Statistical estimates of the transcription factor binding sites cluster in Arabidopsis and distant plant species genomes

Orlov Y.L.*, Dergilev A.I., Dobrovolskaya O.B.

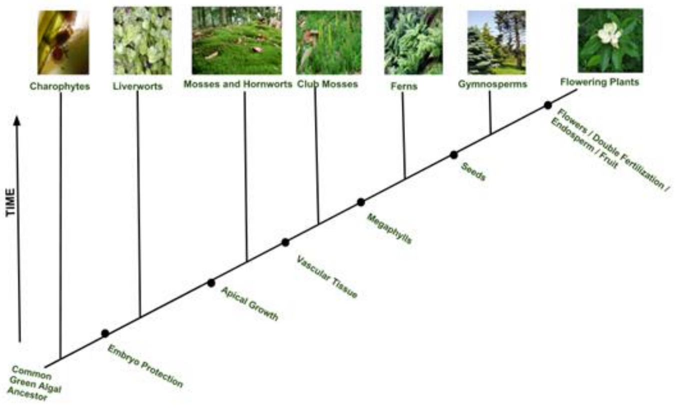
[1]Novosibirsk State University, Novosibirsk, Russia
[2]Peoples' Friendship University of Russia (RUDN University), Moscow, Russia
[3]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

The development of high-performance sequencing technologies allows to study the binding sites of protein transcription factors in genome scale. The clusters of transcription factor binding sites determine regulatory gene networks and evolutionary patterns. We discuss statistical estimates of the binding sites clusters found.

Novosibirsk, 2021

The problem of plant genome analysis and transcription regulation

The growth of data volume on the experimentally determined binding sites raises qualitatively new problems for the analysis of gene expression, determining target genes for transcription factors. However, such data were not investigated in plant genomes in detail comparing to mammalian genomes. Plant genomes remain an insufficiently studied object, although they have complex molecular regulatory mechanisms of gene expression and response to the environmental stresses.



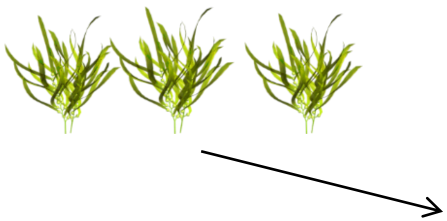
1. *Arabidopsis thaliana (flowering plant)*
   (5 chromosomes, ~ 150 Mb)
2. *Physcomitrella patens (seaweed)*
   (27 chromosomes, ~ 1 Gb)
3. *Klebsormidium subtile (Charophyta, green seaweed)*
   (17 chromosomes, ~ 500 Mb)

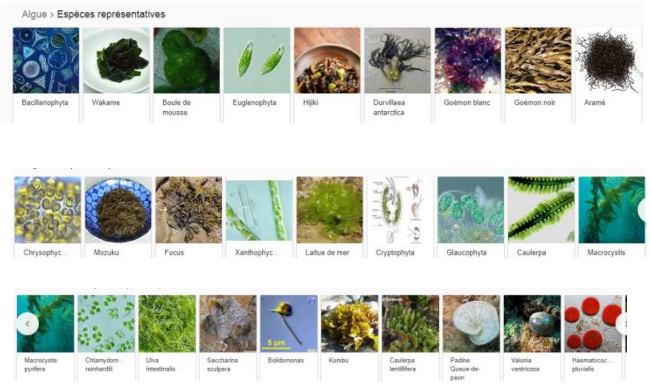
We used ChIP-seq profile peaks to study the transcription factor binding in three plants, including *Arabidopsis thaliana, Physcomitrella patens,* and *Chlamydomonas reinhardtii*.

Using experimental ChIP-seq data on transcription factor binding in plants, Morpheus tool

It is important to develop new software tools for the analysis of the transcription factor binding sites location, their clustering in a model genome, visualization, and statistical estimates for such clusters.

