

Workshop

Arthur I. Dergilev, Nina G. Orlova, Oxana B. Dobrovolskaya and Yuriy L. Orlov*

Statistical estimates of multiple transcription factors binding in the model plant genomes based on ChIP-seq data

<https://doi.org/10.1515/jib-2020-0036>

Received November 16, 2020; accepted November 25, 2021; published online December 21, 2021

Abstract: The development of high-throughput genomic sequencing coupled with chromatin immunoprecipitation technologies allows studying the binding sites of the protein transcription factors (TF) in the genome scale. The growth of data volume on the experimentally determined binding sites raises qualitatively new problems for the analysis of gene expression regulation, prediction of transcription factors target genes, and regulatory gene networks reconstruction. Genome regulation remains an insufficiently studied though plants have complex molecular regulatory mechanisms of gene expression and response to environmental stresses. It is important to develop new software tools for the analysis of the TF binding sites location and their clustering in the plant genomes, visualization, and the following statistical estimates. This study presents application of the analysis of multiple TF binding profiles in three evolutionarily distant model plant organisms. The construction and analysis of non-random ChIP-seq binding clusters of the different TFs in mammalian embryonic stem cells were discussed earlier using similar bioinformatics approaches. Such clusters of TF binding sites may indicate the gene regulatory regions, enhancers and gene transcription regulatory hubs. It can be used for analysis of the gene promoters as well as a background for transcription networks reconstruction. We discuss the statistical estimates of the TF binding sites clusters in the model plant genomes. The distributions of the number of different TFs per binding cluster follow same power law distribution for all the genomes studied. The binding clusters in *Arabidopsis thaliana* genome were discussed here in detail.

Keywords: ChIP-seq; gene expression; plant genomes; regulatory gene networks; transcription factor binding sites; transcription regulation.

*Corresponding author: **Yuriy L. Orlov**, Novosibirsk State University, 630090 Novosibirsk, Russia; Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia; Agrarian and Technological Institute, Peoples' Friendship University of Russia, 117198 Moscow, Russia; and The Digital Health Institute, I.M.Sechenov First Moscow State Medical University (Sechenov University), 119991 Moscow, Russia, E-mail: orlov@bionet.nsc.ru.

<https://orcid.org/0000-0003-0587-1609>

Arthur I. Dergilev, Novosibirsk State University, 630090 Novosibirsk, Russia; and Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia.

Nina G. Orlova, Financial University under the Government of the Russian Federation, 125993 Moscow, Russia; and Moscow State Technical University of Civil Aviation, 125993 Moscow, Russia

Oxana B. Dobrovolskaya, Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia; and Agrarian and Technological Institute, Peoples' Friendship University of Russia, 117198 Moscow, Russia <https://orcid.org/0000-0002-4487-7360>

1 Introduction

The regulation of gene transcription plays a key role in cell functioning [1]. The study of transcription regulation based on the transcription factor (TF) binding data is one of the most developed bioinformatics fields. On-going acquisition of the data on gene expression and binding sites obtained by sequencing techniques such as ChIP-seq, BS-seq, DNaseI-seq, ATAC-seq, NOMe-seq [2, 3] demands development of data integration. Qualitatively new problems arise for combinatorial gene expression regulation based on multiple TFs binding. Does multiple TF binding in a gene region determine organism development program, or it is just random effects not related to gene function [4]? Statistical estimates of multiple TFs binding may highlight fundamental gene expression features common for any eukaryotic genome [5]. Linking the gene promoter regions and high-throughput TF binding data we may reconstruct regulatory gene networks in the genome scale [1]. The location of a TF binding site in genomic sequence may be remote from the gene transcription start, thus preventing reliable TF target gene prediction. Despite the development of data sophisticated analysis techniques, the prediction of the regulatory regions of transcription, based only on the nucleotide sequence or single binding events remains a complex problem.

New problems arise in the analysis of the joint location (co-localization) of the binding sites of different protein transcription factors in the same loci [6]. There are cooperative regulatory effects with the simultaneous binding of various transcription factors in protein complexes to DNA [7, 8]. It is necessary to develop the tools and methods of the sequencing data processing, evaluation of binding effects of transcription factors [7, 9–11].

We used experimental ChIP-seq data to model TF binding as the alternative to sequence-based prediction. Co-location of binding sites of two or more different factors in the promoter region of the same target gene may define an element of a regulatory gene network. Several protein transcription factors may bind to the same gene promoter region that corresponds to network interactions “TFs–gene”. One TF may have several gene targets that corresponds to “TF–genes” interactions. Protein TF may bind to the promoter of its own gene, thus forming a regulatory contour in gene network. Thus, a regulatory gene network can be reconstructed from a set of the binding sites of several different TFs. Such a regulatory network might be conserved between species. The examples of the mutual regulation of several pluripotency factors in embryonic stem cells in mouse were shown [12]. The dynamic models of regulatory gene networks have also been developed for the pluripotency factors binding to the same regulatory regions [13]. Possible overlapping of binding sites on the same nucleotide sequence of multiple TF binding clusters, and mutual location of the binding sites may follow certain regularities and rules to be found by statistics of DNA oligonucleotides [14, 15].

A physical explanation of the observed binding patterns on DNA may be due to the protein-protein interactions between the TFs bound as multimer complex. Protein transcription factors can directly contact each other, simultaneously bind to DNA in neighboring regions, or compete for binding to the same DNA sequence in time [7]. ChIP-seq data generally do not provide an experimental answer to the question: “Do the TFs bind the same genome locus simultaneously or the binding events differ in time?” [7, 12]. Thus, we can only estimate frequency of the interactions, but not prove cooperative binding of the proteins to the same DNA site.

The binding regulatory region could be assessed taking into Account Topologically Associated Domains (TAD) in the chromosomes, which could be revealed using sequencing technologies based on chromosome conformation capture (3C, with genome-wide variants Hi-C and ChIA-PET) [11, 16–18]. Clusters of the binding sites on the chromosomes mark topological domains studied mainly in mammalian genomes [17]. Chromosome conformation data (Hi-C) were obtained initially for Arabidopsis, and recently for other crop plants models. TADs are absent in Arabidopsis but they are found in more complex plant crop genomes including rice [19].

