR1(1.57304e-06) |
| gimme\_132\_MEME\_3\_w10 |  | 92% (12) | 11% (15) | 8.12 |  |
| gimme\_13\_BioProspector\_w10\_3 |  | 85% (11) | 13% (17) | 6.57 |  |
| gimme\_143\_MEME\_4\_w12 |  | 85% (11) | 12% (16) | 6.98 | AT5G11260, TED 5(0.000468975) |
| gimme\_146\_MEME\_7\_w12 |  | 92% (12) | 16% (21) | 5.80 |  |
| gimme\_16\_BioProspector\_w12\_1 |  | 85% (11) | 20% (27) | 4.14 |  |
| gimme\_92\_MDmodule\_Motif.12.3 |  | 85% (11) | 20% (26) | 4.30 |  |

Table 1 shows us that 4 of our original 12 motifs do appear to have known pollen related functions. Out of all the results two stand out, gimme\_105\_Improbizer\_AACACACGTTTATTAGATGTTT and DME\_GRHTGDTGA having 8 and 3 respectively similar pollen related transcription factor binding sites is a strong indicator that this is a motif worth experimental study.

One of the motifs (gimme\_105\_Improbizer\_AACACACGTTTATTAGATGTTT) appears in 100% of the pollen-specific HRGP genes. **Additionally, there are 36 (see supplemental materials for a list of these genes). However, none of these motifs passed Filter A.**

A similarity analysis revealed that the following pairs of motifs capture strongly related patterns. This may indicate that each related pair captures binding site variation for a transcription factor.

|  |  |
| --- | --- |
| gimme\_146\_MEME\_7\_w12 | DME\_MARKGDSRGA |
| gimme\_13\_BioProspector\_w10\_3 | gimme\_132\_MEME\_3\_w10 |
| gimme\_16\_BioProspector\_w12\_1 | gimme\_143\_MEME\_4\_w12 |
| gimme\_102\_Improbizer\_GCGTTATACCCGAGGATCAG | DME\_GADGAYKAS |
| gimme\_92\_MDmodule\_Motif.12.3 | gimme\_16\_BioProspector\_w12\_1 |

Figure 2 does not show any motifs that are strongly non-HRGP (i.e. the top left corner). This indicates that none of the motifs we originally generated are non-pollen specific. In a random set of motifs, we would expect a roughly uniform distribution of motifs. This is obviously not the case however as there is a large empty area. This is certainly interesting and worth examining in future work.

|  |  |
| --- | --- |
| Motif Name | Conserved Homologs |
| gimme\_132\_MEME\_3\_w10 | [AT3G01700,fgenesh2\_kg.3\_\_46\_\_AT3G01700.1], [AT4G34440,Al\_scaffold\_0007\_610], [AT3G18810,scaffold\_302227.1], [AT2G18470,Al\_scaffold\_0003\_3824] |
| gimme\_16\_BioProspector\_w12\_1 | [AT3G01700,fgenesh2\_kg.3\_\_46\_\_AT3G01700.1], [AT1G49270,fgenesh1\_pm.C\_scaffold\_1003128], [AT3G18810,scaffold\_302227.1], [AT2G18470,Al\_scaffold\_0003\_3824] |
| gimme\_13\_BioProspector\_w10\_3 | [AT3G18810,scaffold\_302227.1], [AT2G18470,Al\_scaffold\_0003\_3824] |
| DME\_GADGAYKAS | [AT4G34440,Al\_scaffold\_0007\_610], [AT3G18810,scaffold\_302227.1] |
| gimme\_143\_MEME\_4\_w12 | [AT1G49270,fgenesh1\_pm.C\_scaffold\_1003128], [AT4G34440,Al\_scaffold\_0007\_610], [AT3G18810,scaffold\_302227.1] |
| gimme\_104\_Improbizer\_GTGCAACGGAGAGT | [AT3G01700,fgenesh2\_kg.3\_\_46\_\_AT3G01700.1], [AT4G34440,Al\_scaffold\_0007\_610] |
| gimme\_146\_MEME\_7\_w12 | [AT3G01700,fgenesh2\_kg.3\_\_46\_\_AT3G01700.1] |
| gimme\_105\_Improbizer\_AACACACGTTTATTAGATGTTT | [AT1G24520,scaffold\_103015.1], [AT1G49270,fgenesh1\_pm.C\_scaffold\_1003128], [AT2G18470,Al\_scaffold\_0003\_3824] |
| gimme\_92\_MDmodule\_Motif.12.3 | [AT2G18470,Al\_scaffold\_0003\_3824] |
| DME\_GRHTGDTGA | [AT1G24520,scaffold\_103015.1], [AT3G18810,scaffold\_302227.1] |
| DME\_MARKGDSRGA | [AT3G01700,fgenesh2\_kg.3\_\_46\_\_AT3G01700.1], [AT1G24520,scaffold\_103015.1], [AT4G34440,Al\_scaffold\_0007\_610], [AT2G18470,Al\_scaffold\_0003\_3824] |
| gimme\_102\_Improbizer\_GCGTTATACCCGAGGATCAG | [AT3G18810,scaffold\_302227.1] |

All the 12 motifs are conserved in at least one HRGP homolog. And 3 motifs are conserved in 4 HRGP homologs.

**FILTER B RESULTS:**

Out of the 99 pollen-expressed TFs from PlantTFDB, there were 6 TFs are pollen-specific (i.e. Tau >= 0.85). They are: AT2G32460,AT5G58850,AT3G61910,AT2G17180,AT5G04390,AT5G61620.

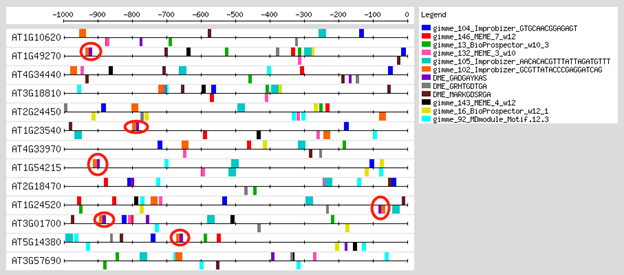
Table 3 presents the list of putative promoter motifs for pollen-specific HRGPs for Filter B.

