The word landscape of the non-coding segments of the *Arabidopsis thaliana* genome

Jens Lichtenberg*1, Alper Yilmaz2, Joshua D Welch1, Kyle Kurz1, Xiaoyu Liang1, Frank Drews1, Klaus Ecker1, Stephen S Lee3, Matt Geisler4, Erich Grotewold2 and Lonnie R Welch1,5,6

Address: 1Bioinformatics Laboratory, School of Electrical Engineering and Computer Science, Ohio University, Athens, Ohio, USA, 2Department of Plant Cellular and Molecular Biology, Plant Biotechnology Center, The Ohio State University, Columbus, Ohio, USA, 3Department of Statistics, University of Idaho, Moscow, Idaho, USA, 4Department of Plant Biology, Southern Illinois University, Carbondale, Illinois, USA, 5Biomedical Engineering Program, Ohio University, Athens, Ohio, USA and 6Molecular and Cellular Biology Program, Ohio University, Athens, Ohio, USA

E-mail: Jens Lichtenberg* - lichtenj@ohio.edu; Alper Yilmaz - yilmaz.11@osu.edu; Joshua D Welch - jw156605@ohio.edu; Kyle Kurz - kk372703@ohio.edu; Xiaoyu Liang - xl187007@ohio.edu; Frank Drews - drews@ohio.edu; Klaus Ecker - ecker@ohio.edu; Stephen S Lee - stevel@uidaho.edu; Matt Geisler - mgeisler@plant.siu.edu; Erich Grotewold - grotewold.1@osu.edu; Lonnie R Welch - welch@ohio.edu

*Corresponding author

Published: 8 October 2009

BMC Genomics 2009, 10:463 doi: 10.1186/1471-2164-10-463

Received: 23 January 2009
Accepted: 8 October 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/463

© 2009 Lichtenberg et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Abstract**

**Background:** Genome sequences can be conceptualized as arrangements of motifs or words. The frequencies and positional distributions of these words within particular non-coding genomic segments provide important insights into how the words function in processes such as mRNA stability and regulation of gene expression.

**Results:** Using an enumerative word discovery approach, we investigated the frequencies and positional distributions of all 65,536 different 8-letter words in the genome of *Arabidopsis thaliana*. Focusing on promoter regions, introns, and 3' and 5' untranslated regions (3'UTRs and 5'UTRs), we compared word frequencies in these segments to genome-wide frequencies. The statistically interesting words in each segment were clustered with similar words to generate motif logos. We investigated whether words were clustered at particular locations or were distributed randomly within each genomic segment, and we classified the words using gene expression information from public repositories. Finally, we investigated whether particular sets of words appeared together more frequently than others.

**Conclusion:** Our studies provide a detailed view of the word composition of several segments of the non-coding portion of the *Arabidopsis* genome. Each segment contains a unique word-based signature. The respective signatures consist of the sets of enriched words, 'unwords', and word pairs within a segment, as well as the preferential locations and functional classifications for the signature words. Additionally, the positional distributions of enriched words within the segments highlight possible functional elements, and the co-associations of words in promoter regions likely represent the formation of higher order regulatory modules. This work is an important step toward fully cataloguing the functional elements of the *Arabidopsis* genome.
Background

All genomes are composed of nucleotides, which are represented abstractly as letters (Adenine (A), Guanine (G), Cytosine (C), and Thymine (T)). Strings of such letters can be conceptualized as words, which provide the blueprints for organisms. Each word is found a specific number of times in a particular genome. Note that the expected frequency of a word is inversely related to the word's length. Some nucleotides appear more frequently than others (e.g. A/T in Arabidopsis), giving each genome a distinct (G+C)% content and biasing expected word frequencies. Higher order frequencies (dinucleotide and trinucleotide) also show distinct biases beyond those expected for single nucleotide frequencies [1].

Distinct selective pressures shape words positioned in different genomic regions. For example, a word in an open reading frame (ORF) has a direct influence on the primary amino acid sequence of a protein and hence is under strong selective pressure. In contrast, words in introns are likely to be under more relaxed selective constraints, unless they are important for gene functions, for example by providing docking sites for splicing factors [2] or for enzymes involved in the post-transcriptional processing of a transcript [3,4]. The gene sections corresponding to the 5’ and 3’ untranslated regions (5’UTRs and 3’UTRs, respectively) are also likely to be under less selective constraints than the ORFs, yet signatures of strong selection in UTRs have been described (reviewed in [5]). The constant formation of DNA microsatellites through slippage by the replication machinery, and the action of viruses and transposons, also complicate the word landscape, especially in regions with lower selective constraints (such as introns, UTRs and intergenic regions) [6,7].

This manuscript describes the results of a genome-wide analysis to discover putative regulatory words. Within this context, we define the cis-regulatory apparatus as all the DNA segments that are located proximal to a gene, and that also contribute to the gene’s expression. It is the function of transcription factors, miRNAs, or other molecules that interact with DNA, to interpret the words (sequence code) hardwired in the cis-regulatory apparatus and to ‘execute’ them, thereby generating signals to the basal transcription machinery that result in changes to the rate of RNA production by the corresponding DNA-dependent RNA polymerases. When located upstream of the transcription start site (TSS), the cis-regulatory apparatus is often referred to as the promoter of a gene.

Promoters are typically divided into three regions: core, proximal and distal. The core promoter, a region at location [+1;-100] relative to the TSS, performs a central role in the formation of pre-initiation transcriptional complexes. Immediately upstream of the core promoter is the proximal promoter, which is located at position [-101;-1000] relative to the TSS and serves as a docking site for transcription factors. The distal promoter is located at [-1001;-3000] relative to the TSS and contains the regulatory elements that are commonly known as enhancers and silencers. The participation of a particular DNA segment in the regulation of gene expression can only be demonstrated experimentally. Thus, understanding the rules at play in deciphering the transcriptional regulatory code remains one of the most significant challenges in biology today.

Although most regulatory elements are present in the UTRs and upstream regions, due to their proximity to the TSS, studies have shown the presence of regulatory elements in introns, and, to a much lesser extent, in coding regions [2,8-16]. Building on this knowledge, a segment-based analysis was performed that is focused on non-coding regions within the open reading frames (i.e., introns) and flanking non-coding regions (i.e., UTRs and upstream regions). The coding regions were omitted from this analysis because they are under other selection pressures corresponding to the amino acid sequences of the proteins they produce, and thus they are subjected to biases other than regulation.

Arabidopsis thaliana provides an ideal reference organism to investigate the word landscape of a plant genome, and to relate said landscape to important biological features. The Arabidopsis genome consists of 125 Mbp arranged into five chromosomes [17,18]. The genome is well annotated and regions corresponding to introns, 3’UTRs, 5’UTRs, and intergenic genomic spaces are all available from The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org) [19].

Many studies have characterized Arabidopsis DNA sequence motifs that participate in the regulation of particular genes (e.g., [20-23]), and public databases such as AthaMap [24] and AGRIS [25] provide comprehensive collections of cis-regulatory elements likely to participate in the regulation of gene expression. However, a systematic analysis of all the words present in the Arabidopsis genome is still lacking.

To analyze the different segments of the Arabidopsis genome, an enumerative word discovery approach was used to detect statistically overrepresented words. Similar approaches have been successfully applied over the last decade in the area of motif discovery [26-37]. In a 2005 study, Tompa et al. [38] showed that enumerative methods outperformed heuristic methods in many cases. They are particularly applicable in this research, because they allow the study of the entire ‘word landscape’ of a genomic data set.
Our approach scans the sequences and produces a set of words and word frequencies. This information is employed by a Markov model to compute expected word frequencies. Words with unexpectedly high frequencies are putative functional elements, and thus they are further characterized by comparing word frequencies and positions to gene induction or suppression using the method of Geisler et al. [39]. Additionally, clusters of similar words are formed and used to create motifs for putative transcription factor binding sites. Sequences that contain the same functional elements are grouped together into putative ‘nodes’ of regulatory networks. Words that co-occur often are identified as putative transcription factor binding modules.

Results and Discussion

Distribution of 8-letter words in the Arabidopsis genome

To determine the word distributions in the segments of the Arabidopsis thaliana genome that contribute to the cis-regulatory apparatus, a comprehensive analysis of 8-letter words in the entire genome was conducted and compared with segments corresponding to non-coding regions. Words of length 6-16 were examined and the complete results have been made available via AGRIS http://arabidopsis.med.ohio-state.edu/ [25,40]. This article reports findings for words of length eight because they correspond to the typical DNA sequence length recognized by transcription factors (~64,000) to reduce false positive results, while retaining sufficient word counts to be statistically informative.

Overrepresented Words

All 8-letter words present in the segments were identified and scored using observed:expected frequency ratios (O/E). Specifically, each word was scored and ranked by using the function \( S \ln(S/E) \), where \( S \) is the number of sequences that contained the word, ‘in’ is the natural logarithm, and \( E \) is the number of sequences in which the word was expected to occur. Words discovered in the whole genome were analyzed using the \( S \ln(S/E) \) score, with \( O \) referring to the overall occurrence of a word across the entire genome and \( E \) representing the expected occurrence of that word. The 25 top-ranked

Table 1: Segment characteristics for Arabidopsis thaliana

| Data Set          | # Sequences/Chromosomes | Min. Seq. Length | Max. Seq. Length | Mean Seq. Length | Std. Deviation | Total Nucleotides | Genome Percentage |
|-------------------|--------------------------|------------------|------------------|------------------|----------------|-------------------|------------------|
| 3’ UTRs           | 19,771                   | 8                | 3,118            | 228,134          | 152,106        | 4,510,410         | 3.78             |
| 5’ UTRs           | 18,585                   | 8                | 3,214            | 140,088          | 130,288        | 2,603,531         | 2.18             |
| Introns           | 118,319                  | 8                | 10,234           | 164,446          | 178,484        | 19,457,029        | 16.32            |
| Core Promoters    | 27,023                   | 100              | 100              | 100              | 0              | 2,702,300         | 2.27             |
| Proximal Promoters| 27,023                   | 900              | 900              | 900              | 0              | 24,320,700        | 20.41            |
| Distal Promoters  | 27,025                   | 1,371            | 2,000            | 1,999.96         | 5.01105        | 54,048,839        | 45.35            |
| Genome-wide       | 5                        | 18,585,000       | 30,432,600       | 23,837,300       | 4,432,780      | 119,186,497       | 100.00           |

Overview of the characteristics properties for non-coding segments and the entire genome for Arabidopsis thaliana. The number of sequence refers to the respective number of unique sequences in the specific segment. In case of the entire genome the sequences are the complete chromosomes. Min. Seq. Length refers to the length of the shortest sequence in the set, while Max. Seq. Length refers to the length of the longest sequence in the set. Mean Seq. Length provides the average length of the sequences in the set, while Std. Deviation describes the deviation from said mean. Finally Total Nucleotides describes the total number of nucleotides contained within the sequences of the set and Genome Percentage elaborates on the relationship between the nucleotide count of the set versus the entire genome.

Some sequences in the segments are shorter than 8 nucleotides. Since these sequences cannot harbour any putative regulatory elements in the context of this study, the sequences are removed from the table. For the 3’ UTRs this results in a total of 179 nt being omitted, for 5’ UTRs 1207 nt and for introns 26 nt. They are however included in the calculation of the background for the different segments since they contribute to the overall nucleotide distribution.
words, corresponding to ~0.04% of all possible words, which also corresponds to ~0.04% of the discovered words, were taken as an exemplary subset of the results and further examined (see Tables 2, 3, 4, 5, 6, 7, &8 and Additional files 1, 2, 3, 4, 5, 6, &7).