In general, the problem of joint gene expression regulation by a set of TFs has not been sufficiently studied [6, 20]. For plant genomes, such studies are presented in separate databases [21–25]. Despite different models of TF-DNA binding, statistical estimates of the distribution of the number of different TF in binding clusters

were not discussed. We applied earlier developed computer scripts [9] to analyze ChIP-seq data, calculate clusters and visualize them in the form of heat maps in plants.

We used ChIP-seq data of three plants, including *Arabidopsis thaliana*, *Physcomitrella patens*, and *Chlamydomonas reinhardtii* to model TF binding clusters in their genomes and estimate the distribution of the TF in the clusters found.

2 Materials and methods

2.1 Current version of TFBS tool analysis portable

A set of scripts in the Python language has been developed earlier for computer analysis of clusters of binding sites for transcription factors (see Figure 1) with a graphical shell implemented in the Qt 5 environment [9]. We compiled it to the cross-platform software tool ‘TFBS Tools Analysis’.

For the full usage of TFBS Analysis Portable, additional software is required: an installed Python interpreter of version 3.5 or higher (<https://www.python.org/downloads/windows/>), programming environment R version 3.4.1 or higher (<https://www.r-project.org/>), Git Bash for Windows (<https://gitforwindows.org/>).

TFBS analysis portable allows conveniently convert and analyze the TF binding sites data in genome scale, namely:

1. Load data (TF binding sites coordinates in a simplified BED format)
2. Load the chromosome coordinates in GFF 3 data format (option including subsequent data conversion using an additional script).

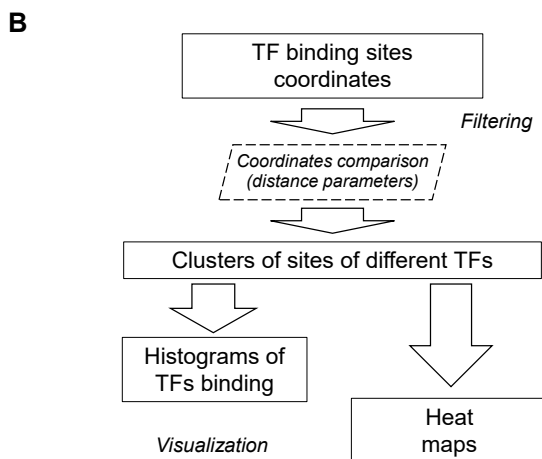
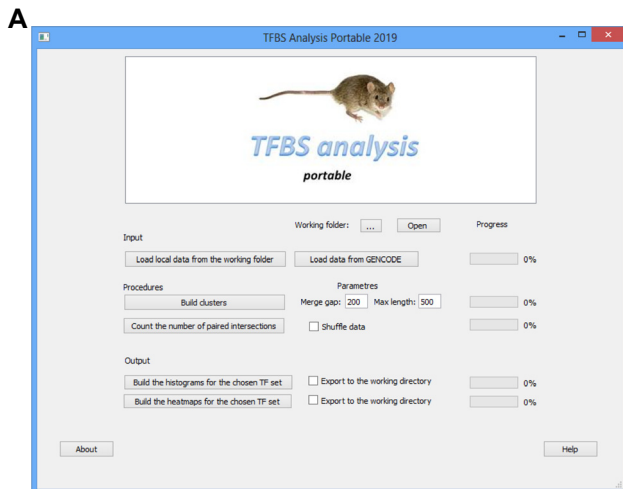


Figure 1: TFBS analysis portable.

A – toolbox main window; B – block-scheme of the data processing.

3. Build coordinates of the clusters of TFs binding sites iteratively based on coordinates of single sites (adjusting the parameters of the gap between the coordinates of the sites and the maximum cluster region size, in nucleotides).
4. Calculate pairwise intersections of the TFs binding loci in the genome.
5. Construct the statistical histograms of the cluster occurrence for each TF separately and in the groups.
6. Build heat maps of TF co-binding in the given genome.

The program did not filter the initial data on the quality of DNA reads, on the basis of which the binding sites of transcription factors in the genome were determined. However, such data pre-processing is necessary [26, 27] and can be included in the general pipeline of analysis and subsequent refinement of site clusters. The detection of binding sites from DNA reads in ChIP-seq experiments provides a set of genome loci for a given TF. ChIP-seq experiments give several sets of TF binding sites in genome (chromosome) coordinates. Comparing the binding coordinates we construct a set of common loci, occupied by two or more TF – i.e. a binding cluster. The width of such clusters in genome is limited by the distance parameter (normally less than 500 nucleotides). The size 500 nt refers to the standard promoter size. It corresponds to the distance where proteins bound to DNA can physically interact to each other. This parameter may vary in the program runs. We adjusted it by the calculation of distance between peaks in ChIP-seq genomes profiles to 200 nt. The number of different TF in the binding cluster is not limited (up to all available TF data set in the genome under study).

2.2 The data and algorithms

We download available ChIP-seq data from UCSC Genome Browser (<http://genome.ucsc.edu/>) and the plant genome databases [28]. The TFBS Analysis Portable tool has flexibility in relation to various genomes differing by genome size and number of chromosomes. Currently, this code is available upon request to the authors. The data in the form of BED files were downloaded from the databases: NCBI (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24568>), PlantTFDB (<http://planttfdb.cbi.pku.edu.cn>, see also <http://ucsc.gao-lab.org/>) [29], PlantRegMap (<http://plantregmap.cbi.pku.edu.cn>) [28].

In total we have downloaded the ChIP-seq peaks for 28 transcription factors for *A. thaliana* (flowering plants) [30], 18 factors for *P. patens* (moss) and eight factors for *C. reinhardtii* (green algae, plant general type) [28, 29].

These plants present taxons from a simple unicellular alga to a more complex flowering plant. The evolutionary relationship between these plants can be seen on the phylogenetic tree shown in Figure 2.