**Table 3. List of putative promoter motifs for pollen-specific HRGPs. “GATA9 is the closest homolog of GATA12”, [5] GATA12 "levels were high in mature pollen grains but diminished in the germinated pollen grains and their pollen tubes" [6] “Characterization of a novel gene encoding a putative single zinc-finger protein, ZIM, expressed during the reproductive phase in Arabidopsis thaliana.”[6] (Allan, Yichao: The previous quotes need clarification/context)**

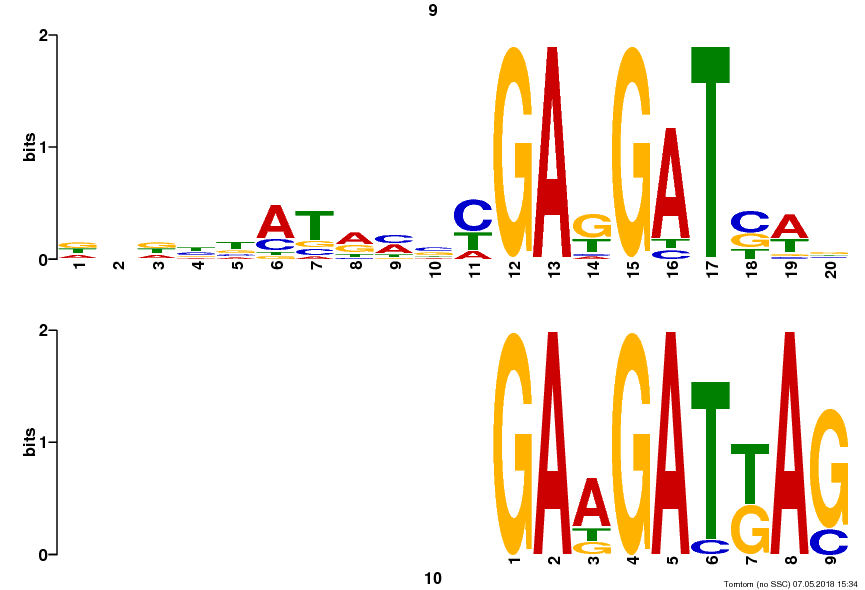
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Motif Name | Motif Logo | Fore Coverage | Back Coverage | Relative Frequency | TFBS match (p-value) | TF TAIR ID |
| Motif\_1 |  | 76.9% | 9.8% | 7.8 | GATA9 (0.0005) | AT4G32890 |
| Motif\_2 |  | 84.6% | 12.1% | 7.0 | ZML1 (0.0009) | AT3G21175 |
| Motif\_3 |  | 76.9% | 12.1% | 6.3 | GATA9 (0.0018) | AT4G32890 |

# DISCUSSION

**FILTER A:**

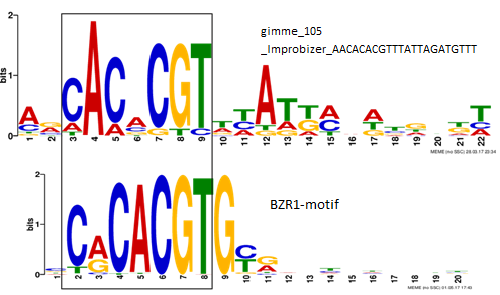
We investigated patterns of the locations of the motif occurrences in the HRGP promoter sequences and found one interesting pattern (see figure 3). The motifs gimme\_102\_Improbizer\_GCGTTATACCCGAGGATCAG and DME\_GADGAYKAS co-occur frequently. Further investigation shows that the two motifs are very similar (see figure 4).

**Fig 3. Visualization of motif occurrences in HRGP gene promotor regions. This figure displays the occurrence of each motif relative to each other. Weather the mark is above or below the line indicates which strand the motif occurs on. Circled in red are two motifs that occur together frequently, these are gimme\_102\_Improbizer\_GCGTTATACCCGAGGATCAG and DME\_GADGAYKAS. Their co-occurrence is caused by the two motifs are so similarities (see figure 4). The lack of other patterns in the motif occurrence indicates that there is no significant interaction between the 12 motifs.**



**Fig 4. Alignment of gimme\_102\_Improbizer\_GCGTTATACCCGAGGATCAG (top) and DME\_GADGAYKAS(bottom). This alignment has a p-value of 2.10e-04.**

Once 12 motifs were selected we used Tomtom[7] to search for known *Arabidopsis thaliana* transcription factor binding sites related to our 12 motifs. Our search returned 129 statistically significant (i.e., having a p-value of less than 9e-3) related known binding sites from ArabidopsisDAPv1 (113) and CIS\_BP\_Arabidopsis\_thaliana (16). After manual research of the 129 binding sites we selected 11 that were known to have pollen related functions.

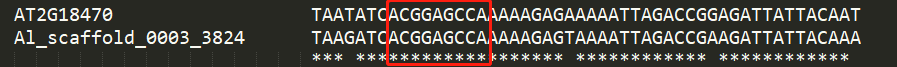
We notice that one of the motif is highly similar (p-value = 1.57304e-06) to the known (Brassinazole Resistant 1) BZR1 binding motif. brassinosteroid (BR) is an important class of steroid hormones in plants that regulates gene expression and cell development[3], [4]. BZR1 is a key transcription factor in the BR signaling pathway, where the binding of BR to a cell surface receptor kinase (BRI1) directly regulates the phosphorylation of BZR1, which then binds to the promoters of BR responsive genes [3]. BR is first discovered in pollen where it regulates cell elongation. Although it is later found in all tissues, the highest abundance is found primarily in pollen, seeds and fruit [4]. The discovery of BZR1-like binding site in the promoters of pollen-specific HRGP suggests that these HRGPs are likely to be regulated by the BR signaling pathway. Indeed, cell wall modification is reported to be one of the major functions targeted by the BR pathway [4]. The consensus sequence of BR-responsive element (BRRE) is CGTGYG reported in [4] and NCRCACGTGYNNNNNNNNNN reported in [5]. The consensus of is NNCACNCGTNNDDDHNDNNNNH. Given that gimme\_105\_Improbizer\_AACACACGTTTATTAGATGTTT is significantly similar to BZR1 binding motif, it is likely to be a variant of BR-responsive element (BRRE) with a common core element CACNCGT (one mismatch in the middle).