A detailed analysis of the words identified a minimal overlap between the sets of overrepresented words for the different segments. Specifically, considering the list of top 25 words discovered in any of the six segments (and in the genome wide analysis), 175 words were unique to one specific set, 15

### Table 2: The top 25 words in 3'UTRs

| Word     | Unmasked | Masked | Unmasked | Masked | Unmasked | Masked | Unmasked | Masked |
|----------|----------|--------|----------|--------|----------|--------|----------|--------|
| TTTTGT   | 2264     | 2066.82| 2488     | 2306.27| 2066.89  | 2279   | 2066.82  | 2306.27|
| TTTTGG   | 2171     | 1981.63| 2404     | 2203.7 | 1981.63  | 2183   | 1981.63  | 2203.7 |
| TTTTTGTA | 598      | 491.47 | 603      | 516.44 | 504.21   | 621    | 531.22   | 109.069|
| TTTTTTGA | 491      | 390.18 | 504      | 408.46 | 487.02   | 621    | 531.22   | 109.069|
| AATTATTG | 588      | 491.47 | 603      | 516.44 | 504.21   | 621    | 531.22   | 109.069|
| ATTTTT   | 482      | 387.67 | 498      | 405.79 | 406.16   | 510    | 426.06   | 494.317|
| ATTTTT   | 491      | 390.18 | 504      | 408.46 | 487.02   | 621    | 531.22   | 109.069|
| TTTTCT   | 1721     | 1628.11| 1839     | 1786.09| 1722     | 1625.78| 1847     | 1798.84|
| AAAAAAT  | 564      | 474.52 | 580      | 498.32 | 498.32   | 580    | 505.26   | 97.177 |
| TTTTTCT  | 1721     | 1628.11| 1839     | 1786.09| 1722     | 1625.78| 1847     | 1798.84|
| TTTTGTT  | 2171     | 1981.63| 2404     | 2203.7 | 1981.63  | 2183   | 1981.63  | 2203.7 |
| TTTTTTG  | 491      | 390.18 | 504      | 408.46 | 487.02   | 621    | 531.22   | 109.069|
| AATTAT   | 588      | 491.47 | 603      | 516.44 | 504.21   | 621    | 531.22   | 109.069|
| ATTTTT   | 482      | 387.67 | 498      | 405.79 | 406.16   | 510    | 426.06   | 494.317|
| ATTTTT   | 491      | 390.18 | 504      | 408.46 | 487.02   | 621    | 531.22   | 109.069|
| TTTTCT   | 1721     | 1628.11| 1839     | 1786.09| 1722     | 1625.78| 1847     | 1798.84|
| AAAAAAT  | 564      | 474.52 | 580      | 498.32 | 498.32   | 580    | 505.26   | 97.177 |
| TTTTTCT  | 1721     | 1628.11| 1839     | 1786.09| 1722     | 1625.78| 1847     | 1798.84|
| TTTTGTT  | 2171     | 1981.63| 2404     | 2203.7 | 1981.63  | 2183   | 1981.63  | 2203.7 |
| TTTTTTG  | 491      | 390.18 | 504      | 408.46 | 487.02   | 621    | 531.22   | 109.069|

Top 25 overrepresented words for the 3'Untranslated Regions in Arabidopsis thaliana. The Word attribute describes the short nucleotide sequence associated with a putative word. S and SE describe the number of sequences a word occurs in and the number of sequences the word was expected to occur in respectively, while O and EO describe the total number of occurrences and the expected total number of occurrences. The score SnSES describes a statistical coverage of the sequences analyzed in the set and is based on a Markov Chain Background Model. Each set of attributes was computed for the masked as well as the unmasked version of the corresponding segment with the emphasis placed on the unmasked version (i.e. sorting of the table based on the unmasked SnSES score). Further information for the word is provided through its reverse complement (RevComp) and the position of the reverse complement in the set of results (RC_Pos) as well as a notion describing if the word is a genomic palindrome (Pal). Finally, PValues describes a p-value that is assigned in order to provide statistical insight allowing the determination if a word is relevant or was discovered as interesting by random chance.
words occurred uniquely in two sets, 7 in three sets, 4 in four sets and none in five sets. Only two words (ATTTTTTA, and AATATATT) were shared in six out of seven sets (neither word was present in the 5'UTR set). Note that the word AATATATT has a significant similarity to the sequence of the TATA-box, a regulatory element that is (1) often found in core promoters and (2) known to contribute to the correct positioning of the core transcriptional machinery [42]. It is conceivable that the absence of AATATATT in the 5'UTR set prevents the initiation of transcription at incorrect sites.

| Word      | Unmasked | Masked | Unmasked |
|-----------|----------|--------|----------|
| CTCTTCTC  | 871      | 614.433| 992      |
| CTTCTCTC  | 1154     | 1003.84| 1293     |
| AAAAAAAA  | 1051     | 920.535| 1134     |
| TTTCTTCA  | 611      | 492.734| 631      |
| GAGAAGAG  | 316      | 211.511| 360      |
| TTTCTTC   | 611      | 492.734| 631      |
| AACAAAAA  | 1078     | 972.588| 1154     |
| GAGAAGAG  | 316      | 211.511| 360      |
| TTTCTTC   | 883      | 771.778| 929      |
| CTCTCTTT  | 1229     | 1116.97| 1351     |
| AAAAAAAA  | 1078     | 972.588| 1154     |
| TTTCTTCT  | 1421     | 1308.64| 1554     |
| AGAAAAAA  | 1078     | 972.588| 1154     |
| ATCTCTCA  | 332      | 243.705| 342      |
| AAAAAACA  | 759      | 663.266| 803      |
| TTTTTTCT  | 1020     | 923.944| 1116     |
| AGAGAAAG  | 589      | 496.468| 634      |
| TTTTCTTT  | 1229     | 1116.97| 1351     |
| AAAAAAACA | 589      | 496.468| 634      |
| ACAAAAAA  | 845      | 754.352| 901      |
| TAAAAAAG  | 231      | 152.899| 238      |
| CAAAACC   | 351      | 268.225| 358      |
| AAGAAAAA  | 351      | 268.225| 358      |
| CCTCTCTT  | 907      | 817.386| 946      |
| TTCTCTCA  | 473      | 387.786| 484      |

Table 3: The top 25 words in 5'UTRs

Top 25 overrepresented words for the 5'Untranslated Regions in Arabidopsis thaliana. The Word attribute describes the short nucleotide sequence associated with a putative word. S and ES describe the number of sequences a word occurs in and the number of sequences the word was expected to occur in respectively, while O and EO describe the total number of occurrences and the expected total number of occurrences. The score SlnSES describes a statistical coverage of the sequences analyzed in the set and is based on a Markov Chain Background Model. Each set of attributes was computed for the masked as well as the unmasked version of the corresponding segment with the emphasis placed on the unmasked version (i.e. sorting of the table based on the unmasked SlnSES score).

Further information for the word is provided through its reverse complement (RevComp) and the position of the reverse complement in the set of results (RC_Pos) as well as a notion describing if the word is a genomic palindrome (Pal).

Finally, PValues describes a p-value that is assigned in order to provide statistical insight allowing the determination if a word is relevant or was discovered as interesting by random chance.
Table 4: The top 25 words in Introns

| Word      | Unmasked | Masked | Unmasked |
|-----------|----------|--------|----------|
| TTTTTGTT  | 10048    | 9365.74| 11094    |
| TTTTCTT   | 9144     | 8495.68| 10201    |
| CTCTTTTCT | 2764     | 2170.42| 2314.32  |
| TTTTTTGA  | 2673     | 2130.13| 2422.33  |
| TTTTGAGC  | 3505     | 2959.4 | 3179.19  |
| TTTTTTGG  | 7618     | 7067.97| 8198     |
| TTTTCTTT  | 9256     | 8733.23| 10299    |
| TGTTTTTT  | 7487     | 6984.58| 8028     |
| CTCTCTTT  | 3193     | 2716.79| 3289     |
| ATTTTTTA  | 2508     | 2044.78| 2645     |
| TTTTTTCC  | 3166     | 2702.47| 3253     |
| TGTTTCAG  | 2215     | 1790.21| 2239     |
| GTTTTTTG  | 2029     | 1611.17| 2092     |
| TTTTTGTT  | 12142    | 11689.3| 13879    |
| TTTGTTTT  | 11017    | 10569.9| 12527    |
| CTTTTTTA  | 2324     | 1828.76| 2282     |
| AATTATT    | 2022     | 1642.55| 2143     |
| ATTATTCA   | 2411     | 2030.35| 2467     |
| ATTTTTCC   | 2810     | 2425.9 | 2881     |
| CATTATT    | 2402     | 2023.84| 2481     |
| TTTTTTCT   | 7674     | 7280.17| 8254     |
| GTTTTGCA   | 1922     | 1563.72| 1933     |
| TTTATT     | 5647     | 5276.08| 6142     |

The top 25 overrepresented words for the Introns in Arabidopsis thaliana. The Word attribute describes the short nucleotide sequence associated with a putative word. S and ES describe the number of sequences a word occurs in and the number of sequences the word was expected to occur in respectively, while O and EO describe the total number of occurrences and the expected total number of occurrences. The score SlnSES describes a statistical coverage of the sequences analyzed in the set and is based on a Markov Chain Background Model. Each set of attributes was computed for the masked as well as the unmasked version of the corresponding segment with the emphasis placed on the unmasked version (i.e. sorting of the table based on the unmasked SlnSES score). Further information for the word is provided through its reverse complement (RevComp) and the position of the reverse complement in the set of results (RC_Pos) as well as a notion describing if the word is a genomic palindrome (Pal). Finally, PValues describes a p-value that is assigned in order to provide statistical insight allowing the determination if a word is relevant or was discovered as interesting by random chance.
Table 5: The top 25 words in Core Promoters

| Word     | S   | ES  | O   | EO  | SlnSES | S   | ES  | O   | EO  | SlnSES | RevComp | RC_Pos | Pal | PValues       |
|----------|-----|-----|-----|-----|--------|-----|-----|-----|-----|--------|---------|--------|-----|---------------|
| TATAAATA | 1355| 1071.69 | 1369 | 1175.57 | 317.831 | 1300 | 1029.92 | 1311 | 1128.85 | 302.753 | TATTTATA | 69 No    | 2.02E-08 |
| CTATAAT  | 712 | 474.27 | 716 | 514.446 | 289.286 | 704 | 464.711 | 708 | 503.987 | 292.416 | ATTTATAG | 2504 No  | 7.77E-16 |
| CTATATAA | 636 | 410.261 | 638 | 444.486 | 278.826 | 626 | 450.579 | 628 | 488.533 | 205.839 | TTATATAG | 18530 No | 1.11E-16 |
| ATATAAAC | 560 | 350.797 | 560 | 379.643 | 261.928 | 554 | 347.685 | 554 | 376.253 | 158.835 | ATTTTTTA | 12 No    | -2.22E-16 |
| ATATATAC | 544 | 394.869 | 559 | 427.688 | 174.295 | 507 | 330.093 | 515 | 357.099 | 217.573 | GTATATAT | 5651 No  | 7.41E-10 |
| AATATATT | 300 | 181.346 | 300 | 195.646 | 151.012 | 287 | 195.452 | 287 | 210.918 | 110.256 | AATATATT | 6 Yes    | 2.74E-12 |
| TTATATAA | 524 | 397.031 | 529 | 430.047 | 145.398 | 514 | 430.79  | 518 | 466.905 | 90.7739 | TTATATAG | 7 Yes    | 2.22E-06 |
| AAGAAAAA | 1261| 1129.24 | 1318 | 1240.05 | 139.165 | 1189 | 1063   | 1238 | 1165.84 | 133.189 | TTTTTCTT | 25 No    | 0.014544 |
| ATATAAAG | 378 | 262.861 | 380 | 284.014 | 137.316 | 375 | 261.811 | 377 | 282.19  | 135.643 | CTTTATAT | 377 No   | 3.41E-08 |
| TATATAAA | 1260| 1131.11 | 1276 | 1242.15 | 135.966 | 1234 | 1102.41 | 1250 | 1209.97 | 139.143 | TTATATAG | 1458 No  | 0.171817 |
| AGAAAAAA | 1127| 1000.04 | 1170 | 1095.49 | 134.693 | 1063 | 936.863 | 1099 | 1025.06 | 134.271 | TTTTTTCT | 31 No    | 0.01331 |
| ATTTTTTA | 312 | 204.097 | 315 | 220.282 | 123.294 | 314 | 207.163 | 302 | 223.604 | 116.302 | TTTTTATA | 13 Yes   | 0.001019 |
| TTATATAA | 688 | 568.245 | 696 | 617.461 | 131.571 | 658 | 543.865 | 665 | 590.725 | 125.351 | TTTTTAAA | 13 Yes   | 0.001019 |
| CTTCTCTC | 402 | 294.202 | 429 | 318.061 | 25.949 | 371 | 277.661 | 390 | 300.087 | 107.516 | GAGAAGAG | 444 No   | 1.97E-09 |
| ACAAAAAA | 958 | 840.585 | 988 | 918.052 | 25.949 | 917 | 799.552 | 939 | 872.564 | 125.681 | TTTTTGT  | 45 No    | 0.011607 |
| ATAAATAC | 578 | 466.039 | 582 | 505.44  | 25.949 | 574 | 459.992 | 578 | 498.825 | 127.095 | GTATTTAT | 14072 No | 0.000465 |
| TTATAAAA | 507 | 397.553 | 508 | 430.617 | 123.294 | 490 | 386.47  | 491 | 418.525 | 116.302 | TTTTTAAT | 945 No   | 0.000153 |
| AAATTTAA | 718 | 609.913 | 745 | 663.251 | 117.144 | 682 | 578.03  | 705 | 628.206 | 112.806 | TTTAATTT | 96 No    | 0.000967 |
| GCCCATTA | 374 | 273.89  | 396 | 295.991 | 161.512 | 372 | 272.658 | 394 | 294.653 | 151.571 | TAATGGGC | 190 No   | 1.82E-08 |
| AAAAAACCA | 893 | 787.368 | 924 | 859.073 | 112.42 | 849 | 736.927 | 874 | 803.277 | 120.193 | TGGTTTTT | 33 No    | 0.014723 |
| TAAAAAAA | 805 | 701.565 | 828 | 764.227 | 110.71 | 768 | 667.112 | 788 | 726.227 | 108.159 | TTTTTATAA | 27 No   | 0.01177 |
| ATTTAAA  | 708 | 609.58  | 719 | 662.885 | 105.969 | 671 | 581.412 | 681 | 631.921 | 96.1611 | TTTTTAAAT | 316 No   | 0.016276 |
| GCCCAATA | 322 | 231.782 | 340 | 250.291 | 105.859 | 321 | 228.286 | 337 | 246.5  | 109.41  | TATTGGGC | 130 No   | 4.26E-08 |