Note the detailed databases on genome-wide TF binding developed earlier – TRRD, GTRD, and HOCOMOCO [31, 32]. However, data on TF binding in plants are rather scattered [21, 22, 30, 33]. For comparison with predicted binding sites in Arabidopsis we used AthaMap resource [34] and JASPAR [35]. We plan to integrate novel available data on TF binding in plants [36].

Transcription factor binding sites tend to be found together, especially in gene promoter regions. See example for *A. thaliana* gene AT1G01040 from plant genome browser (Figure 3) (<http://ucsc.gao-lab.org/>). Conserved TF binding sites overall each other in chromosome coordinates.

To construct the clusters, we have iteratively compared the coordinates of the binding sites for each TF. The ChIP-seq peaks (summit position in the given ChIP-seq profile) were counted as the binding site positions. Thus, we used experimental data instead of sequence-based prediction. However, the tool can take as input data any set of genome coordinates, both experimental and predicted. The distances between the sites (peaks) were compared, counting for the new cluster if the distance was less than 200 nt. Figure 4 shows scheme of cluster counting by TF binding coordinates.

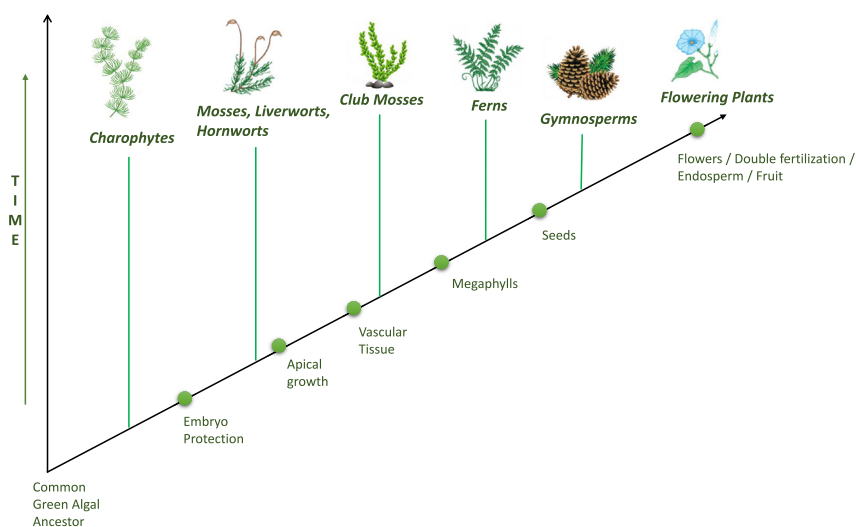


Figure 2: A schematic representation of a phylogenetic tree representing the process of plant evolution over time. The order of genomes studied is from *mosses and algae* (*Chlamydomonas reinhardtii*, and *Physcomitrella patens*) to more complex flowering plants (*Arabidopsis thaliana*).

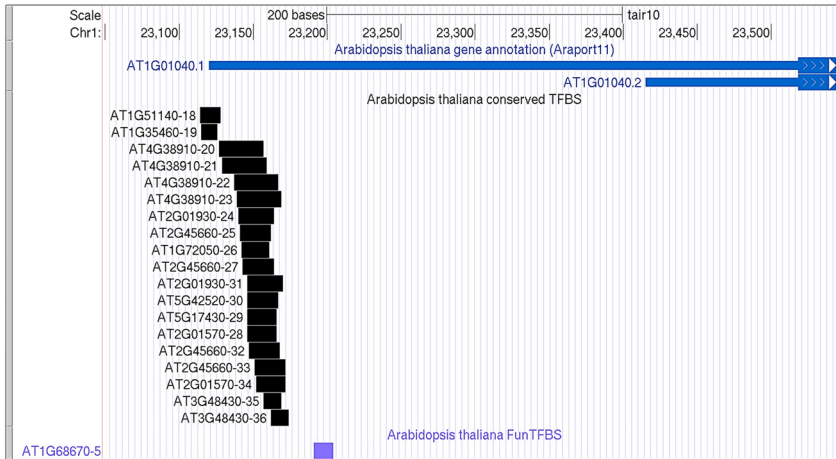


Figure 3: The example of multiple transcription factors binding in the *A. thaliana* genome in promoter region of AT1G01040 (DCL1) gene (the plant genome browser, <http://ucsc.gao-lab.org/>).

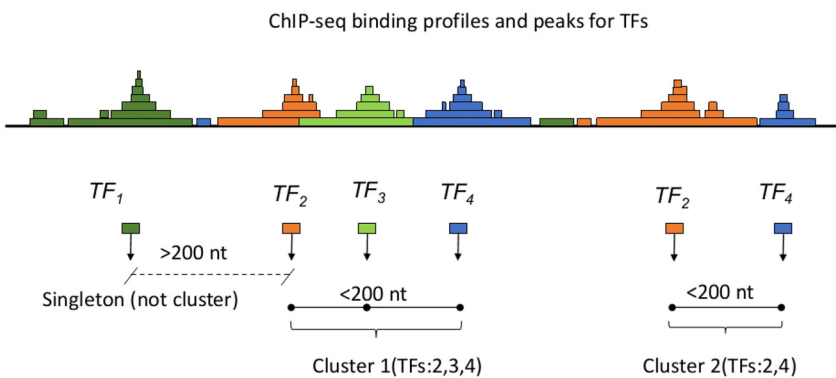


Figure 4: Example of the processing of ChIP-seq profiles (peaks) for the TFs binding to count binding clusters in genome coordinates.

The parameter for the clusters construction is 200 nt gap between the experimentally determined positions of the binding sites in genome coordinates. If taking a longer distance (500 nt) many not related TF sites would be counted in clusters. If considering only neighboring TF binding sites, with a distance less than 20 nt in between, the maximum cluster size is reduced while maintaining the shape of the distribution of the number of sites (data not shown). We had followed the algorithm of cluster construction described first in [12] and later extended in [6].