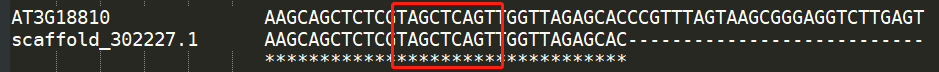
**Table 2. Known transcription factor binding sites related to the motifs with pollen related functions. This table shows information to identify the binding sites related to the 12 discovered motifs, as well as short summary of pollen related function associated with that transcription factor binding site.**



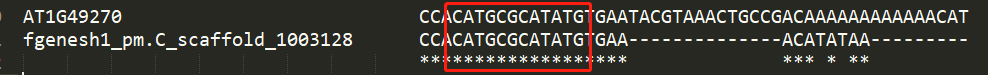
(Allan, Yichao: I assume Figure 5 shows homolog of the 3 pollen-specific HRGP genes in *Arabidopsis Lyrata. Is that correct? Need a more thorough description of the results)* Figure 5 shows the homologs of the 3 pollen-specific HRGP genes in *Arabidopsis Lyrata*.

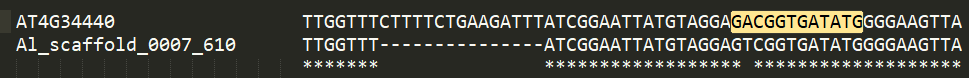
Motif 1

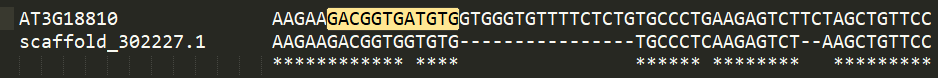




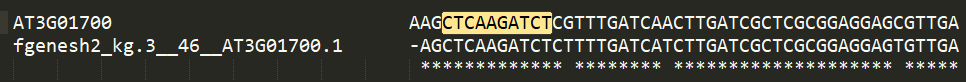
Motif 2

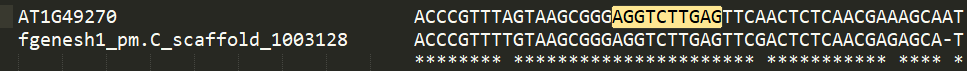






Motif 3





**Figure 5. Conserved motif sites.** The actual location of the 3 reported motifs is highlighted.

Figure 6 shows the patterns of the locations of the 3 reported motifs in the HRGP promoter sequences.

# 

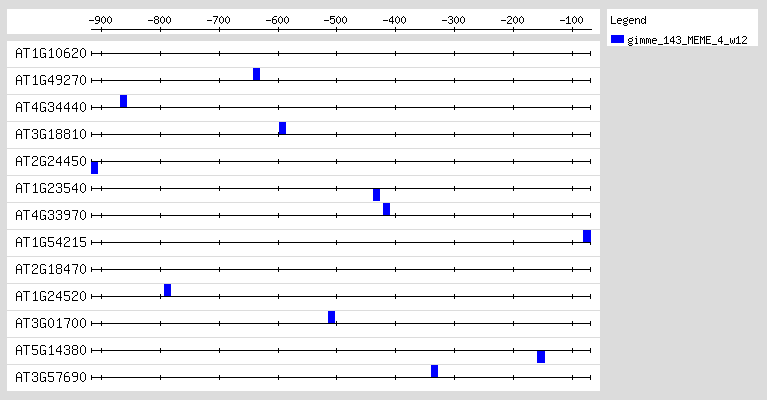
**Figure 6. Sequence visualization of motifs**

**OVERLAP OF FILTER A AND FILTER B RESULTS:**

Finally, we used TomTom to check for similarities between motifs reported using strategies “Filter A” and “Filter B”. Only one motif was found that showed high similarity (Yichao: what is the p-value/significance value for the TomTom match?). The motif is shown in table 4. Figure 7 shows the patterns of the locations of this motif in the HRGP promoter sequences.

**Table 4. Overlapped motif in Table 2 and Table 3.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Motif Name | Motif Logo | Foreground Coverage | Background Coverage | Relative Frequency | Matched TFBS  p-value |
| gimme\_143\_MEME\_4\_w12 |  | 85% (11) | 12% (16) | 6.98 | TGA10 (6.28e-03)  GBF5 (3.06e-03) |
|  |  |  |  |  |  |



**Figure 7. Sequence visualization of gimme\_143\_meme\_4\_w12**

# Methods

The overall bioinformatics workflow was presented in Figure 1. The first step of the workflow consists of identifying polllen-specific HRGP genes. We utilized the public Arabidopsis gene expression database Araport11, containing 133 RNA-seq datasets The gene expression profile was retrieved from Araport11 using the python API, intermine.webservice [16]. A list of 166 HRGP genes was retrieved from [17] . To determine pollen-specific expression, the tissue specificity index, Tau was calculated for the 166 HRGP genes. In a recent benchmarking comparison, Tau was found to be the most robust and biologically relevant method. An HRGP gene was called pollen-specific if its Tau value was greater than 0.85 and the maximally expressed tissue was pollen. All expression values were log-transformed before calculating Tau; values less than 1 were set to 0. Using these criteria, 13 pollen-specific HRGP genes were identified. The background gene set consists of 132 HRGP genes that are never expressed in pollen. Table 5 lists the pollen-specific HRGP genes.

**Table 5. List of pollen-specific HRGP genes. \*These genes have been reported to be pollen-specific in [17].a Gene Name is adopted from [17], where some genes are renamed by the authors to indicate their protein sequence properties. bTissue specificity index Tau is calculated using the formula presented in [18]. cExpression is represented using the median value after log 2 transformation. dExpression value is compared to number of standard deviation (std) away from the mean value in all genes’ expression profile in pollen. Extremely high expressed genes have more than 3\*std away from the mean and high expressed genes have more than 2\*std but less than 3\*std away from the mean.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TAIR ID | Gene Namea | Tissue Specificity Index (Tau) b | Expression in Pollen c | Level of Expression d |
| At1g10620\* | PERK11 | 0.968 | 7.003 | Extremely high |
| At1g49270\* | PERK7 | 0.960 | 8.711 | Extremely high |
| At4g34440\* | PERK5 | 0.948 | 7.039 | Extremely high |
| At3g18810\* | PERK6 | 0.941 | 9.343 | Extremely high |
| At2g24450\* | FLA3 | 0.936 | 12.284 | Extremely high |
| At1g23540\* | PERK12 | 0.936 | 7.141 | Extremely high |
| At4g33970\* | PEX4 | 0.909 | 9.893 | Extremely high |
| At1g54215 | EXT32 | 0.908 | 5.898 | High |
| At2g18470\* | PERK4 | 0.880 | 9.803 | Extremely high |
| At1g24520\* | AGP50C | 0.879 | 13.274 | Extremely high |
| At3g01700\* | AGP11C | 0.872 | 13.079 | Extremely high |
| At5g14380\* | AGP6C | 0.862 | 13.008 | Extremely high |
| At3g57690\* | AGP23P | 0.856 | 14.220 | Extremely high |