Top 25 overrepresented words for the core promoter regions in *Arabidopsis thaliana*. The **Word** attribute describes the short nucleotide sequence associated with a putative word. **S** and **ES** describe the number of sequences a word occurs in and the number of sequences the word was expected to occur in respectively, while **O** and **EO** describe the total number of occurrences and the expected total number of occurrences. The score **SlnSES** describes a statistical coverage of the sequences analyzed in the set and is based on a Markov Chain Background Model. Each set of attributes was computed for the masked as well as the unmasked version of the corresponding segment with the emphasis placed on the unmasked version (i.e. sorting of the table based on the unmasked **SlnSES** score). Further information for the word is provided through its reverse complement (RevComp) and the position of the reverse complement in the set of results (RC_Pos) as well as a notion describing if the word is a genomic palindrome (Pal). Finally, **PValues** describes a p-value that is assigned in order to provide statistical insight allowing the determination if a word is relevant or was discovered as interesting by random chance.
Table 6: The top 25 words in Proximal Promoters

| Word       | S    | ES    | O    | EO    | SlnSES | S    | ES    | EO    | SlnSES | RevComp | RC_Pos | Pal   | PValues  |
|------------|------|-------|------|-------|--------|------|-------|-------|--------|---------|--------|-------|----------|
| TAAAAAT    | 4249 | 3411.11 | 4837 | 3674.74 | 933.272 | 3681 | 3028.65 | 4071 | 3237.18 | 718.039 | ATTTTTTA | 1 No    | 0      |
| ATTTTTTA   | 3876 | 3135.31 | 4372 | 3358.5 | 822.01 | 3313 | 2758.58 | 3636 | 2932.38 | 606.738 | TAAAAAT | 0 No    | 2.22E-16|
| TTATATAA   | 3094 | 2505.92 | 3390 | 2650.31 | 652.239 | 2934 | 2508.38 | 2712 | 2632.02 | 211.674 | TTATATAA | 2 Yes   | 7.77E-16|
| GAAAAAAG   | 2066 | 1652.5 | 2182 | 1718.49 | 461.395 | 1956 | 1621.19 | 2053 | 1684.9 | 367.226 | CTTTTTTC | 5 No    | 1.11E-16|
| CTTTTTTC   | 1960 | 1578.31 | 2072 | 1638.97 | 424.512 | 1869 | 1559.58 | 1969 | 1618.92 | 338.269 | GAAAAAAG | 4 No    | 1.11E-16|
| AAAATTTG   | 2975 | 2595.17 | 3208 | 2749.61 | 406.363 | 2938 | 2681.48 | 2938 | 2881.32 | 173.09 | AAAATTTT | 9 No    | -6.66E-16|
| TAAAATTT   | 4339 | 3951.48 | 5058 | 4305.15 | 405.93 | 3764 | 3489.93 | 4214 | 3603.07 | 439.888 | AAATTTTA | 10 No   | -6.66E-16|
| TAATTTTT   | 4656 | 4272.02 | 5336 | 4686.12 | 400.73 | 4125 | 3726.41 | 4609 | 4040.78 | 419.188 | AATATATT | 19 No   | 0      |
| CAATTTTT   | 2872 | 2499.79 | 3110 | 2643.5 | 398.63 | 2633 | 2389.32 | 2829 | 2870.98 | 390.785 | AAAATTTT | 7 No    | 6.66E-16|
| AAAAAATTG  | 2374 | 2075.6 | 2517 | 2176.47 | 318.891 | 2230 | 1927.32 | 2354 | 2015.09 | 325.288 | ATTTTTCA | 21 No   | 5.64E-13|
| AAAATTTT   | 3874 | 3607.85 | 4265 | 3902.57 | 275.738 | 3494 | 3280.06 | 3823 | 3524 | 220.77 | GATTTTTT | 68 No   | 5.63E-09|
| CATTCCCC    | 1675 | 1426.93 | 1760 | 1477.44 | 1553.49 | 1856 | 1612.42 | 234.336 | 254.025 | TAAGAAAT | 23 No   | 2.52E-11|
| TAAGAATT    | 1895 | 1645.36 | 1990 | 1710.83 | 267.683 | 1773 | 1553.49 | 1856 | 1612.42 | 234.336 | ATTTTTCA | 12 No   | 1.04E-10|
| GAAAAAAA    | 2679 | 2426.86 | 2853 | 2562.63 | 264.801 | 2506 | 2338.07 | 2643 | 2354.4 | 283.363 | TTTTTTC | 98 No   | 9.20E-09|
| AAAATTTA   | 4735 | 4477.84 | 5547 | 4933.58 | 264.404 | 4109 | 3862.67 | 4667 | 4200.51 | 254.025 | TATTTTTT | 8 No    | 1.33E-15|
| CAAATTTT   | 3347 | 3092.9 | 3655 | 3310.2 | 264.267 | 3054 | 2796.42 | 3304 | 2974.88 | 269.093 | AAAATTTT | 60 No   | 1.95E-09|
| ATTTTTC    | 2338 | 2088.5 | 2489 | 2190.56 | 263.846 | 2169 | 1928.62 | 2295 | 2065.6 | 254.769 | TGAAAAAT | 13 No   | 2.29E-10|
| TTTTTTGG    | 3369 | 3120.79 | 3724 | 3341.96 | 257.829 | 3050 | 2802.67 | 3330 | 2981.91 | 257.935 | TTTTTTGG | 28 No   | 4.49E-11|
| ATTTCTTA    | 1947 | 1705.79 | 2052 | 1775.75 | 257.518 | 1800 | 1598.57 | 1900 | 1660.66 | 213.623 | TAAGAAAT | 16 No   | 8.37E-11|

**Top 25 overrepresented words for the proximal promoters in Arabidopsis thaliana.** The Word attribute describes the short nucleotide sequence associated with a putative word. S and ES describe the number of sequences a word occurs in and the number of sequences the word was expected to occur in respectively, while O and EO describe the total number of occurrences and the expected total number of occurrences. The score SlnSES describes a statistical coverage of the sequences analyzed in the set and is based on a Markov Chain Background Model. Each set of attributes was computed for the masked as well as the unmasked version of the corresponding segment with the emphasis placed on the unmasked SlnSES score. Further information for the word is provided through its reverse complement (RevComp) and the position of the reverse complement in the set of results (RC_Pos) as well as a notion describing if the word is a genomic palindrome (Pal). Finally, PValues describes a p-value that is assigned in order to provide statistical insight allowing the determination if a word is relevant or was discovered as interesting by random chance.
The large differences between the various sets of words provide evidence for the existence of segment-specific signatures. Of additional interest is the uniqueness of the word-based genomic signatures in comparison to the signature for the entire *Arabidopsis* genome. Clearly, the segments’ signatures distinguish them from each other and from the entire genome.

In addition to uniquely characterizing each segment, the top words discovered in each data set also have a strong

| Word            | Unmasked | Masked | Unmasked |
|-----------------|----------|--------|----------|
| ATTTTTTA        | 5789     | 4874.02| 7202     |
| TAAAAAAT        | 5865     | 4983.57| 7314     |
| GAAAAAG         | 3578     | 2825.77| 3921     |
| CTTTTTTTC       | 3546     | 2878.92| 3904     |
| TTATATAA        | 4781     | 4107.17| 5656     |
| AATATATT        | 5432     | 4895.21| 6702     |
| CAAGAAAC        | 2910     | 2459.44| 3187     |
| GTTTCTTG        | 2912     | 2482.93| 3182     |
| GAAAAATG        | 3158     | 2736.51| 3416     |
| GTTTTTGAG       | 3516     | 3093.27| 3830     |
| GAAAAC          | 3013     | 2605.34| 3240     |
| CAATTATT        | 4457     | 4041.77| 4991     |
| ATTTTGTA        | 4098     | 3689.96| 4626     |
| TCAAAAAC        | 3414     | 3011.29| 3688     |
| GAAGAAAG        | 3851     | 3448.55| 4291     |
| GTTATG          | 2173     | 1793.07| 2293     |
| CTATTTC         | 1618     | 1250.45| 1676     |
| GTTTAAG         | 1957     | 1584.64| 2054     |
| ATTTTCA         | 4081     | 3695.36| 4496     |
| TAAGAAGT        | 1465     | 1112.41| 1517     |
| CTTGTTTC        | 2351     | 1980.52| 2504     |
| CAAAAG          | 3391     | 3011.99| 3696     |
| TAGAAAT         | 3556     | 3178.38| 3887     |
| ATTTCTCA        | 2716     | 2348.17| 2896     |

Top 25 overrepresented words for the distal promoters in *Arabidopsis thaliana*. The *Word* attribute describes the short nucleotide sequence associated with a putative word. *S* and *ES* describe the number of sequences a word occurs in and the number of sequences the word was expected to occur in respectively, while *O* and *EO* describe the total number of occurrences and the expected total number of occurrences. The score *SnSES* describes a statistical coverage of the sequences analyzed in the set and is based on a Markov Chain Background Model. Each set of attributes was computed for the masked as well as the unmasked version of the corresponding segment with the emphasis placed on the unmasked version (i.e. sorting of the table based on the unmasked SnSES score). Further information for the word is provided through its reverse complement (*RevComp*) and the position of the reverse complement in the set of results (*RC_Pos*) as well as a notion describing if the word is a genomic palindrome (*Pal*). Finally, *PValues* describes a *p*-value that is assigned in order to provide statistical insight allowing the determination if a word is relevant or was discovered as interesting by random chance.

The large differences between the various sets of words provide evidence for the existence of segment-specific signatures. Of additional interest is the uniqueness of the word-based genomic signatures in comparison to the signature for the entire *Arabidopsis* genome. Clearly, the segments’ signatures distinguish them from each other and from the entire genome.