2.3 Functional features of the studied data

We used all available ChIP-seq data for TFs binding in the chosen model plants. These TFs are important for plant development and were previously studied [21]. The following sets of TFs ChIP-seq data were counted:

A. thaliana: AP2, ARF, B3, BBR, BPC, bZIP, C2H2, CH3, CAMTA, CPP, DOF, E2F/DP, ERF, FAR1, G2-like, GATA, GRAS, HSF, HD-ZIP, LBD, LFY, MADS, MYB, NAC, SRS, TCP, WRKY, YABBY.

P. patens: ARF, bZIP, C3H, CAMTA, CPP, WOX, Dof, E2F/DP, ERF, G2, GATA, HD-ZIP, LBD, LFY, MYB, NAC, TCP, WRKY.

C. reinhardtii: C2H2, C3H, CPP, Dof, ERF, G2-like, GATA, MYB.

Consider functions of these transcription factors in brief. AP2/ERF regulate plant development and response to various types of biotic and environmental stress. ARF is a key regulator of virtually all aspects of plant growth and development, from embryogenesis to aging.

In plants, transcription factors of the leucine zipper motif (bZIP) regulate the processes of protection against pathogens, transmission of light and stress signals, seed maturation and flower development. C2H2 is a sequence-specific DNA binding protein regulating transcription. The GATA factors form a class of transcription regulators present in mushrooms, animals and plants which generally recognize a consensus sequence WGATAR [37].

The plant-specific transcription factor LEAFY (LFY) plays a central, evolutionarily conserved role in the transition from vegetative growth to flowering. MADS are DNA-binding proteins that play an essential role in a variety of biological processes in the development and flower plants [38]. TCP are plant-specific transcription factors that play an important regulatory role in plant growth [39]. WRKY are one of the largest families of transcriptional regulators found exclusively in plants. They perform a variety of biological functions in relation to plant resistance to diseases, response to abiotic stress and processes controlled by hormones.

The main function of the YABBY genes is to determine the fate of abaxial cells in the lateral organs of *Arabidopsis* [40]. In particular, YABBY-regulated WRKYs play an important role in the regulation of plant stress responses, both as activators and repressors. Overall, the TFs used in the study were compiled by ChIP-seq data availability in the genomes, not restricting to the function.

3 Results and discussion

Using coordinates of ChIP-seq binding peaks we have iteratively compared the binding sites coordinates and constructed the coordinates of the clusters for every genome. The statistical histograms have been built for all the plants studied. Figure 5 shows the number of clusters depending on the cluster size (the number of sites it contains).

The probability of clustering several binding sites decreases by Power law distribution. One can see it from the observed histograms in log-scale (with some variability). Clusters of sites with a size of four and higher in the genome can no longer be explained by random causes, as it was shown previously earlier theoretically and using computer simulations [12]. The number of different TF in a binding cluster generally obeys the distribution of the probability of obtaining group of elements of a given size [41, 42]. However, the observed distribution of the number of different TF per cluster differs from the expected by chance: it follows power-law distribution with longer right-tail (clusters of very large size not expected by chance).

The figure shows the presence of large clusters (two, three or more different TFs). There are up to several hundred binding clusters, for each of the considered plants.

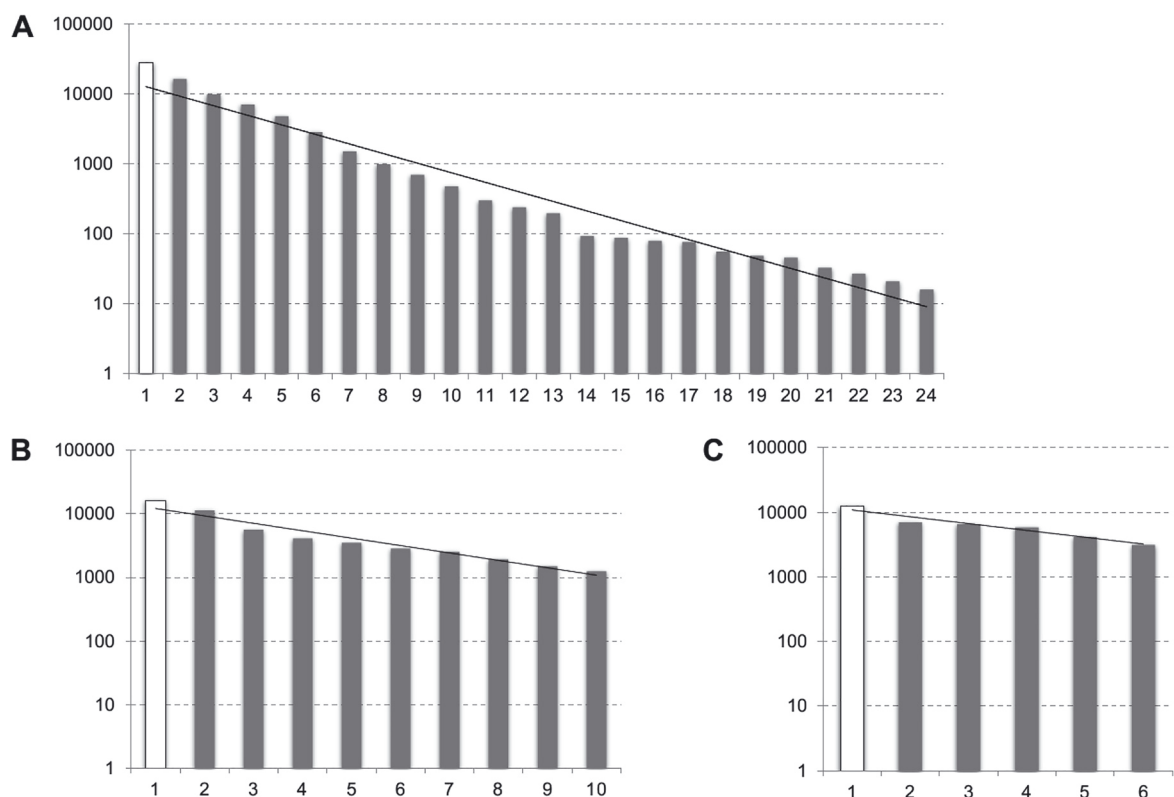


Figure 5: Dependence of the number of clusters of binding sites on cluster size.