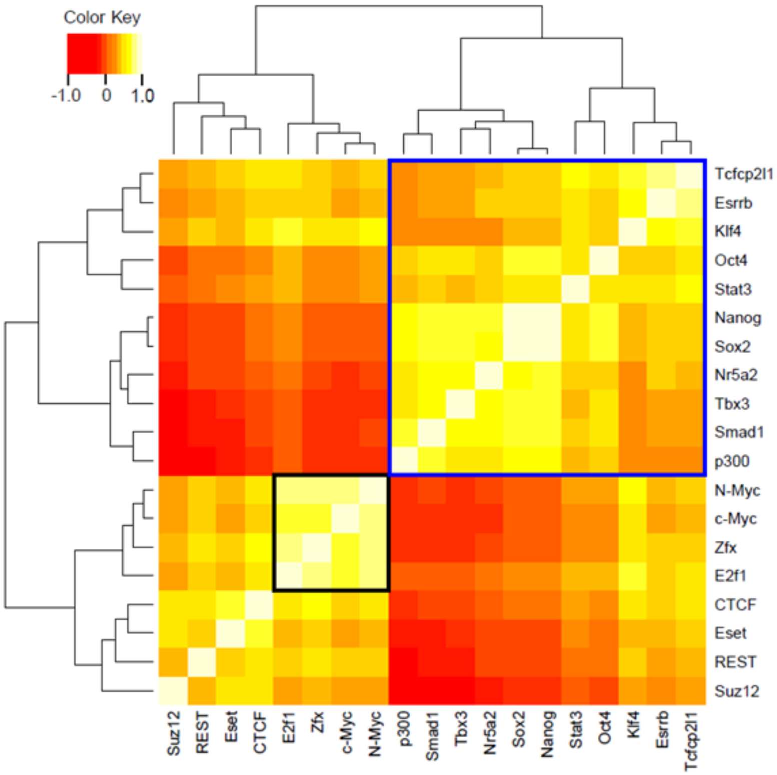
The existence of non-random clusters of the binding sites for 3 and more different factors identified by ChIP-seq was shown previously. Such clusters of sites could be used for gene promoter and enhancer prediction. This work presents a new application for the analysis of transcription factor binding sites in several evolutionarily distant model plant organisms.
We present the applications of computer scripts to analyze ChIP-seq data, description of clusters and visualization in the heatmaps format.