In addition to uniquely characterizing each segment, the top words discovered in each data set also have a strong
Table 8: The top 25 words in the entire genome

| Word       | Unmasked | Masked | Unmasked |
|------------|----------|--------|----------|
|            | S  ES O  OlnOEO S  ES O  OlnOEO RevComp RC_Pos Pal PValues |
|AAAAAAAA    | 5 5 128631 119310 9675.67 5 5 101229 95334 6073.66 TTTTTTTT 1 No 0 |
|TTTTTTTT    | 5 5 126533 117302 9585.11 5 5 98883 93091.2 5968.36 AAAAAAAA 0 No 1.67E-15 |
|TATATAT     | 5 5 58215 49385.7 9575.32 5 5 29264 27159.9 2183.54 TATATATA 2 Yes 3.89E-15 |
|ATATATAT    | 5 5 59429 53453 6298.28 5 5 30192 29596.8 601.111 ATATATAT 3 Yes 3.00E-15 |
|AAAAAAT     | 5 5 14823 11276.3 4053.8 5 5 11492 9148.23 2621.21 ATTTTTTTA 5 No 4.44E-16 |
|TTTTTATA    | 5 5 14743 11385.1 3810.52 5 5 30192 29596.8 601.111 ATATATAT 3 Yes 3.89E-15 |
|TTTTTTTA    | 5 5 14743 11385.1 3810.52 5 5 30192 29596.8 601.111 ATATATAT 3 Yes 3.89E-15 |
|GAAGAAGA    | 5 5 30102 26908.7 3375.68 5 5 22784 20523.6 2380.33 TCTTCTTTT 7 No 0 |
|TCTTCTTC    | 5 5 30267 27090.3 3356.11 5 5 23044 20902.7 2247.42 GAAGAAGA 6 No 0 |
|TTTTAAAA    | 5 5 29354 26314.9 3208.24 5 5 19409 17519.9 1987.46 TTTTTTTT 8 Yes 2.55E-15 |
|AAATATATT   | 5 5 14170 11335.3 3140.06 5 5 11168 10179.5 1035.06 AATATATT 9 Yes 1.11E-16 |
|TTTTCTTT    | 5 5 31066 28174.8 3034.69 5 5 26876 24223.6 2571.58 AAAGAAAAA 11 No 0 |
|AAAGAAAAA   | 5 5 31033 28187.3 2984.8 5 5 26861 24520.1 2469 TTTTTTTET 10 No 1.11E-16 |
|AGAGAGAG    | 5 5 19376 16630.5 2960.63 5 5 12615 11397.8 1280.05 CTCTCTCT 16 No 1.11E-16 |
|TCTCTCTCT   | 5 5 19179 16519.7 2862.73 5 5 12912 11634.1 1345.64 GAGAGAGA 14 No 4.44E-16 |
|GAGAGAGAG   | 5 5 20064 17413.4 2842.81 5 5 13136 11970.7 1220.21 TCTCTCTC 13 No 1.89E-15 |
|AAGAGAGA    | 5 5 32397 29731.9 2781.12 5 5 24352 23296.2 1079.35 TCTTTCTT 19 No 0 |
|CTCTCTCT    | 5 5 18513 15956.1 2751.61 5 5 12312 11212.7 1151.45 AGAGAGAG 12 No 1.11E-16 |
|AGAGAGAG    | 5 5 26477 24049.7 2545.91 5 5 19161 18013.6 1183.17 CTCTCTCT 20 No 8.88E-16 |
|TTATATAA    | 5 5 11402 9138.11 2523.66 5 5 9262 8518.12 775.46 TTATATAA 18 Yes 1.11E-15 |
|TTCTCTCT    | 5 5 32333 29910 2518.58 5 5 24550 23579.9 989.811 AAGAGAGA 15 No 0 |
|CTCTCTCT    | 5 5 26463 24183.9 2383.23 5 5 19432 18332.3 1132.03 AAGAGAGA 17 No 0 |
|TTTTTTTT    | 5 5 30561 28331 2315.57 5 5 26516 24717.1 1862.84 AAGAAAAA 22 No 0 |
|AAAGAAAAA   | 5 5 30461 28234.7 2311.9 5 5 26488 24756.8 1790.32 TTTTTTTT 21 No 4.44E-16 |
|TTTTTTTT    | 5 5 32141 29931 2289.6 5 5 27813 26102.2 1765.71 AAAAAAA 36 No 8.88E-16 |

Top 25 overrepresented words for the entire genome of Arabidopsis thaliana. The Word attribute describes the short nucleotide sequence associated with a putative word. S and ES describe the number of chromosomes a word occurs in and the number of chromosomes the word was expected to occur in respectively, while O and EO describe the total number of occurrences and the expected total number of occurrences. The score OlnOEO describes a statistical overrepresentation of the word in the genome and is based on a Markov Chain Background Model. Each set of attributes was computed for the masked as well as the unmasked version of the corresponding segment with the emphasis placed on the unmasked (i.e. sorting of the table based on the unmasked OlnOEO score).

Further information for the word is provided through its reverse complement (RevComp) and the position of the reverse complement in the set of results (RC_Pos) as well as a notion describing if the word is a genomic palindrome (Pal).

Finally, PValues describes a p-value that is assigned in order to provide statistical insight allowing the determination if a word is relevant or was discovered as interesting by random chance.
probability of being functional regulatory elements. This argument was strengthened by a functional analysis, which is described later in this section.

**Missing Words**

Another interesting component of our word-based signature is the set of words NOT contained within the different segments (see Tables 9, 10, 11, &12 and Additional files 8, 9, 10, &11), referred to as **unwords** [43] or **nullomers** [44,45]. The absence of words in particular segments is an interesting phenomenon and may represent negative selection pressure or increased mutation rates specific to these words, or structural constraints on DNA [44]. Thus, the missing word sets, which show unwords and their associated scores, serve as important 'fingerprints' for the segments.

Table 9: Words not detected in the 3'UTRs

| #WORD     | E_S   | E    |
|-----------|-------|------|
| CTAGCAGG  | 5.98269 | 6.17391 |
| ACTGCCAG  | 4.99319 | 5.1526  |
| CGCCTGAT  | 4.97776 | 5.13667 |
| GCGTCCGA  | 4.52742 | 4.67187 |
| GGGTGCGC  | 4.5248  | 4.66917 |
| ACTCCGCC  | 4.38831 | 4.5283  |
| CCGTTCGC  | 4.25101 | 4.3866  |
| ACACGCGC  | 4.21714 | 4.35165 |
| CCGCTCA   | 4.193   | 4.32673 |
| CTGGGCGT  | 4.06873 | 4.19847 |
| GACCTCGC  | 3.71851 | 3.83704 |
| GGCACGTA  | 3.68699 | 3.80451 |
| GCACCCCA  | 3.6084  | 3.7234  |
| GCACCTTC  | 3.59671 | 3.71134 |
| CGCAGCCA  | 3.54333 | 3.65625 |
| CCGGGTCG  | 3.53385 | 3.64646 |
| GGGTCGCC  | 3.52066 | 3.63636 |
| GACGGCCT  | 3.35465 | 3.46154 |
| GCGCAGCC  | 3.31181 | 3.41732 |
| CGTCCGCT  | 3.28252 | 3.3871  |
| CTGGGCGC  | 3.2624  | 3.36634 |
| GCGGACCT  | 3.25626 | 3.36 |
| ATACGCC   | 3.18816 | 3.28972 |
| AGCGCTCC  | 2.98494 | 3.08    |
| TAGCGCGG  | 2.98494 | 3.08    |

Top 25 words that were expected to occur in the 3'UTR but are not part of the sequences. Each word is identified through its nucleotide sequence and contains information about the expected number of sequences it was computed to occur in (E_S) as well as the expected number of total occurrences in the set of sequences (E). The words are sorted by their expected sequence occurrence.

Table 10: Words not detected in the 5'UTRs

| #WORD     | E_S   | E    |
|-----------|-------|------|
| GGAACGTC  | 5.1333 | 5.40909 |
| GAGGACCC  | 5.02658| 5.29661 |
| GCCCTATA  | 5.015  | 5.2844 |
| CCGTACCT  | 4.98236| 5.25   |
| GCGAGTAT  | 4.94491| 5.21053 |
| TATCGCAC  | 4.83088| 5.09034 |
| GTGGTGGG  | 4.69443| 4.94652 |
| GCGGAGTG  | 4.66421| 4.91468 |
| AGTACAGC  | 4.51745| 4.76   |
| GTGCGATG  | 4.4368 | 4.675  |
| GTCCGAGG  | 4.41572| 4.65278 |
| CCGCCGTTG | 4.3768 | 4.61176 |
| GTGCCGGGG | 4.16843| 4.39216 |
| GTGCTGGG  | 4.13122| 4.35294 |
| TAGTGCAC  | 4.12843| 4.35   |
| TACCGCCC  | 4.08277| 4.30189 |
| GCCTACGC  | 4.03144| 4.24779 |
| CACGGGGG  | 3.94494| 4.15663 |
| GCAGGCTG  | 3.90217| 4.11155 |
| CCGCTTAG  | 3.77819| 3.98089 |
| CAGCCAGC  | 3.74709| 3.94811 |
| TGAACGGG  | 3.74703| 3.94805 |
| CGTACTGC  | 3.74638| 3.94737 |
| GTGGCCGGG | 3.68013| 3.87755 |
| AGTCCTGG  | 3.67692| 3.87417 |

Top 25 words that were expected to occur in the 5'UTR but are not part of the sequences. Each word is identified through its nucleotide sequence and contains information about the expected number of sequences it was computed to occur in (E_S) as well as the expected number of total occurrences in the set of sequences (E). The words are sorted by their expected sequence occurrence.
Any biologically required sequence experiences evolutionary pressure (in this case purifying selection) resulting in a narrowing of the range of allowable sequence mutations. Often, a word and various mutations of the word exhibit the same functionality. To incorporate this into our analysis, clusters were built around each of the top overrepresented words, forming groups of words that are similar to each ‘seed word.’ Word similarity was measured through the Hamming distance metric, which models single point mutations. A Hamming distance of 1 was used to form the clusters. Each cluster is depicted via a sequence logo, providing a visual motif of the characteristics of the cluster.

### Table 11: Words not detected in the Introns

| #WORD            | E_S   | E   |
|------------------|-------|-----|
| CGCCGACA         | 6.1805| 6.4557 |
| CCCGGGAG         | 4.57278| 4.77632 |
| CCGGCCCC         | 4.46781| 4.66667 |
| CGCCCCCC         | 4.45254| 4.65072 |
| GCCCACCG         | 4.16782| 4.35331 |
| GCCGCAGGG        | 3.47686| 3.63158 |
| CGAGGGGG         | 3.34433| 3.49315 |
| AAGCCGCC         | 3.17737| 3.31875 |
| CGCCAGCG         | 2.99188| 3.125 |
| CGCTCGCC         | 2.91507| 3.04478 |
| GCGTCGCC         | 2.8245| 2.95017 |
| CCGGACGC         | 2.48216| 2.59259 |
| CGGGGGGC         | 2.25483| 2.35514 |
| CCGCCGCC         | 2.16189| 2.25806 |
| TCGGGGCC         | 2.11021| 2.20408 |
| GGGCAAGA         | 2.02051| 2.11039 |
| CGCTCCGC         | 2.00514| 2.09434 |
| CGCCACGC         | 1.99945| 2.0884 |
| TGCCGCAG         | 1.9539| 2.04082 |
| GGTGGCGG         | 1.92911| 2.01493 |
| CGGGGCC         | 1.90464| 1.98936 |
| CGGGCGGA         | 1.86163| 1.94444 |
| GCGCGAGG         | 1.83299| 1.91453 |
| GGCGCGGC         | 1.79662| 1.87654 |
| CGGCCGGG         | 1.73887| 1.81622 |

Top 25 words that were expected to occur in the introns but are not part of the sequences. Each word is identified through its nucleotide sequence and contains information about the expected number of sequences it was computed to occur in (E_S) as well as the expected number of total occurrences in the set of sequences (E). The words are sorted by their expected sequence occurrence.

### Table 12: Words not detected in the Core Promoters

| #WORD            | E_S   | E   |
|------------------|-------|-----|
| CGCACCC         | 5.86109| 6.3029 |
| GTCCGAAC         | 5.46787| 5.88 |
| GCCCTATG         | 5.23895| 5.6338 |
| GGACGTCC         | 4.98873| 5.36471 |
| GCCCTAG          | 4.47129| 4.80822 |
| CGCGAGCG         | 4.35999| 4.68852 |
| GATCCCCC         | 3.92081| 4.21622 |
| GGCGCAT          | 3.82028| 4.10811 |
| TACCCAGG         | 3.80429| 4.09091 |
| GGCCCATG         | 3.67267| 3.94937 |
| CGCATCCG         | 3.66922| 3.94565 |
| CACGCGCA         | 3.56933| 3.83824 |
| CGGGCCGC         | 3.51312| 3.77778 |
| CGCGGTCG         | 3.51079| 3.77528 |
| AGGCGCTC         | 3.50922| 3.77358 |
| GGGCGCGTC         | 3.49296| 3.7561 |
| AGGCCCTG         | 3.45587| 3.71622 |
| CGCGACGC         | 3.30648| 3.55556 |
| AGTGGGCG         | 3.29952| 3.54808 |
| GGCGGTTC         | 3.26995| 3.51628 |
| CGCGCAAG         | 3.25481| 3.5 |
| ACCCGCGT         | 3.22635| 3.46939 |
| TTACCCCG         | 3.22482| 3.46774 |
| CCGGTCGCC         | 3.18249| 3.42222 |
| TAGGCGCC         | 3.18249| 3.42222 |

Top 25 words that were expected to occur in the core promoters but are not part of the sequences. Each word is identified through its nucleotide sequence and contains information about the expected number of sequences it was computed to occur in (E_S) as well as the expected number of total occurrences in the set of sequences (E). The words are sorted by their expected sequence occurrence.
Selected clusters and the corresponding sequence logos are shown in Additional file 12. Two representative motifs are presented for each segment. Motifs for each segment were chosen in order to provide a variety of examples of putative binding sites for the non-coding segments.