A – *Arabidopsis thaliana*, B – *Physcomitrella patens*, C – *Chlamydomonas reinhardtii*. Axis X shows the number of sites in the cluster, axis Y – the number of such clusters. Single binding sites (size 1, singletons) are shown by the hollow column. The values are displayed on semi-logarithmic scale.

It is interesting to note the existence of very complex clusters containing binding sites for almost all transcription factors available (10 or more TFs). For *A. thaliana*, the maximum cluster size is 24. For *P. patens* – the maximal size is 10. For *Chlamydomonas Reinhardtii* – 18.

Thus, such regions could act as super-enhancers. The term super enhancer was introduced in the literature for gene regulatory regions mediating expression of many genes in different tissues first for stem cells [43]. Super-enhancers are large clusters of transcriptional enhancers that are co-occupied by multiple lineage-specific transcription factors driving expression of genes that define cell identity. For plants such enhancer may drive development programs, switch plant cell response to external factors. If we consider structure of gene regulatory network the gene having such large TF binding cluster in its promoter region will be seen as a network hub with multiple links.

Statistical histograms were constructed showing the individual occurrence of each factor for a given set of plants (Figure 6).

It is seen from the figures that the number of clusters decreases exponentially with the size of the cluster. This dependence approximately follows Power law distribution. It is theoretically expected. At the same time, there is no genomic region containing simultaneously the sites of all factors for each plant which indicates their functional differences.

Let's consider the functions of some factors found in the binding clusters.

It can be seen that the LFY factor, which is responsible for the growth of flowers in plants, tends to occur in the Arabidopsis genome most often among all the others, while the SRS group factors are least expressed [30].

In the *P. patens* genome, LFY is again quantitatively in the first place, while G2, which participates in the process of the cell cycle and influences cell differentiation, occurs less frequently than others.

The CPP family of factors in *C. reinhardtii* tends to occur more often, and the factors of the group G2 like are less common.

For each pair of TFs we count number of binding clusters and put this integer number to a matrix. It is a symmetrical matrix. At the next stage for each pair of transcription factors we calculated the Pearson linear correlation coefficient by the rows (vectors of the integer numbers) from this matrix [9]. Each cell of such a symmetric matrix contains the number of binding sites (overlapping in a given genomic range up to 200 nt) of

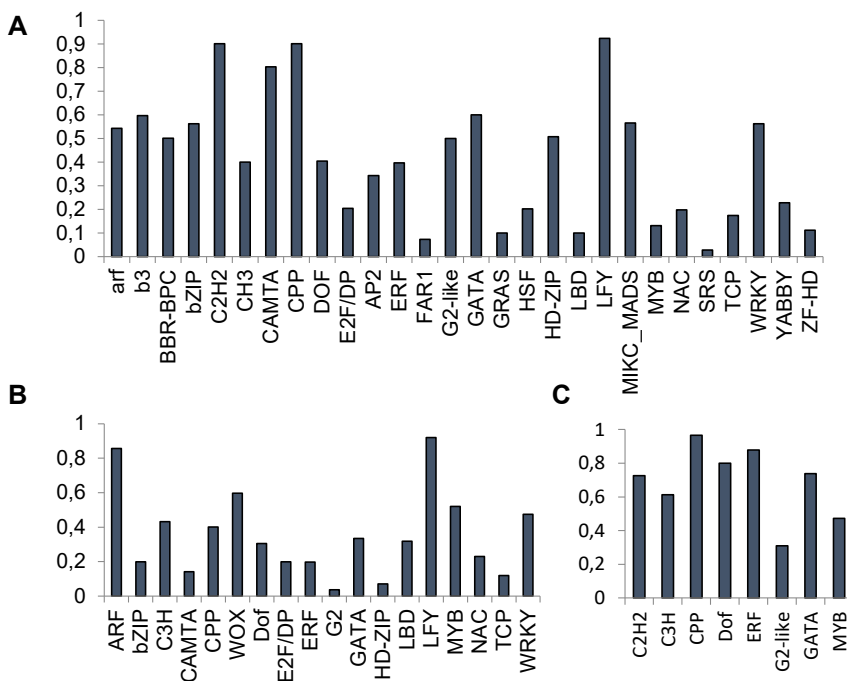


Figure 6: The occurrence of each transcription factor in the clusters (in the range [0;1]). A – *Arabidopsis thaliana*, B – *Physcomitrella patens*, C – *Chlamydomonas reinhardtii*. The name of the factor is on the horizontal axis, and its occurrence is shown on the vertical axis.

two factors. It is possible to use rank correlation coefficients as well as linear coefficients to follow the work by Chen et al. [12]. Thus, the matrix of integer numbers was translated into a correlation matrix (float numbers in [0;1] interval). The correlation coefficient gives normalization of the values for the sets of transcription factors sites of different sizes. Such a matrix provides a measure of association, between the binding sites for the pairs of transcription factors. The higher the correlation coefficient is, the closer the location of the binding sites of the transcription factors in the genome. The matrix can be represented in the form of a heat map, where the color intensity shows the correlation (the correlation coefficient).

Using this measure in the R programming environment (<https://www.r-project.org/>) the sites of the transcription factors were clustered with the following visualization. Figure 7 presents the heatmaps that highlight the intensity of joint binding of transcription factors in the genomes studied.

Figure 7 shows that the same factors GATA and MYB are in the largest cluster in all the genomes.

In the investigated set of 28 factors of the plant *A. thaliana*, the heatmaps allow delineating two major groups of TFs: {GATA, ARF, BBR-BPC, HSF, LBD, FAR1, C2H2, GRAS, MYB, TCP, E2F/DP, HD-ZIP, CH3, NAC, AP2, bZIP, WRKY} and {B3, G2-like, ZF-HD, MADS, DOF, YABBY}, or, in other words, relating to the GATA and related G2-like transcription factors.