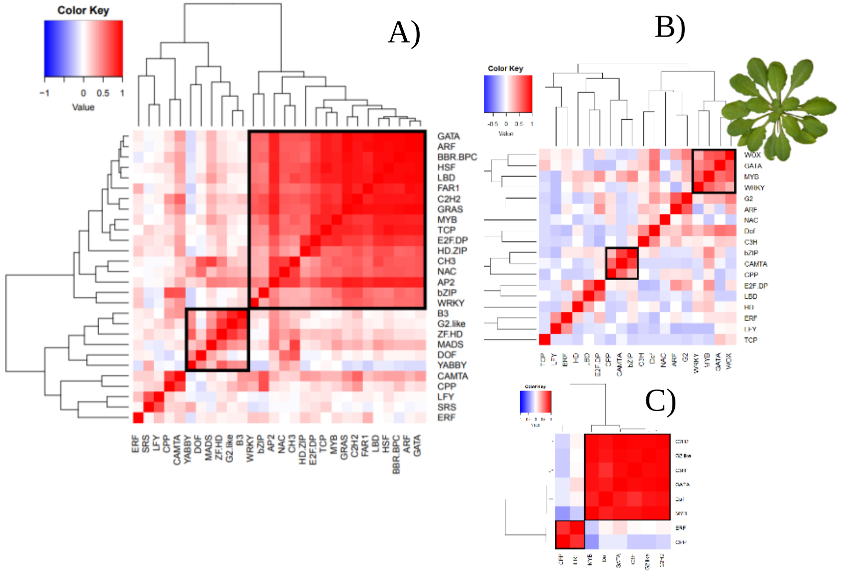
Integration of bioinformatics data

- **Heatmap for transcription factor bindings in mammalian genome**



19 transcription factors in mouse genome

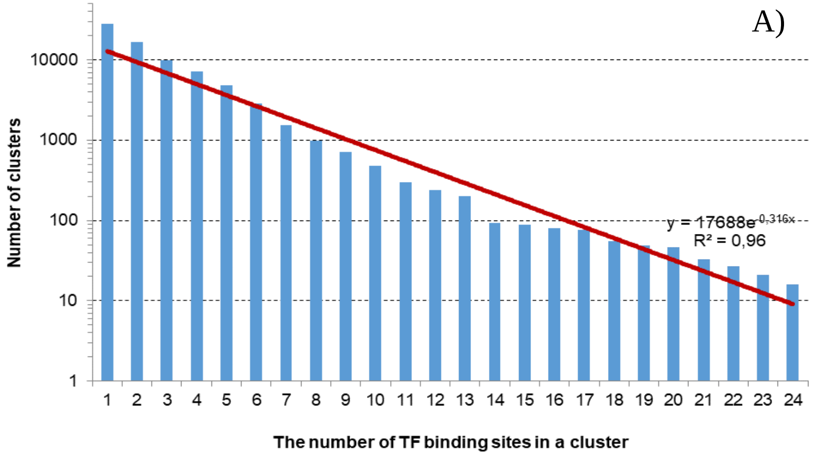
- **Heatmaps of transcription factor binding in the plant genomes**

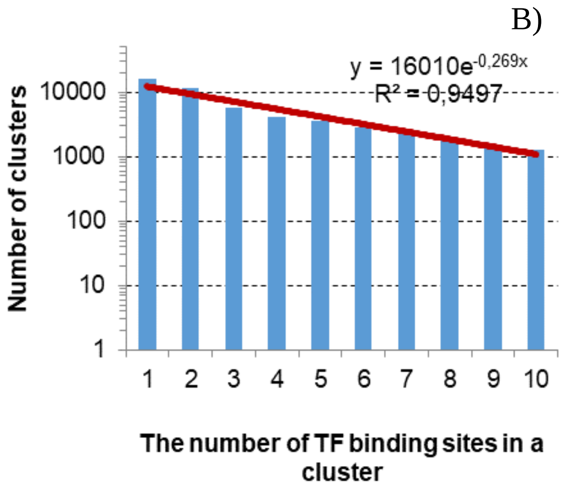


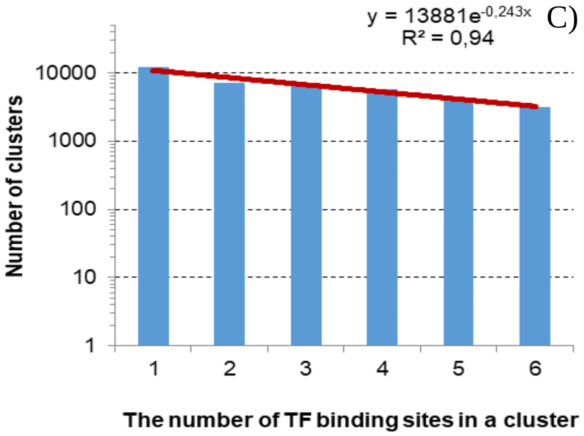
Heat maps of the joint localization of transcription factors for plants.
A - Arabidopsis Thaliana, B - Physcomitrella patens, C - Chlamydomonas Reinhardtii.
Red color means more frequent occurrence.
The squares highlight the groups of factors that tend to occur together most frequently.

The tool for transcription factor location analysis (using ChIP-seq data) is developed

Distribution of the number of different transcription factors in the plant genomes







Dependence of the number of clusters of binding sites on cluster size.
A - *Arabidopsis thaliana*, B - *Physcomitrella patens*, C - *Chlamydomonas reinhardtii*.
On the horizontal axis, the number of sites in the cluster, on the vertical axis, the number of clusters. The graphics are displayed with a logarithmic scale.

The distribution of the number of transcription factors in the genomes has the same shape (in logarithmic scale)