To perform motif discovery, we then constructed a list of 132 HRGP genes that were never expressed in pollen. Promoters were retrieved from Ensemble Plant v35 Biomart web interface using gene stable ID, Flank (Gene) Coding Region, and Upstream flank 1000 bp. To identify regulatory motifs for pollen-specific HRGPs, an ensemble method of de novo motif discovery was applied. GimmeMotifs is a python package that integrates multiple motif discovery tools, including MEME, Weeder, BioProspector, AMD, Homer, GADEM, MDmodule, Improbizer. The above tools are generative motif discovery tools, meaning that a background set of sequences is not required. Therefore, two additional discriminative motif discovery tools were used; they were DECOD, and DME. It is important to keep a diverse type of algorithms in an ensemble setting. MEME, Improbizer are expectation maximization based algorithms. BioProspector is a Gibbs sampling based algorithm. GADEM is based on genetic algorithm. Weeder, DECOD, and DME are word enumeration based algorithms. AMD, Homer, and MDmodule contain a series of optimization steps, including local optimizations and filtering processes. Motif width ranging from 6 nt to 16 nt were searched.

Our bioinformatics workflow now diverges to support two different motif selection filter strategies. In Figure 1, these two strategies are referred to as “Filter A” (red) and “Filter B” (yellow). The main difference between those two filters is in the sets of motifs that are considered before applying any further filtering. Filter A considers the set of all the generated motfis and then filters based on given thresholds, Filter B first reduces this set by only considering motifs that show statistically significant similarity to TFBSs for transcription factors known to be expressed in pollen.

Filter A is based on thresholds for minimum occurrences of motifs in pollen-specific genes and maximum occurrences in non-pollen specific genes. The minimum threshold was set to 11 and the maximum threshold to 30. This allows us to find a large set of possible motifs from all HRGP genes and eliminate any that prove to occur in non-pollen tissues, leaving pollen specific motifs. This resulted overall in 12 motifs.

Filter B (yellow) is based on a more sophisticated filtering strategy. First, used the motif database from plantTFDB[19] to match the 3519 motifs to known TFBSs where the corresponding TFs were expressed in pollen. Out of 619 TFs in *Arabidopsis thaliana*, 99 TFs were expressed in pollen. Significantly similar motifs were identified using Tomtom with a p-value less than 0.005. The remaining 1341 motifs were further examined by conservation analysis. The homologs of the 13 pollen-specific HRGP genes in *Arabidopsis Lyrata* were retrieved from Ensembl Plants Biomart. Then motifs were filtered if they occurred in less than 10 promoters of the pollen-specific HRGPs or more than 19 promoters of the non-pollen HRGPs. Lastly, redundant motifs were removed (retaining motifs that covered the largest amount of the foreground sequences); 3 putative motifs were identified. Note that the motifs that were removed due to redundancy may be variations of a binding site for the same TF. Thus, we report the redundant motifs in the rows of the following table:

|  |  |  |
| --- | --- | --- |
| **Motif 1** | **Motif 2** | **Motif 3** |
| gimme\_143\_MEME\_4\_w12 | DECOD\_Motif4\_13 | DME\_AHGGTGRWR |
| DME\_ARRTCYKVRG | gimme\_30\_AMD\_Motif5 |  |

The similar motifs in the first row (gimme\_143\_MEME\_4\_w12, DECOD\_Motif4\_13 and DME\_AHGGTGRWR) cover all 13 of the promoters. However, all 3 co-occur in only 6/13 of the promoters, thus lending credibility to hypothesis that these may be variants of the binding site of the same TF.

However, the similar motifs in the second row (DME\_ARRTCYKVRG and gimme\_30\_AMD\_Motif5)

co-occur in 9/13 of the foreground promoters, possibly indicating that they are the same binding site.

Conservation analysis was performed using the method adopted by Roy et, al[9]. Orthologous information between *Thaliana* and *Lyrata* were retrieved from Ensembl Plant Biomart v39. CLUSTALW2 was used to do multiple sequence alignment with gap open penalty of 10 and extension penalty of 0.1. A motif was defined as conserved if it occurred at the same position in the orthologous promoter alignment.

# Conclusion

Based on the results for the two different filter strategies, the Filter A strategy, which considers only motifs that match known TFBSs for pollen-specific transcription factors, identified 12 motifs, 4 of which we would recommend further examination of: gimme\_105\_Improbizer\_AACACACGTTTATTAGATGTTT, DME\_GRHTGDTGA, DME\_GADGAYKAS, and gimme\_143\_MEME\_4\_w12 in that order. These were all found to have a related known transcription factor binding site that has a known function related to pollen or pollen development. gimme\_105\_Improbizer\_AACACACGTTTATTAGATGTTT shows the most promise with 8 significant transcription factor binging site matches. Additionally, it’s related to CAMTA5 which is known to be pollen specific. Filter B, which does not restrict the original set of generated motitfs, resulted in 3 putative motifs which we recommend for further examination. Finally, one motif (gimme\_143\_MEME\_4\_w12) was found that was reported by both filter strategies.

# References

[1] Zambelli F, Pesole G, Pavesi G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Briefings in Bioinformatics. 2013;14(2):225-237. doi:10.1093/bib/bbs016.

[2] Simon J. van Heeringen, Gert Jan C. Veenstra; GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments, Bioinformatics, Volume 27, Issue 2, 15 January 2011, Pages 270–271, <https://doi.org/10.1093/bioinformatics/btq636>

[3] Peter Huggins, Shan Zhong, Idit Shiff, Rachel Beckerman, Oleg Laptenko, Carol Prives, Marcel H. Schulz, Itamar Simon, Ziv Bar-Joseph; DECOD: fast and accurate discriminative DNA motif finding, Bioinformatics, Volume 27, Issue 17, 1 September 2011, Pages 2361–2367, <https://doi.org/10.1093/bioinformatics/btr412>

[4] Smith AD, Sumazin P, Zhang MQ. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(5):1560-1565. doi:10.1073/pnas.0406123102.

[5] P. Ravindran, V. Verma, P. Stamm, and P. P. Kumar, “A Novel RGL2–DOF6 Complex Contributes to Primary Seed Dormancy in Arabidopsis thaliana by Regulating a GATA Transcription Factor,” *Mol. Plant*, vol. 10, no. 10, pp. 1307–1320, 2017.