It is also interesting to note that the LFY factor tends to occur most often among all the others, but it correlates weakly with other factors (except with SRS). Among the selected 28 factors, GATA, ARF, BBR-BPC and HSF tend to occur together more often.

In the studied set of 18 factors of the *P. patens* plant, the situation looks different: the factors are rather weakly correlated with each other, so it is difficult to distinguish large groups of co-occurring factors. Moreover, WOX [44] and GATA [45] occur together more often.

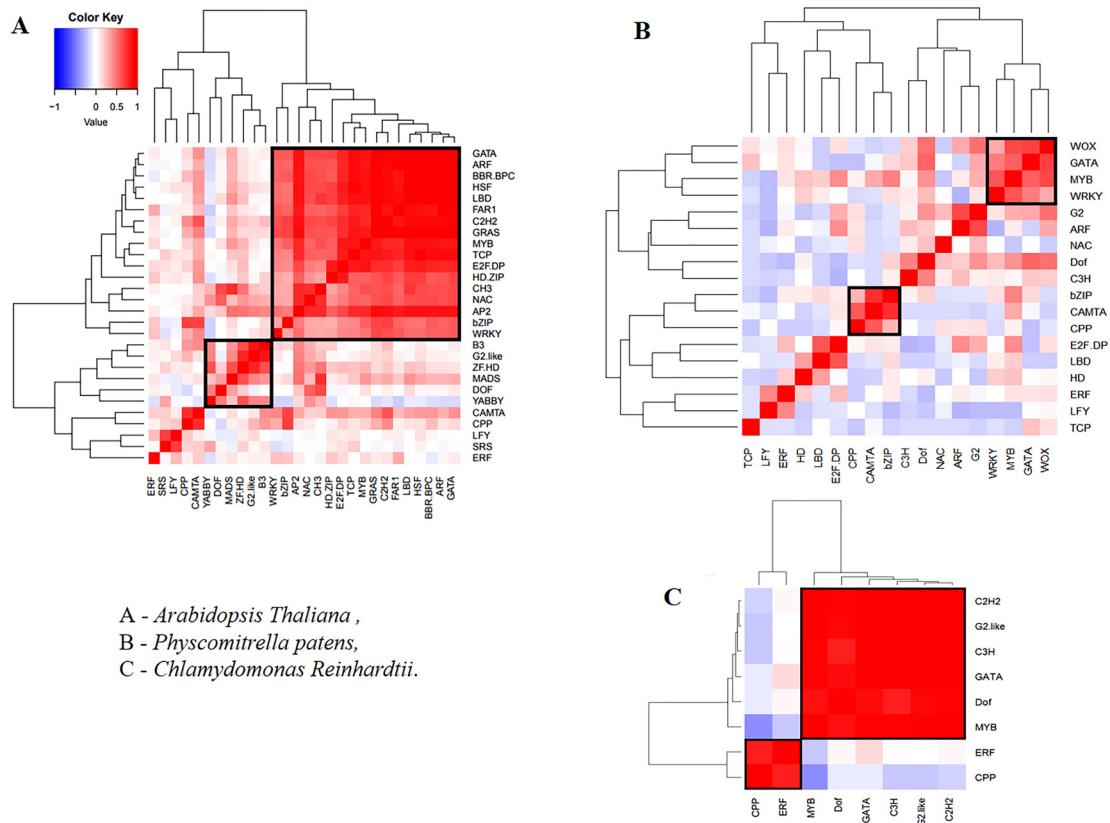


Figure 7: Heat map of joint localization of TFs in the plants studied. Red is a more frequent occurrence. The squares highlight the groups of factors that tend to occur together most frequently.

It is known that the interaction between GATA and the C/EBP family of transcription factors (which includes the bZIP factor considered in the work and some others) is critical for the GATA-mediated suppression of adipocyte differentiation [45]. It is interesting to note that C/EBP interacts with the CCAAT box motif in several gene promoters. They are characterized by a highly conserved basic leucine zipper (bZIP) domain at the C-terminus.

In the set of eight factors of the *C. reinhardtii* plant, two groups are clearly distinguished: {C2H2, C3H, CPP, Dof, G 2-like, GATA, MYB} and {CPP, ERF}.

In general, it can be noted that the evolutionarily more ancient factors GATA and MYB are presented in the clusters of sites in all studied plant species. The flowering plant-specific LFY factor is presented in binding clusters in Arabidopsis. Clustering of sites in the genome for various transcription factors can be affected by the context features of the nucleotide sequence (nucleotide composition and motifs) of a regulatory region, by the structure of the protein that determines the type of binding to DNA, and has an evolutionary constraint [14]. At the same time, epigenetic modifications, and the binding of TF to DNA in a specific tissue and plant organ should be considered. These sequencing data are currently insufficient for such assessments [21].

The general revealed regularity of the existence of clusters of binding sites in eukaryotic genomes, shown for the mouse and human genomes [12], is confirmed for plants. The general pattern in the distribution of site cluster sizes corresponds to the number of links in models of complex biological networks [46], including networks of protein–protein interactions [47]. Clusters of TF binding sites can be found using the proposed statistical estimates. Visualization of such cluster using heat maps is a convenient bioinformatics method [10].

4 Conclusions

The developed computer approach for clustering of variable sets of sites can be applied to a wide range of problems of assessing functional elements in the genome. In addition to TF binding the sites may include areas of low text complexity, tandem repeats, and CpG islands [48]. Analysis of functional clusters makes it possible to statistically describe enhancers, annotate genomes, in cases where experimental information is not available. Integration of experimental genomic information, Big Data, in general, represents an important problem in bioinformatics, requiring the integration of existing software tools and approaches [49–51]. The presented statistical analysis of clusters of binding sites in plant genomes helps in solving the problem, developing approaches to the study of the evolutionary origin of enhancers and gene networks [52, 53]. The exact positioning of sites in the clusters is affected by both nucleosome packing [54] and spatial restrictions on the sequence length in topologically associated domain, where protein transcription factors interact with DNA [17], thus arranging the binding sites by complex rules yet to be revealed.