The presented motifs correspond to well-known regulatory elements and complex motifs, which represent sets of putative regulatory elements. Of particular interest in Additional file 12 are the word-based clusters for the core promoters (in the left column) which correspond to the TATA-box. Also known as the Goldberg-Hogness box [46], the TATA-box is a well-characterized regulatory element appearing 31 bp upstream of the transcription start site in 30% of the promoter sequences in *Arabidopsis* [23]. The core promoters also contain another interesting motif, (CGACGTCG), which is involved in stress response in *Arabidopsis thaliana* [22]. An extensive functional characterization is described later in this section.

**Word Location Distribution**

The locations of a particular word within a segment can provide insight into functional properties of the word. For example, functional TATA motifs are located at a specific distance upstream of the TSS [23,46]. We identified the segment-specific locations of the seed words of the clusters shown in Additional file 12. Being selected for their high complexity, these words are expected to exhibit a distribution bias, manifesting as peaks in the scatterplots of the distribution across sequences, as shown in Figures 1, 2, 3 and 4.

---

**Figure 1**

**Word location distribution across introns.** Word location distributions for interesting words within the introns. The occurrences are shown on a log-scale in order to allow a comparison between the different segments as well as the words visualized for the entire genome.
The Figures contain histograms showing the numbers of occurrences of specific words at each point along the sequences. For uniformity, sequence lengths are normalized to the range [1;100]. Strong peaks can indeed be found for the words selected in the intron, core promoter, and proximal promoter regions. The peaks detected for the intron segment are at both the 5' and 3' ends of the introns, which means that the words occur in close proximity to flanking exons. The close proximity to the intron-exon boundaries is expected for splicing regulatory sequences [2,8-16]. The peaks exhibited in core and proximal promoters are not surprising. The distributions of words locations in these segments are expected to show clustering, due to positional conservation of locations of cis-regulatory elements [23]. Of particular interest is the location of the peak for the first word chosen for the core promoter distribution (TATAATA), the TATA-box. A location of around 31 bp upstream from the TSS corresponds to the findings in [23].

Interestingly, we also detect strong peaks for the example words chosen for the genome wide word landscape, possibly indicating an important chromosomal feature that is not yet understood.

**Word Co-occurrences**

Genes are usually controlled by a combination of multiple transcription factors, or by transcription factor complexes binding to different sites embedded in the genes' regulatory non-coding regions. In order to detect the interacting transcription factor binding sites of a complex, we examined the positional relationships of
words. The top 25 overrepresented words were paired, and the overrepresentation of each pair was determined using a Markovian background model of order 6. The top 25 overrepresented word pairs for each segment are displayed in Tables 13, 14, 15, 16, 17 and 18 (see also Additional files 13, 14, 15, 16, 17, &18). The limited overlap between the word pairs of different segments indicates additional unique word-based signatures for genomic segments.

**AGRIS Lookup**

The AGRIS database [25] contains a large collection of known regulatory elements for *Arabidopsis thaliana*. The words discovered in this study were compared to the regulatory elements of equal or lesser length in AGRIS. Table 19 provides the overview of the motifs and their locations within the results.

**Functional Categorizations of Words**

In order to reveal biological meanings of overrepresented words, we established associations between the overrepresented words and biological functions of the genes that harbour a particular word in their corresponding segment (Table 1). For a single word, all the genes that contain that word in their selected segment were found and the corresponding overrepresented Gene Ontology (GO) terms were identified. Overrepresentation of a GO term is determined by using the *Arabidopsis* gene GO term distributions as a background model. The developmental and experimental parameters that perturb the

---

**Figure 3**

*Word location distribution across proximal promoters*. Word location distributions for interesting words within proximal promoters. The occurrences are shown on a log-scale in order to allow a comparison between the different segments as well as the words visualized for the entire genome.
expression of genes harbouring a particular word was determined by comparing the number of induced, suppressed or neutral genes, to that expected by chance in a collection of 1305 tissue and stress microarrays from the public domain. Significant enrichment or depletion of induced or suppressed genes has been shown to correlate strongly with factors affecting regulation of a cis-regulatory element [39].

As shown in Figures 5, 6, 7, 8, 9 and 10, we identified overrepresented functional categories (y-axis) of genes that carry a particular word (x-axis, top panel) in either their 3’UTR (Figure 5), 5’UTR (Figure 6), intron (Figure 7), or promoter regions (Core, Proximal and Distal Promoters, Figures 8, 9 and 10, respectively). The red squares depict overrepresented categories with lowest p-value, calculated for each segment separately, smaller than 10E-16. For example, the word GTTTTTGA was significantly enriched in the 3’UTRs of genes that participate in the GO category “Protein Synthesis” (including the sub-categories ribosome biogenesis, ribosomal proteins, translation), and is correlated with genes suppressed in flowers and early stage siliques (p-value 4E-14). Based on microarray expression of micro-dissected tissues (see methods), the word TGTTTTTT is present in the 3’UTR of genes induced in roots (p-value 1E-8), in the atrichoblast (hairless) cell files of the root (p-value 7E-25), the root cortex (p-value 2E-23), endodermis (p-value 2E-51), and lateral root cap (p-value 4E-20). The word CTCTCTTT, enriched in introns, was correlated with differential induction in cotyledons (p-value 8E-20), suppressed in young flowers, especially

Figure 4
Word location distribution across the entire genome. Word location distributions for interesting words within the genome. The occurrences are shown on a log-scale in order to allow a comparison between the different segments as well as the words visualized for the entire genome.
carpals (p-value 1E-14) and heart stage embryos (p-value 3E-20). Surprisingly, the presence of these words in the UTRs and introns were strongly correlated with tissue specific profiles, but were only weakly enriched or strongly depleted for responses by hormones, biotic and abiotic stresses. There was no significant correlation to any of the 1305 surveyed conditions if the words were located in the 1000 bp upstream or downstream regions. This is strikingly different to the well characterized abscisic acid responsive element (ABRE) (CACGTGTC) \[22\], which when found in the 1000 bp 5' upstream region, was strongly correlated to induction by 10 μM abscisic acid (ABA) (p-value 4E-49), cold, salt and drought stresses (p-values < 1E-40), in flowers (p-value

### Table 13: Co-occurrence in 3'UTRs

| Word1   | Word2   | S   | E_s | S*ln(S/E_s) |
|---------|---------|-----|-----|-------------|
| TTCTTTTT | TTCTTTTT | 322 | 238.5802 | 96.5504     |
| TGTTTTTT | TTCTTTTT | 283 | 217.7183 | 74.2154     |
| TTTTCTTT | TTTTCTCT | 260 |
| TTTTCTTT | TTCTTTCT | 326 |
| TTTTCTTT | TTCTTTTT | 270 |
| TTCTTTTT | TTTTTTCC | 278 |
| TTTTTGTT | TTCTTTTT | 161 |
| TTATTTTT | TTCTTTTT | 211 |
| TTCTTTTT | TTCTTTTT | 290 |
| TGTTTTTT | TTCTTTTT | 239 |
| TTTTCTTT | TTTTTTCT | 217 |
| TTCTTTTT | TTCTTTTT | 112 |
| TGTTTTTT | TTCTTTTT | 129 |
| TTTTTGTT | TTCTTTTT | 148 |
| TTTTTCTT | TTCTTTTT | 128 |
| TTTTTTTT | TGTTTTTT | 265 |
| TTTTTTTT | TTCTTTTT | 170 |
| TTTTTTTT | TTCTTTTT | 136 |
| TTTTTTTT | TTCTTTTT | 127 |
| TTTTTTTT | TTCTTTTT | 285 |
| TTTTTTTT | TTCTTTTT | 137 |
| TTTTTTTT | TTCTTTTT | 127 |
| TTTTTTTT | TTCTTTTT | 216 |
| TTTTTTTT | TTCTTTTT | 161 |
| ATTTTTTA | TTTTTTTA | 82 |

Overrepresented non-overlapping word-pairs detected in the 3'Untranslated Regions of Arabidopsis thaliana. A word-pair is characterized through the two nucleotide sequences associated with it (Word1 and Word2), the number of sequences the pair occurs in (S) as well as the expected number of sequences (E_s) and a statistical score symbolizing the overrepresentation of the word-pair in the specific sequence set (S*ln(S/E_s)).

### Table 14: Co-occurrence in 5'UTRs

| Word1   | Word2   | S   | E_s | S*ln(S/E_s) |
|---------|---------|-----|-----|-------------|
| CTCTCTCT | CTTTCTCT | 209 |
| TTCTCTTC | CTCTTTCT | 214 |
| TTCTCTTC | CTCTTTCT | 198 |
| TTCTCTTC | CTCTTTCT | 97  |
| TTCTCTTC | TTCTTTTT | 97  |
| TTCTCTTC | TTCTTTTT | 122 |
| TTCTCTTC | TTCTTTTT | 290 |
| TTCTCTTC | TTCTTTTT | 239 |
| TTCTCTTC | TTCTTTTT | 112 |
| TTCTCTTC | TTCTTTTT | 129 |
| TTCTCTTC | TTCTTTTT | 148 |
| TTCTCTTC | TTCTTTTT | 128 |
| TTCTCTTC | TTCTTTTT | 265 |
| TTCTCTTC | TTCTTTTT | 170 |
| TTCTCTTC | TTCTTTTT | 136 |
| TTCTCTTC | TTCTTTTT | 127 |
| TTCTCTTC | TTCTTTTT | 285 |
| TTCTCTTC | TTCTTTTT | 137 |
| TTCTCTTC | TTCTTTTT | 127 |
| TTCTCTTC | TTCTTTTT | 216 |
| TTCTCTTC | TTCTTTTT | 161 |
| TTCTCTTC | TTCTTTTT | 82  |

Overrepresented non-overlapping word-pairs detected in the 5'Untranslated Regions of Arabidopsis thaliana. A word-pair is characterized through the two nucleotide sequences associated with it (Word1 and Word2), the number of sequences the pair occurs in (S) as well as the expected number of sequences (E_s) and a statistical score symbolizing the overrepresentation of the word-pair in the specific sequence set (S*ln(S/E_s)).

\[22\]
(1E-31), and suppressed in roots (p-value 4E-7) but no significant correlations were observed when ABRE was present in the 3'UTRs, 5'UTRs or introns. We also analyzed primary promoter regions where most of the basal promoter elements are expected to be located. The frequency of words is calculated as described above, and genes that contain the high scoring word in their primary promoter region were queried for enriched biological function. For example, GCCCATTA is found in core promoter regions of genes preferably involved in ribosome biogenesis and translation. Genes with this word in the upstream promoter are significantly depleted for response to all hormones, biotic and abiotic stresses (typically p-value 1E-8 or better). In other words, overrepresented non-overlapping word-pairs detected in the introns of Arabidopsis thaliana. A word-pair is characterized through the two nucleotide sequences associated with it (Word1 and Word2), the number of sequences the pair occurs in (S) as well as the expected number of sequences (ES) and a statistical score symbolizing the overrepresentation of the word-pair in the specific sequence set (S*ln(S/ES)).