[6] A. Nishii, M. Takemura, H. Fujita, M. Shikata, A. Yokota, and T. Kohchi, “Characterization of a novel gene encoding a putative single zinc-finger protein, ZIM, expressed during the reproductive phase in Arabidopsis thaliana,” *Biosci. Biotechnol. Biochem.*, vol. 64, no. 7, p. 1402—1409, Jul. 2000.

[7] Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biology. 2007;8(2):R24. doi:10.1186/gb-2007-8-2-r24

[8] Mitsuda, N., et al. (2003). "Arabidopsis CAMTA family proteins enhance V-PPase expression in pollen." Plant Cell Physiol 44(10): 975-981.

[9] Murmu, J., et al. (2010). "Arabidopsis Basic Leucine-Zipper Transcription Factors TGA9 and TGA10 Interact with Floral Glutaredoxins ROXY1 and ROXY2 and Are Redundantly Required for Anther Development." Plant Physiol 154(3): 1492-1504.

[10] (2013-02-11). "Locus: AT2G04038." 2018, from https://www.arabidopsis.org/servlets/TairObject?type=locus&name=At2g04038.

[11] Weltmeier, F., et al. (2009). "Expression patterns within the Arabidopsis C/S1 bZIP transcription factor network: availability of heterodimerization partners controls gene expression during stress response and development." Plant Mol Biol 69(1-2): 107-119.

[12] Ye, Q., et al. (2010). "Brassinosteroids control male fertility by regulating the expression of key genes involved in Arabidopsis anther and pollen development." Proceedings of the National Academy of Sciences 107(13): 6100-6105.

[13] Skubacz, A., et al. (2016). "The Role and Regulation of ABI5 (ABA-Insensitive 5) in Plant Development, Abiotic Stress Responses and Phytohormone Crosstalk." Front Plant Sci 7.

[14] (2017-10-19). "Locus: AT2G24260." 2018, from <https://www.arabidopsis.org/servlets/TairObject?type=locus&name=At2g24260>.1504.

[15] Rampey, R. A., et al. (2006). "An Arabidopsis Basic Helix-Loop-Helix Leucine Zipper Protein Modulates Metal Homeostasis and Auxin Conjugate Responsiveness." Genetics 174(4): 1841-1857.

[16] Cheng, Chia-Yi & Krishnakumar, Vivek & Chan, Agnes & Schobel, Seth & Town, Chris. (2016). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. bioRxiv. 10.1101/047308.

[17] Showalter, A. M. Keppler, B., Lichtenberg, J, Gu, D. and Welch, L. R., A bioinformatics approach to the identification, classification, and analysis of hydroxyproline-rich glycoproteins., *Plant Physiol.*, vol. 153, no. 2, pp. 485–513, Jun. 2010.

[18] Yanai I., et al., “Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification,” Bioinformatics, vol. 21, no. 5, p. 650, 2005.

[19] J. Jin *et al.*, “PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D1040–D1045, 2017.

Supplement File 1, PWMs of all the motifs with 100% foreground coverage.

MEME version 4.4

ALPHABET= ACGT

strands: + -

Background letter frequencies (from web form):

A 0.25000 C 0.25000 G 0.25000 T 0.25000

MOTIF gimme\_100\_Improbizer\_ACGAAAGAGAGAGAAAAG

letter-probability matrix: alength= 4 w= 18 nsites= 100 E= 0.001

0.382999468002 0.306999772001 0.00300098799605 0.306999772001

0.306999772001 0.382999468002 0.00300098799605 0.306999772001

0.382999468002 0.0790006839973 0.382999468002 0.155000379998

0.534998860005 0.231000076 0.0790006839973 0.155000379998

0.534998860005 0.231000076 0.0790006839973 0.155000379998

0.458999164003 0.00300098799605 0.231000076 0.306999772001

0.00300399197601 0.0790797634444 0.914912252604 0.00300399197601

0.914912252604 0.00300399197601 0.0790797634444 0.00300399197601

0.00300399197601 0.00300399197601 0.990988024072 0.00300399197601

0.990988024072 0.00300399197601 0.00300399197601 0.00300399197601

0.00300399197601 0.00300399197601 0.914912252604 0.0790797634444

0.914912252604 0.00300399197601 0.00300399197601 0.0790797634444

0.0790797634444 0.155155534913 0.762760709667 0.00300399197601

0.610998556006 0.0790006839973 0.0790006839973 0.231000076

0.534998860005 0.00300098799605 0.231000076 0.231000076

0.458999164003 0.00300098799605 0.458999164003 0.0790006839973

0.534998860005 0.0790006839973 0.231000076 0.155000379998

0.231000076 0.0790006839973 0.382999468002 0.306999772001

MOTIF gimme\_140\_MEME\_1\_w12

letter-probability matrix: alength= 4 w= 12 nsites= 100 E= 0.001

0.07697631498 7.68994155644e-05 0.922869886189 7.68994155644e-05

0.692171639496 7.68994155644e-05 0.307674561673 7.68994155644e-05

7.68994155644e-05 7.68994155644e-05 0.845970470624 0.153875730544

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.153875730544 7.68994155644e-05 0.845970470624 7.68994155644e-05

0.922869886189 7.68994155644e-05 0.07697631498 7.68994155644e-05

0.153875730544 7.68994155644e-05 0.845970470624 7.68994155644e-05

0.76907105506 7.68994155644e-05 0.153875730544 0.07697631498

7.68994155644e-05 7.68994155644e-05 0.922869886189 0.07697631498

0.76907105506 7.68994155644e-05 0.07697631498 0.153875730544

0.845970470624 7.68994155644e-05 7.68994155644e-05 0.153875730544

0.461473392802 7.68994155644e-05 0.538372808367 7.68994155644e-05

MOTIF gimme\_141\_MEME\_2\_w12

letter-probability matrix: alength= 4 w= 12 nsites= 100 E= 0.001

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.922869886189 7.68994155644e-05 0.07697631498 7.68994155644e-05

0.922869886189 0.07697631498 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.76907105506 7.68994155644e-05 0.153875730544 0.07697631498

0.461473392802 7.68994155644e-05 0.461473392802 0.07697631498

0.922869886189 7.68994155644e-05 0.07697631498 7.68994155644e-05

0.845970470624 7.68994155644e-05 0.153875730544 7.68994155644e-05

0.153875730544 7.68994155644e-05 0.845970470624 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.384573977238 0.615272223931 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