For the considered plant genomes, analysis of an expanded set of binding sites for transcription factors confirmed a broad joint clustering of binding sites for transcription factors of the GATA family, that is, the factors characterized by their ability to bind to the GATA DNA sequence. Contrary to expectations, the LFY factor weakly correlates with the other factors studied.

In general, the proposed statistical estimates allow identifying non-random clusters of TF binding sites in plant genomes. The distribution of site clusters by size shows the general patterns of the clusters site formation in eukaryotic genomes. Such gene regulatory regions in the plant genome should be further investigated by experimental methods to identify cooperative interactions, to determine the functional role of the TF binding clusters, including response of the plant to stress conditions [55].

Acknowledgments: The authors are grateful to A. E. Dresvyannikova, R. O. Babenko, V. A. Kononov for scientific discussion and assistance in this work.

Author contribution: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Research funding: The publication has been prepared with the support of the RUDN University Strategic Academic Leadership Program (Recipients: YO, OD).

Conflict of interest statement: Authors state no conflict of interest.

References

1. Ignatieva EV, Podkolodnaya OA, Orlov YL, Vasiliev GV, Kolchanov NA. Regulatory genomics: combined experimental and computational approaches. *Russ J Genet* 2015;51:334–52.
2. Chen X, Bhadauria V, Ma B. CHIP-seq: a powerful tool for studying protein-DNA interactions in plants. *Curr Issues Mol Biol* 2018;27:171–80.
3. Orlov Y, Xu H, Afonnikov D, Lim B, Heng JC, Yuan P, et al. Computer and statistical analysis of transcription factor binding and chromatin modifications by ChIP-seq data in embryonic stem cell. *J Integr Bioinform* 2012;9:211.
4. Kazemian M, Pham H, Wolfe SA, Brodsky MH, Sinha S. Widespread evidence of cooperative DNA binding by transcription factors in *Drosophila* development. *Nucleic Acids Res* 2013;41:8237–52.
5. Ramsey SA. An empirical prior improves accuracy for bayesian estimation of transcription factor binding site frequencies within gene promoters. *Bioinf Biol Insights* 2016;9(4 Suppl):59–69.
6. Tsukanov AV, Orlova NG, Dergilev AI, Orlov YL. Programs for statistical analysis, clusterization and visualization of genome distribution of transcription factor binding sites. *Vestnik NSU. Series: Information Technologies* 2018;16:51–63.
7. Khoueiry P, Girardot C, Ciglar L, Peng P-C, Gustafson EH, Sinha S, et al. Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. *eLife* 2017;6:e28440.
8. Narasimhan K, Pillay S, Huang YH, Huang YH, Jayabal S, Udayasuryan B, et al. DNA-mediated cooperativity facilitates the co-selection of cryptic enhancer sequences by SOX2 and PAX6 transcription factors. *Nucleic Acids Res* 2015;43:1513–1528.
9. Dergilev AI, Spitsina AM, Chadaeva IV, Svichkarev AV, Naumenko FM, Kulakova EV, et al. Computer analysis of colocalization of the TFs' binding sites in the genome according to the ChIP-seq data. *Russ J Genet: Appl Res* 2017;7:513–22.
10. Kulakova EV, Spitsina AM, Bogomolov AG, Orlova NG, Dergilev AI, Chadaeva IV, et al. Program for analysis of genome distribution of chromosome contacts in cell nucleus by the data obtained using ChIA-PET and Hi-C technologies. *Prog Syst: Theory Appl* 2017;8:219–42.
11. Kulakova EV, Spitsina AM, Orlova NG, Dergilev AI, Svichkarev AV, Safronova NS, et al. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. *Prog Syst: Theory Appl* 2015;62:129–48.
12. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 2008;133:1106–17.
13. Akberdin IR, Omelyanchuk NA, Fadeev SI, Leskova NE, Oschepkova EA, Kazantsev FV, et al. Pluripotency gene network dynamics: system views from parametric analysis. *PLoS One* 2018;13:e0194464.
14. Vityaev EE, Orlov YL, Vishnevsky OV, Belenok AS, Kolchanov NA. Computer system “Gene Discovery” to search for patterns in eukaryotic regulatory nucleotide sequences. *Mol Biol* 2001;35:810–7.
15. Vityaev EE, Orlov YL, Vishnevsky OV, Pozdnyakov MA, Kolchanov NA. Computer system “Gene Discovery” for promoter structure analysis. *Silico Biol* 2002;2:257–62.
16. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 2009;462:58–64.
17. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 2012;148:84–98.
18. Orlov YL, Thierry O, Bogomolov AG, Tsukanov AV, Kulakova EV, Galieva ER, et al. [Computer methods of analysis of chromosome contacts in the cell nucleus based on sequencing technology data]. *Biomed Khim* 2017;63:418–22.
19. Grob S. Three-dimensional chromosome organization in flowering plants. *Brief Funct Genomics* 2020;19:83–91.
20. Ibarra IL, Hollmann NM, Klaus B, Augsten S, Velten B, Hennig J, et al. Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Nat Commun* 2020;11:124.
21. Chow CN, Lee TY, Hung YC, Li GZ, Tseng KC, Liu YH, et al. PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Res* 2019;47:D1155–63.
22. Hao Q, Zhang L, Yang Y, Shan Z, Zhou XA. Genome-wide analysis of the WOX gene family and function exploration of GmWOX18 in soybean. *Plants* 2019;8:215.
23. Muhammad II, Kong SL, Akmar Abdullah SN, Munusamy U. RNA-seq and ChIP-seq as complementary approaches for comprehension of plant transcriptional regulatory mechanism. *Int J Mol Sci* 2019;21:167.
24. Orlov YL, Salina EA, Eslami G, Kochetov AV. Plant biology research at BGRS-2018. *BMC Plant Biol* 2019;19(1 Suppl):56.
25. Zakhartsev M, Medvedeva I, Orlov Y, Akberdin I, Krebs O, Schulze WX. Metabolic model of central carbon and energy metabolisms of growing *Arabidopsis thaliana* in relation to sucrose translocation. *BMC Plant Biol* 2016;16:262.