Table 15: Co-occurrence in Introns

| Word1     | Word2     | S   | E_S | S*ln(S/ES) |
|------------|-----------|-----|-----|------------|
| TTTTATTT   | AATATTTA  | 393 | 217.8144 | 231.9354   |
| TTTTATTT   | AATATTTA  | 334 | 186.0726 | 195.3914   |
| TAAAAAAAT  | AATATTTA  | 147 | 39.3119  | 193.8792   |
| TTTTATTT   | TTTTATTT  | 460 | 306.2869 | 187.084    |
| TAAAAAAAT  | TTTTATTT  | 273 | 140.3538 | 181.6284   |
| TAATTTTT   | TTTTATTT  | 238 | 113.2939 | 176.6363   |
| CTCTGTTT   | CTGTGTTT  | 346 | 208.3136 | 175.5583   |
| TTTTATTT   | AATATTTA  | 308 | 175.8151 | 172.6854   |
| TTTTATTT   | TTTTATTT  | 505 | 358.7745 | 172.6415   |
| TAAAAAAAT  | AATATTTA  | 149 | 48.6332  | 166.8264   |
| TAAAAAAAT  | TTTTATTT  | 189 | 79.759   | 163.0573   |
| TAAAAAAAT  | TAATTTTT  | 179 | 73.1119  | 160.2756   |
| TTTTATTT   | TTTTATTT  | 461 | 328.5857 | 156.0948   |
| TTTTATTT   | TTTTATTT  | 238 | 123.6151 | 155.9133   |
| TAAAAAAAT  | TTTTCTTT  | 305 | 185.7949 | 151.1788   |
| TAAAAAAAT  | TTTTATTT  | 230 | 119.9486 | 149.7338   |
| TTTTATTT   | AATATTTA  | 261 | 150.2261 | 144.1709   |
| TATTTTTT   | TTTTATTT  | 300 | 186.1617 | 143.1501   |
| TTTTTTAT   | AATATTTA  | 202 | 99.8493  | 142.3033   |
| TTTTTTAT   | TTTTATTT  | 670 | 542.1648 | 141.8441   |
| TAAAAAAAT  | TTTTGTGT  | 262 | 157.163  | 133.898    |
| TAAATTTT   | TTTTGTGT  | 187 | 91.5206  | 133.6198   |
| ATTTTATA   | TTTTGTGT  | 354 | 243.9756 | 131.769    |
| TAAAAAAAT  | TTTTGTGT  | 357 | 246.9371 | 131.5909   |
| TTTTTATTT  | TTTTGTGT  | 638 | 519.9558 | 130.5312   |

Overrepresented non-overlapping word-pairs detected in the core promoters of Arabidopsis thaliana. A word-pair is characterized through the two nucleotide sequences associated with it (Word1 and Word2), the number of sequences the pair occurs in (S) as well as the expected number of sequences (ES) and a statistical score symbolizing the overrepresentation of the word-pair in the specific sequence set (S*ln(S/ES)).

Table 16: Co-occurrence in Core Promoters

| Word1     | Word2     | S   | E_S | S*ln(S/ES) |
|------------|-----------|-----|-----|------------|
| GCCCATTA   | GCCCATTA  | 32  | 2.3492 | 83.5729    |
| TTTTTCCT   | TTTTTCCT  | 68  | 22.9531 | 73.8516    |
| AAAAAAAA   | AAGAAAAAA | 84  | 41.5798 | 59.069     |
| CTCTTCTT   | CTCTTCTT  | 40  | 9.1626  | 58.95      |
| ATTTTTAA   | ATTTTTAA  | 57  | 22.4453 | 53.1222    |
| AACAAAAA   | AACAAAAA  | 71  | 35.1265 | 49.9645    |
| AACAAAAA   | AACAAAAA  | 66  | 31.075 | 49.6455    |
| ATTTTCAA   | TATATATA  | 30  | 6.1031  | 47.772     |
| AAAAAAA    | TAAAAAA   | 38  | 10.8748 | 47.5432    |
| AAAAAAAA   | AAAAAAAA  | 56  | 24.4921 | 46.3121    |
| AAAAAAT    | AAAAAACA  | 44  | 15.5191 | 45.8533    |
| AACAAAAA   | AACAAAAA  | 77  | 42.5433 | 45.6828    |
| AACAAAAA   | AACAAAAA  | 69  | 37.6758 | 41.7512    |
| TTTTCTTT   | TTTTCTTT  | 40  | 14.2927 | 41.1653    |
| AAAAAAAA   | ATATATA   | 30  | 7.659   | 40.9596    |
| AAAAAAAA   | CTATATA   | 36  | 11.9538 | 39.689     |
| AAAAAAT    | CTATATA   | 30  | 8.0863  | 39.3309    |
| TATATATA   | TATATATA  | 36  | 12.3623 | 38.4793    |
| AAAAAAAA   | TATATATA  | 53  | 25.8324 | 38.0892    |
| TTTTTTTT   | TTTTTTTT  | 38  | 14.0039 | 37.9336    |
| TTTTCTTT   | TTTTCTTT  | 50  | 23.5743 | 37.5932    |
| TTTTCTTT   | TTTTCTTT  | 46  | 20.3942 | 37.416     |
| AAATTTAA   | AAATTTAA  | 44  | 18.9721 | 37.0137    |
| AACAAAAA   | AACAAAAA  | 65  | 36.8225 | 36.938     |
| AACAAAAA   | AACAAAAA  | 41  | 16.8429 | 36.4755    |

Overrepresented non-overlapping word-pairs detected in the core promoters of Arabidopsis thaliana. A word-pair is characterized through the two nucleotide sequences associated with it (Word1 and Word2), the number of sequences the pair occurs in (S) as well as the expected number of sequences (ES) and a statistical score symbolizing the overrepresentation of the word-pair in the specific sequence set (S*ln(S/ES)).
genes harbouring this word in their upstream promoter region tend to be less responsive to stresses than randomly chosen genes. However, the word CTATAAAT was found in core promoter regions of genes preferably functioning as storage facilitating proteins (Figure 8). Genes with this word in the upstream promoter are rapidly induced by 10 nM brassinolide (p-value 1E-9) and by salt stress in roots (p-value 4E-9). These genes are also induced in roots, flowers, pollen, and during seed development, and strongly suppressed in young leaves and cotyledons.

A set of 10 frequently enriched cis-elements was recently provided for the ATH95 gene coexpression.

| Word1  | Word2  | S        | E_s      | S^*ln(S/ES) |
|--------|--------|----------|----------|-------------|
| AAATTTTA | TAAAAAAT | 996      | 489.8445 | 706.8206    |
| ATTTTTTA | TAAAAAAT | 869      | 395.77   | 683.4771    |
| TAAATTTT | TAAAAAAT | 970      | 501.8706 | 639.1852    |
| AAAATTTA | TAAAAAAT | 1040     | 565.2386 | 634.1171    |
| TAAATTTT | TAAAAAAT | 963      | 498.7952 | 633.5171    |
| TAAATTTT | TAAAAAAT | 892      | 458.4645 | 593.7003    |
| AAAATTTA | ATTTTTTA | 868      | 450.2375 | 569.7695    |
| AAAATTTA | ATTTTTTA | 947      | 519.5356 | 568.5445    |
| AAAATTTA | TAAAAAAT | 919      | 496.1801 | 566.4231    |
| TAATTTTT | TAAAAAAT | 965      | 539.2575 | 561.5671    |
| AAAATTTA | ATTTTTTA | 865      | 456.0608 | 553.6894    |
| TAATTTTT | ATTTTTTA | 907      | 495.6552 | 548.0656    |
| AATATATT | TAAAAAAT | 776      | 391.8276 | 530.2646    |
| AAAATTTA | AAATTTTA | 973      | 564.4665 | 529.8015    |
| AAAATTTA | TAAAAAAT | 976      | 567.4415 | 529.3092    |
| AAAAATTTA | TAAAAAAT | 1125     | 707.8947 | 521.1483    |
| AATATATT | TAAAAAAT | 730      | 360.1459 | 515.7708    |
| TAATTTTT | ATTTTTTA | 845      | 461.2912 | 511.4845    |
| AAAATTTA | TAAAAAAT | 1052     | 654.7789 | 498.8066    |
| AAAATTTA | AAAATTA  | 1044     | 651.346  | 492.5318    |
| AAAATTTA | TAAAAAAT | 958      | 574.7807 | 489.4031    |
| AAAATTTA | TAAAAAAT | 993      | 613.4724 | 478.2242    |
| TAATTTTT | TAAAAAAT | 995      | 624.6821 | 463.1724    |
| AAAATTTA | TAAAAAAT | 990      | 621.407  | 461.0615    |
| TTATATAA | TAAAAAAT | 645      | 316.3233 | 459.5531    |

Overrepresented non-overlapping word-pairs detected in the proximal promoters of Arabidopsis thaliana. A word-pair is characterized through the two nucleotide sequences associated with it (Word1 and Word2), the number of sequences the pair occurs in (S) as well as the expected number of sequences (E_s) and a statistical score symbolizing the overrepresentation of the word-pair in the specific sequence set (S^*ln(S/E_s)).

| Word1  | Word2  | S        | E_s      | S^*ln(S/ES) |
|--------|--------|----------|----------|-------------|
| TAAAAAAT | ATTTTTTA | 1855     | 898.8038 | 1344.087    |
| AATATATT | TAAAAAAT | 1759     | 902.7094 | 1173.429    |
| AATATATT | ATTTTTTA | 1692     | 882.8679 | 1100.631    |
| TTATATAA | ATTTTTTA | 1478     | 740.7429 | 1020.99     |
| TTATATAA | TAAAAAAT | 1464     | 757.3903 | 964.8477    |
| AATATATT | TTATATAA | 1447     | 743.9616 | 962.6287    |
| AAAAATTTG | TAAAAAAT | 1301     | 747.7933 | 720.4442    |
| CAATTTTT | TAAAAAAT | 1279     | 745.3293 | 690.6698    |
| AAAAATTTG | ATTTTTTA | 1237     | 731.3568 | 650.0966    |
| ATTTTTGA | ATTTTTTA | 1156     | 665.4975 | 638.3272    |
| CAATTTTT | ATTTTTTA | 1200     | 728.947  | 598.171     |
| TAGAAAT | TAAAAAAT | 1024     | 586.114  | 571.3484    |
| ATTTTTGA | TAAAAAAT | 1108     | 680.4539 | 540.2074    |
| CAATTTTT | AATATATT | 1162     | 732.1145 | 536.7987    |
| ATTTTTCA | ATTTTTTA | 1078     | 666.4705 | 518.3745    |
| AAAAATTTG | AATATATT | 1148     | 734.5348 | 512.627     |
| CAATTTTT | TTATATAA | 1003     | 614.2579 | 491.8069    |
| TAGAAAT | AATATATT | 956      | 575.7221 | 484.8189    |
| ATTTTTCA | ATTTTTTA | 952      | 574.2477 | 481.2399    |
| ATTTTTCA | TAAAAAAT | 964      | 587.1534 | 477.9562    |
| TAGAAAT | ATTTTTTA | 941      | 573.2313 | 466.4103    |
| ATTTTTCA | TAAAAAAT | 1058     | 681.4487 | 465.4297    |
| TGAAATTT | ATTTTTTA | 1020     | 658.2655 | 446.7086    |
| TGAAATTT | TAAAAAAT | 1033     | 673.0593 | 442.5259    |
| AAAAATTTG | TTATATAA | 970      | 616.2886 | 439.9733    |