MOTIF gimme\_136\_MEME\_7\_w10

letter-probability matrix: alength= 4 w= 10 nsites= 100 E= 0.001

0.922869886189 7.68994155644e-05 0.07697631498 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

7.68994155644e-05 0.999769301753 7.68994155644e-05 7.68994155644e-05

0.922869886189 7.68994155644e-05 7.68994155644e-05 0.07697631498

0.692171639496 7.68994155644e-05 0.307674561673 7.68994155644e-05

0.845970470624 7.68994155644e-05 0.153875730544 7.68994155644e-05

7.68994155644e-05 0.153875730544 0.845970470624 7.68994155644e-05

0.692171639496 0.153875730544 0.153875730544 7.68994155644e-05

0.384573977238 0.615272223931 7.68994155644e-05 7.68994155644e-05

MOTIF gimme\_20\_BioProspector\_w12\_5

letter-probability matrix: alength= 4 w= 12 nsites= 100 E= 0.001

0.122600509598 0.348699605202 0.00350098599606 0.525198899204

0.00750172016514 0.00350133612827 0.981495223541 0.00750172016514

0.985495607578 0.00350133612827 0.00350133612827 0.00750172016514

0.352699589202 0.00350098599606 0.463699145203 0.180100279599

0.00750172016514 0.00350133612827 0.981495223541 0.00750172016514

0.985495607578 0.00350133612827 0.00350133612827 0.00750172016514

0.237623811886 0.00350133612827 0.00350133612827 0.755373515858

0.00750096999612 0.118600525598 0.578798684805 0.295099819601

0.985495607578 0.00350133612827 0.00350133612827 0.00750172016514

0.00750096999612 0.00350098599606 0.636298454806 0.352699589202

0.237576292066 0.0610946461568 0.348664738867 0.35266432291

0.410199359203 0.00350098599606 0.118600525598 0.467699129203

MOTIF gimme\_142\_MEME\_3\_w12

letter-probability matrix: alength= 4 w= 12 nsites= 100 E= 0.001

0.461473392802 0.538372808367 7.68994155644e-05 7.68994155644e-05

7.68994155644e-05 0.845970470624 0.153875730544 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.153875730544 0.538372808367 0.230775146109 0.07697631498

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.76907105506 7.68994155644e-05 0.07697631498 0.153875730544

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

7.68994155644e-05 0.538372808367 0.230775146109 0.230775146109

0.692171639496 0.307674561673 7.68994155644e-05 7.68994155644e-05

MOTIF gimme\_133\_MEME\_4\_w10

letter-probability matrix: alength= 4 w= 10 nsites= 100 E= 0.001

7.68994155644e-05 0.461473392802 0.384573977238 0.153875730544

0.384573977238 7.68994155644e-05 0.615272223931 7.68994155644e-05

0.692171639496 7.68994155644e-05 0.307674561673 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

7.68994155644e-05 7.68994155644e-05 0.999769301753 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.307674561673 0.07697631498 0.230775146109 0.384573977238

7.68994155644e-05 7.68994155644e-05 0.845970470624 0.153875730544

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

7.68994155644e-05 0.07697631498 0.922869886189 7.68994155644e-05

MOTIF gimme\_153\_MEME\_4\_w14

letter-probability matrix: alength= 4 w= 14 nsites= 100 E= 0.001

7.68994155644e-05 0.307674561673 7.68994155644e-05 0.692171639496

7.68994155644e-05 7.68994155644e-05 7.68994155644e-05 0.999769301753

7.68994155644e-05 0.307674561673 7.68994155644e-05 0.692171639496

7.68994155644e-05 0.230775146109 7.68994155644e-05 0.76907105506

7.68994155644e-05 0.07697631498 7.68994155644e-05 0.922869886189

0.07697631498 0.230775146109 0.230775146109 0.461473392802

7.68994155644e-05 7.68994155644e-05 0.07697631498 0.922869886189

7.68994155644e-05 0.07697631498 7.68994155644e-05 0.922869886189

7.68994155644e-05 7.68994155644e-05 7.68994155644e-05 0.999769301753

7.68994155644e-05 0.153875730544 7.68994155644e-05 0.845970470624

7.68994155644e-05 0.76907105506 7.68994155644e-05 0.230775146109

7.68994155644e-05 0.153875730544 7.68994155644e-05 0.845970470624

7.68994155644e-05 7.68994155644e-05 7.68994155644e-05 0.999769301753

0.07697631498 0.307674561673 0.461473392802 0.153875730544

MOTIF gimme\_151\_MEME\_2\_w14

letter-probability matrix: alength= 4 w= 14 nsites= 100 E= 0.001

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.384573977238 0.230775146109 0.384573977238 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.845970470624 7.68994155644e-05 0.153875730544 7.68994155644e-05

0.845970470624 0.07697631498 7.68994155644e-05 0.07697631498

0.538372808367 7.68994155644e-05 0.307674561673 0.153875730544

0.76907105506 7.68994155644e-05 0.230775146109 7.68994155644e-05

0.922869886189 7.68994155644e-05 0.07697631498 7.68994155644e-05

0.384573977238 0.07697631498 0.538372808367 7.68994155644e-05

0.845970470624 0.153875730544 7.68994155644e-05 7.68994155644e-05

0.07697631498 0.922869886189 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