26. Abnizova I, te Boekhorst R, Orlov Y. Computational errors and biases of short read next generation sequencing. *J Proteomics Bioinf* 2017;10:1–17.
27. Naumenko FM, Abnizova II, Beka N, Genaev MA, Orlov YL. Novel read density distribution score shows possible aligner artefacts, when mapping a single chromosome. *BMC Genom* 2018;19(3 Suppl):92.
28. Tian F, Yang DC, Meng YQ, Jin J, Gao G. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res* 2020;48:D1104–D1113.
29. Jin J, Zhang H, Kong L, Gao G, Luo J. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res* 2014;42:D1182–7.
30. Fridborg I, Kuusk S, Moritz T, Sundberg E. The Arabidopsis dwarf mutant shi exhibits reduced gibberellin responses conferred by overexpression of a new putative zinc finger protein. *Plant Cell* 1999;11:1019–32.
31. Kolmykov S, Yevshin I, Kulyashov M, Sharipov R, Kondrakhin Y, Makeev VJ, et al. GTRD: an integrated view of transcription regulation. *Nucleic Acids Res* 2021;49:D104–D111.
32. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale CHIP-Seq analysis. *Nucleic Acids Res* 2018;46:D252–9.
33. Gorshkov VM, Samarina LS, Kulyan RV, Malyarovskaya VI, Ryndin AV, Rakhmangulov RS, et al. Challenges of in vitro conservation of citrus germplasm resources. *Vavilov J of Gene Breed* 2019;23:24–8.
34. Hehl R, Norval L, Romanov A, Bülow L. Boosting AthaMap database content with data from protein binding microarrays. *Plant Cell Physiol* 2016;57:e4.
35. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 2004;5:276–87.
36. Di Stefano M, Nützmann HW, Marti-Renom MA, Jost D. Polymer modelling unveils the roles of heterochromatin and nucleolar organizing regions in shaping 3D genome organization in Arabidopsis thaliana. *Nucleic Acids Res* 2021;49:1840–58.
37. Lowry JA, Atchley WR. Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding Domain. *J Mol Evol* 2000;50:103–15.
38. Lai X, Daher H, Galien A, Hugouvieux V, Zubieta C. Structural basis for plant MADS transcription factor oligomerization. *Comput Struct Biotechnol J* 2019;17:946–953.
39. Zheng K, Ni Z, Qu Y, Ca Y, Yang Z, Sun G, et al. Genome-wide identification and expression analyses of TCP transcription factor genes in *Gossypium barbadense*. *Sci Rep* 2018;8. <https://doi.org/10.1038/s41598-018-32626-5>.
40. Villanueva JM, Broadhvest J, Hauser BA, Meister RJ, Schneitz K, Gasser CS. INNER NO OUTER regulates abaxial- adaxial patterning in Arabidopsis ovules. *Genes Dev* 1999;13:3160–9.
41. Kuznetsov VA, Orlov YL, Wei CL, Ruan Y. Computational analysis and modeling of genome-scale avidity distribution of transcription factor binding sites in chip-pet experiments. *Genome Inform* 2007;19:83–94.
42. Wang J, Meng X, Dobrovolskaya OB, Orlov YL, Chen M. Non-coding RNAs and their roles in stress response in plants. *Dev Reprod Biol* 2017;15:301–12.
43. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 2013;153:307–19.
44. Costanzo E, Trehin C, Vandenbussche M. The role of WOX genes in flower development. *Ann Bot* 2014;114:1545–53.
45. Tong Q, Tsai J, Tan G, Dalgin G, Hotamisligil GS. Interaction between GATA and the C/EBP family of transcription factors is critical in GATA-mediated suppression of adipocyte differentiation. *Mol Cell Biol* 2005;25:706–15.
46. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. *Nature* 2000;407:651–4.
47. Rolland T, Taşan M, Charlotiaux B, Pevzner SJ, Zhong Q. A proteome-scale map of the human interactome network. *Cell* 2014;159:1212–26.
48. Babenko VN, Bogomolov AG, Babenko RO, Galieva ER, Orlov YL. CpG islands' clustering uncovers early development genes in the human genome. *Comput Sci Inf Syst* 2018;15:473–85.
49. Chen M, Harrison A, Shanahan H, Orlov Y. Biological Big bytes: integrative analysis of large biological datasets. *J Integr Bioinform* 2017;14. <https://doi.org/10.1515/jib-2017-0052>.
50. Orlov YL, Galieva AG, Orlova NG, Ivanova EN, Mozyleva YA, Anashkina AA. Reconstruction of gene network associated with Parkinson disease for gene targets search. *Biomed Khim* 2021;67:222–30.
51. Orlov YL, Bragin AO, Babenko RO, Dresvyannikova AE, Kovalev SS, Shaderkin IA, et al. Integrated computer analysis of genomic sequencing data based on ICGenomics tool. In: Hu Z, Petoukhov S, He M, editors, *Advances in Intelligent Systems, Computer Science and Digital Economics*. CSDEIS 2019, AISC 1127, International Journal of Intelligent Systems and Applications (IJISA); 2020. pp. 154–64.
52. Dobrovolskaya O, Pont C, Sibout R, Martinek P, Badaeva E, Murat F, et al. FRIZZY PANICLE drives supernumerary spikelets in bread wheat. *Plant Physiol* 2015;167:189–99.
53. He X, Duque TS, Sinha S. Evolutionary origins of transcription factor binding site clusters. *Mol Biol Evol* 2012;29:1059–70.

54. Goh WS, Orlov Y, Li J, Clarke ND. Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. *PLoS Comput Biol* 2010;6:e1000649.
55. Doroshkov AV, Konstantinov DK, Afonnikov DA, Gunbin KV. The evolution of gene regulatory networks controlling *Arabidopsis thaliana* L. trichome development. *BMC Plant Biol* 2019;19(1 Suppl):53.

Supplementary Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/jib-2020-0036>).