Overrepresented non-overlapping word-pairs detected in the distal promoters of Arabidopsis thaliana. A word-pair is characterized through the two nucleotide sequences associated with it (Word1 and Word2), the number of sequences the pair occurs in (S) as well as the expected number of sequences (E_s) and a statistical score symbolizing the overrepresentation of the word-pair in the specific sequence set (S^*ln(S/E_s)).
| Regulatory Element from AGRIS database \cite{25} | 3'UTRs | 5'UTRs | Intron Core Promoters | Proximal Promoters | Distal Promoters |
|-----------------------------------------------|--------|--------|----------------------|--------------------|-----------------|
| | Rank | Score | Rank | Score | Rank | Score | Rank | Score | Rank | Score |
| Bellringer/replumless/pennywise BS3 IN AG | - | - | - | - | 43503 | 0.0479784 | - | - | 64618 | 0.955909 | 56341 | -103.557 |
| CBF1 BS cor15a | - | - | - | - | 48346 | -1.48116 | - | - | 4852 | 1.34988 | 11624 | 24.1708 |
| Octamer promoter motif | - | - | - | - | 41435 | 0.673899 | - | - | 11935 | 1.28979 | 23858 | 4.69741 |
| Bellringer/replumless/pennywise BS1 IN AG | 72 | 67.6311 | 352 | 35.2087 | 574 | 127.468 | 19 | 117.144 | 875 | 1.0759 | 58337 | -186.12 |
| ABRE-like binding site motif | 5445 | 11.7462 | 1138 | 21.7556 | 15242 | 16.0488 | 304 | 41.9698 | 53 | 1.45099 | 109 | 255.929 |
| G-box promoter motif | 1852 | 21.1577 | 1138 | 21.7556 | 12023 | 16.0488 | 304 | 41.9698 | 53 | 1.45099 | 102 | 260.604 |
| DPBFI&2 binding site motif | 3720 | 14.7278 | 2963 | 13.7441 | 3460 | 54.8094 | 355 | 39.8827 | 137 | 1.36496 | 102 | 260.604 |
| MYB1 binding site motif | 4306 | 13.6223 | 446 | 13.7441 | 3460 | 54.8094 | 355 | 39.8827 | 137 | 1.36496 | 102 | 260.604 |
| RAV1-A binding site motif | 568 | 34.0603 | 148 | 49.0095 | 2000 | 73.6726 | 451 | 36.3111 | 135 | 1.0759 | 58337 | -186.12 |
| W-box promoter motif | 751 | 30.7769 | 675 | 27.0198 | 458 | 139.172 | 533 | 34.175 | 176 | 1.19182 | 756 | 131.24 |
| CBF2 binding site motif and GBF1/2/3 BS in ADH1 | - | - | - | - | 34949 | 2.87187 | 540 | 34.0562 | 729 | 1.293 | 998 | 117.554 |
| ARF and ARF1 binding site motif | 976 | 27.5809 | 216 | 42.5544 | 741 | 116.214 | 568 | 33.5619 | 2852 | 1.07934 | 2306 | 80.856 |
| L1-box promoter motif | 2697 | 17.6326 | - | - | 5824 | 38.2912 | 585 | 33.083 | 2889 | 1.05367 | 2235 | 81.9035 |
| GATA promoter motif | 1186 | 25.6353 | 741 | 26.1103 | 1247 | 91.6715 | 802 | 29.355 | 355 | 1.08161 | 1033 | 115.612 |
| ATB2/AtbZIP3/AtbZIP4/GBF5 BS in ProDH | 1757 | 21.6648 | 1225 | 20.9254 | 2890 | 60.5806 | 908 | 27.9139 | 1313 | 1.12688 | 3204 | 67.6808 |
| SORLIP2 | 3658 | 14.8663 | 9024 | 6.91197 | 16361 | 14.6754 | 1006 | 26.5383 | 550 | 1.34186 | 780 | 129.375 |
| MYB binding site promoter | 4762 | 12.8183 | 2462 | 15.1743 | 1897 | 75.734 | 1032 | 26.1692 | 4931 | 1.06605 | 2010 | 86.739 |
| CCA1 binding site motif | 1230 | 25.1325 | 371 | 34.5029 | 5202 | 41.532 | 1225 | 24.4536 | 61990 | 0.99765 | 58013 | -161.161 |
| TGA1 binding site motif | - | - | 13290 | 4.96662 | 12326 | 16.0488 | 304 | 41.9698 | 53 | 1.45099 | 102 | 260.604 |
| SORLIP1 | 5297 | 11.9625 | 6172 | 9.0064 | 11076 | 22.5348 | 1286 | 23.8899 | 4965 | 1.15533 | 4097 | 58.1886 |
| T-box promoter motif | 639 | 32.6567 | 1532 | 19.0267 | 774 | 114.265 | 1325 | 23.5609 | 193 | 1.27522 | 205 | 212.153 |
| Ibox promoter motif | 2156 | 19.649 | 358 | 35.0463 | 3223 | 57.1901 | 1797 | 20.4507 | 1081 | 1.14622 | 628 | 140.679 |
| Box II promoter motif | 1403 | 23.9863 | 4993 | 10.3195 | 1437 | 85.6577 | 1804 | 20.4254 | 1986 | 1.30314 | 669 | 136.891 |
| Hexamer promoter motif | 7590 | 9.4166 | 1616 | 18.5991 | 10347 | 24.0156 | 2225 | 18.6568 | 3477 | 1.24419 | 1252 | 107.567 |
| AtMYC2 BS in RD22 | 1193 | 25.5614 | 4026 | 11.6309 | 3460 | 54.8094 | 2823 | 16.6193 | 646 | 1.21499 | 2073 | 85.133 |
| RAV1-B binding site motif | 7054 | 9.94571 | 8250 | 7.4051 | 11589 | 21.6087 | 2996 | 16.0975 | 6084 | 1.12709 | 2017 | 86.5658 |
| RY-repeat promoter motif | 182 | 49.4382 | - | - | 330 | 132.253 | 3097 | 15.8378 | 72 | 1.29305 | 61 | 302.629 |
| MYB3 binding site motif | 5128 | 12.2348 | 10575 | 6.06616 | 1407 | 86.7638 | 3292 | 15.3953 | 3288 | 1.08324 | 11546 | 24.3649 |
| Bellringer/replumless/pennywise BS2 IN AG | 3126 | 16.2923 | - | - | 64424 | -30.4349 | 3694 | 14.5011 | 62777 | 0.97976 | 58184 | -172.62 |
Table 19: AGRIS Lookup (Continued)

| Element Description | Length | NCBI Accession | Length | Ensembl Accession | Length | GenBank Accession | Length | RefSeq Accession | Length |
|---------------------|--------|----------------|--------|-------------------|--------|-------------------|--------|------------------|--------|
| AtMYB2 BS in RD22   | -      | 6797           | 10.1949| 9630              | 6.55608| 4961              | 42.997 | 3570             | 1.07359 | 3218             | 67.5209 |
| E2F binding site motif and E2F/DP BS in AtCDC6 | -      | 4078           | 26.9446| 20822             | 10.4265| 6359              | 10.5016| 4340             | 1.35349 | 1735             | 93.0802 |
| ERFI BS in AtCBD-A and GCC-box promoter motif | -      | 681            | 8.95512| 16036             | 15.0374| 11248             | 7.01938| 11296            | 1.13624 | 7155             | 38.6247 |
| Z-box promoter motif | -      | 6230           | 8.95512| 16036             | 15.0374| 11248             | 7.01938| 11296            | 1.13624 | 7155             | 38.6247 |
| LTRE promoter motif | -      | 6230           | 8.95512| 16036             | 15.0374| 11248             | 7.01938| 11296            | 1.13624 | 7155             | 38.6247 |
| SORLIP4            | 5170   | 12.1706        | 3175   | 13.3137           | 14017  | 17.6817           | 11614  | 6.82909         | 14984  | 1.04471         | 22267  | 6.5221          |
| ABFs and ABRE binding site motif | 8540   | 8.6035         | 6266   | 8.92287          | 29109  | 5.33319          | 12250  | 6.52158         | 725    | 1.25598         | 1490   | 100.349         |
| PI promoter motif   | 9436   | 7.96403        | -      | -                | -      | -                | -      | -               | -      | -               | -      | -               |

Observations about the regulatory elements (length = 8) contained in the AGRIS database [25].

Gene Function Distribution of Genes Containing Top 25 Words - 3'UTR - log10(p-val)

Figure 5
Cellular functions in 3'UTRs. Enriched functional categories within the set of genes associated with each word in the top 25 words of the 3'UTRs. The lookup was conducted against the MIPS Functional Catalogue Database (FunCatDB) [54].
Figure 6

Cellular functions in 5'UTRs. Enriched functional categories within the set of genes associated with each word in the top 25 words of the 5'UTRs. The lookup was conducted against the MIPS Functional Catalogue Database (FunCatDB) [54].
Figure 7

Cellular functions in introns. Enriched functional categories within the set of genes associated with each word in the top 25 words of the introns. The lookup was conducted against the MIPS Functional Catalogue Database (FunCatDB) [54].
Figure 8

Cellular functions in core promoters. Enriched functional categories within the set of genes associated with each word in the top 25 words of the core promoters. The lookup was conducted against the MIPS Functional Catalogue Database (FunCatDB) [54].
Figure 9

**Cellular functions in proximal promoters.** Enriched functional categories within the set of genes associated with each word in the top 25 words of the proximal promoters. The lookup was conducted against the MIPS Functional Catalogue Database (FunCatDB) [54].

Figure 10

**Cellular functions in distal promoters.** Enriched functional categories within the set of genes associated with each word in the top 25 words of the distal promoters. The lookup was conducted against the MIPS Functional Catalogue Database (FunCatDB) [54].
neighbourhood (AAACCCTA, CITATCCN, GGCCCANN, GCCACGTN, GCCGGAAN, GACCGTNC, AANGTCAA, CNGATCNA, NGCTGTCC, CATGCANN) [47]. Our results show a direct overlap with two of those words (AAACCCTA, NGCTGTCC), which are detected and marked as ‘interesting’ in the 5’UTRs, and the proximal promoters, respectively. Several words were hit partially as members of the ‘interesting’ word clusters (CITATCCN, GCCACGTN, AANGTCAA, CNGATCNA), while others were not represented in the selected word clusters and the top 25 words. While no overlap for GACCGTNC could be found, it is possible to validate the significance of GGCCCANN and GCCGGAAN through the detection of these two words as unwords in the introns, marking them interesting regulatory elements associated with the expression, but not necessarily with the regulation of the associated alternative splicing process.

Conclusion
The analyses described here provide a first view of the word landscape within the non-coding regions of the Arabidopsis thaliana genome. An analysis centred on the statistically interesting words furnishes important insights into the unique elements of each segment. The correlations of particular words with cellular functions or expression patterns provide valuable hypotheses for further experimentation. Correlation between word position and expression also seems strong, with one class of words only present in the 5’/3’UTRs and introns, and another class of words putatively functioning only in the region upstream of the TSS. Words in the first class seem more directed at regulation of tissue and cellular identity, while words which function upstream appear more likely to be involved in environmental responses.

Methods
Word-based genomic signatures are the union of results generated by applying the software pipeline shown in Figure 11. Statistically relevant words are extracted from a set of genomic sequences, and are analyzed to determine similarity, location distribution, groupings, and predicted cellular function.

Sequence Data
This manuscript reports the results of analyzing DNA sequences of Arabidopsis thaliana. The non-coding genomic segments (specifically, the 3’UTRs, 5’UTRs, promoters and introns) and the entire genomic sequence (as complete chromosomes) were obtained from TAIR (release 8) [19]. Both masked and unmasked versions of the genome were analyzed. Ambiguous nucleotides, depicted in the sequences by the letters [R, Y, W, S, K, M], were removed because they represent sequencing anomalies; this resulted in the removal of 0.15% (or 188,820) of the nucleotides.

In this study, only protein-coding genes were considered as genes, and transposable-like, or pseudo-genes, were omitted. Thus, the total number of genes in this study is ~27,000. Due to different lengths and locations of the promoter elements it is possible that, while core promoters can occur for a specific gene, no distal promoter for that gene exists due to the fact that its location would fall into another gene or even outside of a chromosome. The difference in number of genes in 3’UTRs and 5’UTRs sets compared to other sets is due to genes that lack annotated UTR (it is yet to be discovered).

Whenever multiple spliced transcripts were available for a gene, a major transcript was chosen (Atngnnnnn.1) to prevent bias towards genes that contain multiple transcripts. Likewise, only introns of major transcripts were selected.

Word Enumeration and Scoring
The first pipeline stage employs a radix trie data structure [48] to enumerate all subsequences (words) of a specified length in the given DNA input sequences. For each word w, with o total occurrences in s sequences, a word score is computed as $s \cdot \ln(s/E_s(w))$. The expected number of sequences containing word w, $E_s(w)$, is computed as the product of (1) the probability for each observed word to occur anywhere in the input sequences and (2) the total length of the sequences. This model implicitly assumes a binomial model for the word distribution, i.e., that the word probabilities are independent of the positions of the words within the sequences [49,50]. The probability is computed by using a maximum-order homogeneous Markov chain model [49] where the transition probabilities are determined using the Maximum Likelihood Method [50]. (Note that under this model, the (G+C)% biasing is achieved for any order of Markov model greater than or equal to zero, since the frequencies of individual nucleotides are taken into consideration for all orders.) The order of the Markov model was chosen by using a standard chi-square test to assess the appropriateness of Markov chains of orders 0 to 6. To provide the highest precision for computation of expected values, the highest order model that passed the chi-square test was selected. Thus, an order 6 model was selected.