MOTIF gimme\_105\_Improbizer\_AACACACGTTTATTAGATGTTT

letter-probability matrix: alength= 4 w= 22 nsites= 100 E= 0.001

0.610998556006 0.231000076 0.0790006839973 0.0790006839973

0.382999468002 0.155000379998 0.382999468002 0.0790006839973

0.155155534913 0.762760709667 0.00300399197601 0.0790797634444

0.990988024072 0.00300399197601 0.00300399197601 0.00300399197601

0.155155534913 0.838836481135 0.00300399197601 0.00300399197601

0.534998860005 0.231000076 0.155000379998 0.0790006839973

0.00300399197601 0.914912252604 0.0790797634444 0.00300399197601

0.00300399197601 0.00300399197601 0.914912252604 0.0790797634444

0.00300399197601 0.0790797634444 0.00300399197601 0.914912252604

0.306999772001 0.0790006839973 0.155000379998 0.458999164003

0.231000076 0.306999772001 0.00300098799605 0.458999164003

0.838836481135 0.00300399197601 0.0790797634444 0.0790797634444

0.231000076 0.00300098799605 0.155000379998 0.610998556006

0.155000379998 0.00300098799605 0.382999468002 0.458999164003

0.458999164003 0.382999468002 0.00300098799605 0.155000379998

0.306999772001 0.155000379998 0.306999772001 0.231000076

0.458999164003 0.00300098799605 0.231000076 0.306999772001

0.155000379998 0.231000076 0.155000379998 0.458999164003

0.231000076 0.0790006839973 0.458999164003 0.231000076

0.155000379998 0.231000076 0.306999772001 0.306999772001

0.155000379998 0.0790006839973 0.382999468002 0.382999468002

0.231000076 0.231000076 0.00300098799605 0.534998860005

MOTIF gimme\_25\_BioProspector\_w14\_5

letter-probability matrix: alength= 4 w= 14 nsites= 100 E= 0.001

0.130000479998 0.00380098479606 0.85819756721 0.00800096799613

0.374162087143 0.00380060473711 0.430956180557 0.191081127563

0.740324006303 0.247875220977 0.00380060473711 0.00800016798253

0.923404965884 0.00380060473711 0.00380060473711 0.0689938246422

0.191100235599 0.308899764401 0.491999032004 0.00800096799613

0.679298282807 0.00380098479606 0.064800740797 0.2520999916

0.00800016798253 0.00380060473711 0.797118099718 0.191081127563

0.191081127563 0.614037140137 0.00380060473711 0.191081127563

0.00800096799613 0.369999520002 0.491999032004 0.130000479998

0.557243046723 0.00380060473711 0.00380060473711 0.435155743803

0.00800096799613 0.00380098479606 0.491999032004 0.496199015204

0.740324006303 0.00380060473711 0.247875220977 0.00800016798253

0.618236703383 0.00380060473711 0.369962523898 0.00800016798253

0.130000479998 0.064800740797 0.797197811209 0.00800096799613

MOTIF gimme\_157\_MEME\_8\_w14

letter-probability matrix: alength= 4 w= 14 nsites= 100 E= 0.001

7.68994155644e-05 0.230775146109 0.76907105506 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

7.68994155644e-05 7.68994155644e-05 0.922869886189 0.07697631498

0.922869886189 7.68994155644e-05 7.68994155644e-05 0.07697631498

0.538372808367 7.68994155644e-05 0.230775146109 0.230775146109

7.68994155644e-05 0.692171639496 0.307674561673 7.68994155644e-05

0.692171639496 0.07697631498 0.07697631498 0.153875730544

0.461473392802 7.68994155644e-05 0.384573977238 0.153875730544

7.68994155644e-05 0.153875730544 0.845970470624 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.615272223931 0.07697631498 0.230775146109 0.07697631498

0.07697631498 0.538372808367 0.384573977238 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.538372808367 7.68994155644e-05 0.307674561673 0.153875730544

MOTIF gimme\_135\_MEME\_6\_w10

letter-probability matrix: alength= 4 w= 10 nsites= 100 E= 0.001

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.384573977238 7.68994155644e-05 0.615272223931 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.76907105506 0.230775146109 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.999769301753 7.68994155644e-05 7.68994155644e-05 7.68994155644e-05