A p-value for each word (representing the probability of obtaining a score at least as high as the one observed [51]) is calculated by using a binomial word distribution to determine the probability of obtaining at least o repeats in the s input sequences that contain w.
**Word Clustering**

The Word Clustering stage computes a cluster for each of the top scoring words (seed words) identified in the Word Scoring phase. A cluster is computed from a seed word by determining the set of words whose Hamming distance is within a user-specified threshold. A Position/Weight Matrix (PWM) is constructed for each cluster [52], and a sequence logo is created from each PWM using the TFBS module by Lenhart and Wasserman [53]. For example, the PWM for the seed word ATTTTGTA in the 3'UTRs is as follows:

\[
\begin{pmatrix}
0.78 & 0.05 & 0.03 & 0.03 & 0.01 & 0.05 & 0.05 & 0.81 \\
0.05 & 0.03 & 0.02 & 0.01 & 0.01 & 0.03 & 0.02 & 0.03 \\
0.06 & 0.03 & 0.04 & 0.03 & 0.02 & 0.89 & 0.03 & 0.04 \\
0.11 & 0.89 & 0.91 & 0.93 & 0.96 & 0.04 & 0.90 & 0.12 \\
\end{pmatrix}
\]

**Figure 11**

Process Flowchart. Methodology flow applied for the discovery of word-based genomic signatures in non-coding *Arabidopsis thaliana.*
The columns of the PWM correspond to nucleotide positions and the rows correspond to the nucleotides A, C, G, and T, respectively.

**Word Location Distribution**
For selected words from the different segments it was determined if they were clustered at specific locations along the corresponding sequences in which they occur. In order to detect a location bias, representative of such clusters, histograms were created to show the numbers of occurrences of a specific word at each point corresponding to a positional offset from the transcription start site (TSS). For uniformity, sequence lengths were normalized to the range [1;100], to represent the number of nucleotides between the position and the TSS.

**Co-Occurrence Analysis**
The Co-Occurrence Analysis considers all non-overlapping pairs of the top ranked words and computes the expected number of sequences that contain both words. Subsequently, the observed number of sequences that contain both words is determined, and an observed-to-expected ratio is computed (using a binomial word distribution) for each word pair.

**AGRIS Lookup**
Previously published and curated binding site motifs which are equal to or shorter than eight base pairs were extracted from the AGRIS AtcisDB database [25], and were compared with the word lists generated for the different segments. For each motif the corresponding entries in word list were determined and the highest scoring word was identified.

**Determine Cellular Function**
The MIPS Functional Catalogue Database (FunCatDB) [54], was used for determining over-represented cellular functions in each gene list containing a particular word. The workflow of the cellular function analysis, labelled as “Cellular Function” in the larger process flow (Figure 11) is as follows. For each word in the ‘top 25’ lists (Tables 2, 3, 4, 5, 6, 7, &8) we determined the list of genes that contained the word being analyzed in the corresponding region. Then we determined the functional category of each gene by using the functional category scheme (version 2.1) retrieved from FunCatDB. The p-values for enrichment of categories were calculated by statistical tests with the hypergeometric distribution. After filtering out p-values greater than 1E-5, results were visualized by the matrix2png software package [55].

Analysis of the correlation between word location and gene expression was done as described in [39] with the following exceptions. A larger database was constructed from 1305 available raw microarray datasets (Additional file 19) present in NASC affyarrays http://www.arabidopsis.info and the gene expression omnibus http://www.ncbi.nlm.nih.gov/geo/. The p-value was calculated using a chi-squared test comparing genes 2-fold induced, 2-fold suppressed, or neutral between observed (all genes harbouring the word) and expected values (based on genomic average). The Bonferroni correction was used to adjust for multiple hypothesis testing. Microarray sources included a large tissue macro-dissection [56], and the follow-up studies on stress, hormones, and pathogens [57]. We included the laser capture microdissected tissue microarray datasets [58], the gene expression profile of the Arabidopsis root [59], analysis of brassinosteroids [60], and the numerous other experiments found in the collected dataset in the above mentioned repositories. Data were normalized using global scaling of the middle 96% data points, and then noise filtered using a t-test of signal vs. background, and a t-test of signal vs. control.

**Authors’ contributions**
JL contributed in the development of algorithms and models, the implementation of algorithms, generation of the results and drafting of the document. JDW contributed in the development of the models and algorithms and the implementation of the approaches. KK contributed in the development, implementation and testing of models and algorithms. XL contributed in the development of the models and algorithms for co-occurrence analysis and generated the respective data. FD contributed in the development of models and algorithms, and in the implementation of the methods. MG conducted correlation analysis between word presence/location and gene expression pattern. KE contributed the idea of Hamming-distance-based clustering. SSL contributed to the statistical foundations of the scoring model. AY’s contributions include extraction of data sets, functional analysis of words, and writing the manuscript. EG contributed to writing the manuscript and integrating the identified words with existing knowledge on control of gene expression. In addition to architecting the software pipeline employed in this research, LRW contributed to the design, implementation and validation of models and algorithms (especially in the areas of word searching and word scoring) and to the writing of this manuscript.

All authors read and approved the final manuscript.
### Additional material

**Additional file 1**  
Words discovered in 3'UTRs. Entire set of words discovered in the 3'UTRs with occurrences, expected occurrences, scores, reverse complement information and p-value.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S1.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S1.CSV)

**Additional file 2**  
Words discovered in 5'UTRs. Entire set of words discovered in the 5'UTRs with occurrences, expected occurrences, scores, reverse complement information and p-value.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S2.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S2.CSV)

**Additional file 3**  
Words discovered in introns. Entire set of words discovered in the introns with occurrences, expected occurrences, scores, reverse complement information and p-value.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S3.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S3.CSV)

**Additional file 4**  
Words discovered in core promoters. Entire set of words discovered in the core promoters [-100;+1] with occurrences, expected occurrences, scores, reverse complement information and p-value.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S4.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S4.CSV)

**Additional file 5**  
Words discovered in proximal promoters. Entire set of words discovered in the proximal promoters [-1,000;-101] with occurrences, expected occurrences, scores, reverse complement information and p-value.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S5.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S5.CSV)

**Additional file 6**  
Words discovered in distal promoters. Entire set of words discovered in the distal promoters [-3,000;-1,001] with occurrences, expected occurrences, scores, reverse complement information and p-value.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S6.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S6.CSV)

**Additional file 7**  
Words discovered in entire genome. Entire set of words discovered in the complete genome with occurrences, expected occurrences, scores, reverse complement information and p-value.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S7.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S7.CSV)

**Additional file 8**  
Words missed in 3'UTRs. Entire set of words expected to occur but not discovered in the 3'UTRs with expected occurrences.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S8.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S8.CSV)

**Additional file 9**  
Words missed in 5'UTRs. Entire set of words expected to occur but not discovered in the 5'UTRs with expected occurrences.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S9.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S9.CSV)

**Additional file 10**  
Words missed in introns. Entire set of words expected to occur but not discovered in the introns with expected occurrences.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S10.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S10.CSV)

**Additional file 11**  
Words missed in core promoters. Entire set of words expected to occur but not discovered in the core promoters with expected occurrences.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S11.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S11.CSV)

**Additional file 12**  
Word based clusters. Word-based clusters built around 2 overrepresented words of each non-coding segment of Arabidopsis thaliana represented by the word cluster and the sequence logo associated with said cluster. A word in a word cluster is presented through the nucleotide sequence associated with the word, the sequence count, the overall count and the StnSES score.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S12.DOC](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S12.DOC)

**Additional file 13**  
Word co-occurrences in 3'UTRs. Entire set of co-occurring words (taken from the top 25 words) discovered in the 3'UTRs with occurrence, expected occurrences and scores.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S13.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S13.CSV)

**Additional file 14**  
Word co-occurrences in 5'UTRs. Entire set of co-occurring words (taken from the top 25 words) discovered in the 5'UTRs with occurrence, expected occurrences and scores.  
Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S14.CSV](http://www.biomedcentral.com/content/supplementary/1471-2164-10-463-S14.CSV)
Acknowledgements

The Ohio University team acknowledges the support of the Stocker Endowment, Ohio University’s Graduate Research and Education Board (GERB), the Ohio Supercomputer Center, the Choose Ohio First Initiative of the University System of Ohio. We also wish to thank Sarah Wyatt for proofreading the manuscript. The Ohio University team further acknowledges that salaries and research support are provided by state funds appropriated to the Ohio Plant Biotechnology Consortium through The Ohio State University, Ohio Agricultural Research and Development Center. MG acknowledges the support of the plant biology department at Southern Illinois University Carbondale, and would like to thank Patrick Brown and Elisabeth Fitzek for help in assembling and analysis of the microarray database and tools. Funding for EG was provided by National Science Foundation grants MCB-0418891, MCB-0705415 and by state funds appropriated to the Ohio Plant Biotechnology Consortium through The Ohio State University, Ohio Agricultural Research and Development Center. Finally we would like to acknowledge the anonymous reviewers who helped in enhancing the quality of the manuscript presented here.

Additional file 15

Word co-occurrences in introns. Entire set of co-occurring words (taken from the top 25 words) discovered in the introns with occurrence, expected occurrences and scores.

Click here for file
[http://www.biomedcentral.com/content-supplementary/1471-2164-10-463-S15.CSV]

Additional file 16

Word co-occurrences in core promoters. Entire set of co-occurring words (taken from the top 25 words) discovered in the core promoters with occurrence, expected occurrences and scores.

Click here for file
[http://www.biomedcentral.com/content-supplementary/1471-2164-10-463-S16.CSV]

Additional file 17

Word co-occurrences in proximal promoters. Entire set of co-occurring words (taken from the top 25 words) discovered in the proximal promoters with occurrence, expected occurrences and scores.

Click here for file
[http://www.biomedcentral.com/content-supplementary/1471-2164-10-463-S17.CSV]

Additional file 18

Word co-occurrences in distal promoters. Entire set of co-occurring words (taken from the top 25 words) discovered in the distal promoters with occurrence, expected occurrences and scores.

Click here for file
[http://www.biomedcentral.com/content-supplementary/1471-2164-10-463-S18.CSV]

Additional file 19

NASC Microarrays. Entire set of microarray experiments available in NASC that were used for the cellular functional analysis.

Click here for file
[http://www.biomedcentral.com/content-supplementary/1471-2164-10-463-S19.XLS]

References

1. Thijss G, Lescoat M, Marchal K, Rombouts S, De Moor B, Rouze P and Moreau Y: A higher-order background model improves the detection of promoters regulatory elements by Gibbs sampling. Bioinformatics 2001, 17(12):1113–1122.
2. Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I and Conboy JG: Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. Nucleic Acids Res 2001, 29(11):2338–2348.
3. Ham B-K, Brandon JL, Xocomost-centares B, Ringgold V, Lough TJ and Lucas WJ: A Polypyrimidine Tract Binding Protein, Pumpkin RBPSO, Forms the Basis of a Phloem-Mobile Ribonucleoprotein Complex. Plant Cell 2009, 21(1):197–215.
4. Huang N-C and Yu T-S: The sequences of Arabidopsis GA-INSENSITIVE RNA constitute the motifs that are necessary and sufficient for RNA long-distance trafficking. The Plant Journal 2009, 59(6):921–929.
5. Sunyaev SR, Lathe WC 3rd, Ramensky VE and Bork P: SNP frequencies in human genes an excess of rare alleles and differing modes of selection. Trends Genet 2000, 16(8):335–337.
6. Symonds VV and Lloyd AM: An Analysis of Microsatellite Loci in Arabidopsis thaliana: Mutational Dynamics and Application. Genetics 2003, 165(3):1467–1481.
7. Biemont C and Vieira C: What transposable elements tell us about genome organization and evolution: the case of Drosophila. Cytogenet Genome Res 2005, 110:25–34.
8. Nezarvanov U, Umezawa A and Oshima KG: A Regulatory Element within tata Exon 2 of Human Immunodeficiency Virus Type 1 (HIV-1) Are Characteristic of Group M but Not Group O HIV-1 Strains. Journal of Virology 1999, 73 (2):9764–9772.
9. Bornstein P, McKay J, Morishima JK, Devarayalu S and Gelinas RE: A Regulatory Element in the 5'-Flanking region of the human alpha 1(I) collagen gene. Proceedings of the National Academy of Sciences of the United States of America 1997, 84(24):8869–8873.
10. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G and Montoya M, et al: The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res 2003, 31(1):224–228.
11. Hudson ME and Quail PH: Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. Plant Physiology 2003, 133(4):1605–1616.