0.461473392802 7.68994155644e-05 0.538372808367 7.68994155644e-05

0.922869886189 7.68994155644e-05 0.07697631498 7.68994155644e-05

MOTIF DECOD\_Motif1\_14

letter-probability matrix: alength= 4 w= 14 nsites= 100 E= 0.001

0.0000 0.1500 0.2000 0.6500

0.0500 0.4500 0.2000 0.3000

0.2000 0.0500 0.0000 0.7500

0.0000 0.2000 0.0000 0.8000

0.0000 0.7000 0.3000 0.0000

0.2500 0.0000 0.0000 0.7500

0.0500 0.4500 0.0000 0.5000

0.0000 0.2000 0.1500 0.6500

0.0000 0.5000 0.1500 0.3500

0.0500 0.2000 0.0000 0.7500

0.0500 0.4500 0.2000 0.3000

0.0000 0.1500 0.0000 0.8500

0.0000 0.5000 0.0000 0.5000

0.2000 0.3000 0.0500 0.4500

MOTIF DECOD\_Motif1\_15

letter-probability matrix: alength= 4 w= 15 nsites= 100 E= 0.001

0.0000 0.1500 0.2500 0.6000

0.1000 0.2500 0.2500 0.4000

0.0000 0.2000 0.1000 0.7000

0.0500 0.4500 0.2000 0.3000

0.2000 0.0000 0.0000 0.8000

0.0000 0.2000 0.0000 0.8000

0.0000 0.6000 0.4000 0.0000

0.2500 0.0000 0.0000 0.7500

0.0000 0.4500 0.0000 0.5500

0.0000 0.2000 0.1500 0.6500

0.0000 0.5500 0.1000 0.3500

0.1000 0.2000 0.0000 0.7000

0.1500 0.4000 0.2000 0.2500

0.0000 0.1500 0.0000 0.8500

0.0000 0.5000 0.0500 0.4500

MOTIF DECOD\_Motif2\_15

letter-probability matrix: alength= 4 w= 15 nsites= 100 E= 0.001

0.3000 0.0000 0.1000 0.6000

0.0000 0.1000 0.8500 0.0500

0.3500 0.0000 0.2500 0.4000

0.0000 0.0500 0.2000 0.7500

0.0000 0.3500 0.1500 0.5000

0.0500 0.0500 0.2000 0.7000

0.0000 0.0000 0.1500 0.8500

0.1000 0.0000 0.8000 0.1000

0.6500 0.0000 0.1500 0.2000

0.2500 0.1500 0.3500 0.2500

0.0500 0.1500 0.5000 0.3000

0.1500 0.0000 0.2000 0.6500

0.0000 0.0000 0.2000 0.8000

0.0000 0.0500 0.2000 0.7500

0.1500 0.2000 0.3500 0.3000

MOTIF DECOD\_Motif1\_16

letter-probability matrix: alength= 4 w= 16 nsites= 100 E= 0.001

0.4000 0.1500 0.1500 0.3000

0.5000 0.0000 0.4500 0.0500

0.9000 0.0000 0.1000 0.0000

0.3500 0.2500 0.3000 0.1000

0.7000 0.0000 0.2500 0.0500

0.4000 0.1000 0.5000 0.0000

0.7000 0.1000 0.2000 0.0000

0.5500 0.0000 0.4500 0.0000

0.8500 0.0000 0.0000 0.1500

0.1000 0.3500 0.5500 0.0000

0.7500 0.0500 0.2000 0.0000

0.7500 0.0000 0.0500 0.2000

0.3500 0.2000 0.4000 0.0500

0.7500 0.0500 0.2000 0.0000

0.4000 0.2500 0.3000 0.0500

0.6000 0.2000 0.1500 0.0500

MOTIF DECOD\_Motif3\_16

letter-probability matrix: alength= 4 w= 16 nsites= 100 E= 0.001

0.2000 0.3000 0.0000 0.5000

0.0500 0.3000 0.0000 0.6500

0.0000 0.4000 0.0500 0.5500

0.1500 0.5500 0.1500 0.1500

0.4000 0.2500 0.0000 0.3500

0.3000 0.2500 0.1000 0.3500

0.4500 0.5000 0.0000 0.0500

0.0000 0.0000 0.0000 1.0000

0.0500 0.5500 0.2500 0.1500

0.4000 0.0000 0.0000 0.6000

0.0000 0.3500 0.0500 0.6000

0.0500 0.7500 0.0000 0.2000

0.2500 0.4500 0.1500 0.1500

0.1000 0.0000 0.0000 0.9000

0.0000 0.6000 0.2000 0.2000

0.4000 0.1000 0.2000 0.3000

MOTIF DECOD\_Motif2\_16

letter-probability matrix: alength= 4 w= 16 nsites= 100 E= 0.001

0.4500 0.2500 0.1500 0.1500

0.6000 0.0500 0.1000 0.2500

0.2000 0.4000 0.2000 0.2000

0.5000 0.4500 0.0000 0.0500

0.8500 0.1000 0.0500 0.0000

0.8000 0.2000 0.0000 0.0000

0.6000 0.2500 0.1500 0.0000

0.4000 0.4000 0.1000 0.1000

0.3500 0.0500 0.0000 0.6000

0.0500 0.7500 0.1000 0.1000

0.8000 0.1500 0.0000 0.0500

0.7000 0.2000 0.1000 0.0000

0.7500 0.1000 0.1500 0.0000

0.6500 0.2000 0.1500 0.0000

0.4000 0.3500 0.0000 0.2500

0.0500 0.7000 0.1000 0.1500

MOTIF DME\_TWTTTKTTCTT

letter-probability matrix: alength= 4 w= 11 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.5 0.0 0.0 0.5

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.5 0.5

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 1.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_TTTTATTTRKTT

letter-probability matrix: alength= 4 w= 12 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

1.0 0.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.75 0.0 0.25 0.0

0.0 0.0 0.25 0.75

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_TKYKYTTTCT

letter-probability matrix: alength= 4 w= 10 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.0 0.0 0.333333333333 0.666666666667

0.0 0.333333333333 0.0 0.666666666667

0.0 0.0 0.333333333333 0.666666666667

0.0 0.222222222222 0.0 0.777777777778

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 1.0 0.0 0.0

0.0 0.0 0.0 1.0

MOTIF DME\_TTTKYTCTTTT

letter-probability matrix: alength= 4 w= 11 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.25 0.75

0.0 0.25 0.0 0.75

0.0 0.0 0.0 1.0

0.0 1.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_TDTTTTMTTTTTS

letter-probability matrix: alength= 4 w= 13 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.571428571429 0.0 0.285714285714 0.142857142857

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.571428571429 0.428571428571 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.428571428571 0.571428571429 0.0

MOTIF DME\_YTRTTTTATTTTT

letter-probability matrix: alength= 4 w= 13 nsites= 100 E= 0.001

0.0 0.25 0.0 0.75

0.0 0.0 0.0 1.0

0.75 0.0 0.25 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

1.0 0.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_KTTTCTTTTT

letter-probability matrix: alength= 4 w= 10 nsites= 100 E= 0.001

0.0 0.0 0.285714285714 0.714285714286

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 1.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_TRTTTTMTTYTT

letter-probability matrix: alength= 4 w= 12 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.333333333333 0.0 0.666666666667 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.5 0.5 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.166666666667 0.0 0.833333333333

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_KTTTTTTSTTT

letter-probability matrix: alength= 4 w= 11 nsites= 100 E= 0.001

0.0 0.0 0.333333333333 0.666666666667

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.5 0.5 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_KATTTTATTTKT

letter-probability matrix: alength= 4 w= 12 nsites= 100 E= 0.001

0.0 0.0 0.2 0.8

1.0 0.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

1.0 0.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.4 0.6

0.0 0.0 0.0 1.0

MOTIF DME\_TTTTTBTTKTT

letter-probability matrix: alength= 4 w= 11 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.333333333333 0.333333333333 0.333333333333

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.833333333333 0.166666666667

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_AGAAAAARRA

letter-probability matrix: alength= 4 w= 10 nsites= 100 E= 0.001

1.0 0.0 0.0 0.0

0.0 0.0 1.0 0.0

1.0 0.0 0.0 0.0

1.0 0.0 0.0 0.0

1.0 0.0 0.0 0.0

1.0 0.0 0.0 0.0

1.0 0.0 0.0 0.0

0.444444444444 0.0 0.555555555556 0.0

0.777777777778 0.0 0.222222222222 0.0

1.0 0.0 0.0 0.0

MOTIF DME\_TGTTTTVTTT

letter-probability matrix: alength= 4 w= 10 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.0 0.0 1.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.5 0.333333333333 0.166666666667 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_TTTTATTTTTM

letter-probability matrix: alength= 4 w= 11 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

1.0 0.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.5 0.5 0.0 0.0

MOTIF DME\_TKTTTCTDTT

letter-probability matrix: alength= 4 w= 10 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.0 0.0 0.444444444444 0.555555555556

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 1.0 0.0 0.0

0.0 0.0 0.0 1.0

0.111111111111 0.0 0.111111111111 0.777777777778

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_TYTATTTTATTT

letter-probability matrix: alength= 4 w= 12 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.0 0.2 0.0 0.8

0.0 0.0 0.0 1.0

1.0 0.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

1.0 0.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

MOTIF DME\_TRTTTTATTTK

letter-probability matrix: alength= 4 w= 11 nsites= 100 E= 0.001

0.0 0.0 0.0 1.0

0.5 0.0 0.5 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

1.0 0.0 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.166666666667 0.833333333333

MOTIF DME\_RTTTTMTTTTTS

letter-probability matrix: alength= 4 w= 12 nsites= 100 E= 0.001

0.666666666667 0.0 0.333333333333 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.666666666667 0.333333333333 0.0 0.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.0 0.0 1.0

0.0 0.333333333333 0.666666666667 0.0