



HAL
open science

RSAT::Plants: Motif Discovery in ChIP-Seq Peaks of Plant Genomes

Jaime A. Castro-Mondragon, Claire Rioualen, Bruno Contreras-Moreira,
Jacques Van Helden

► **To cite this version:**

Jaime A. Castro-Mondragon, Claire Rioualen, Bruno Contreras-Moreira, Jacques Van Helden. RSAT::Plants: Motif Discovery in ChIP-Seq Peaks of Plant Genomes. Hehl, Reinhard. Plant Synthetic Promoters, 1482, Springer New York, pp.297–322, 2016, 978-1-4939-6394-2 978-1-4939-6396-6. 10.1007/978-1-4939-6396-6_19 . hal-01624368

HAL Id: hal-01624368

<https://amu.hal.science/hal-01624368>

Submitted on 1 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RSAT::Plants: Motif Discovery in ChIP-Seq Peaks of Plant Genomes

Jaime Castro-Mondragon (1, 2)*, Claire Rioualen(1, 2)*, Bruno Contreras-Moreira (3, 4) and Jacques van Helden (1, 2)

1 INSERM, U1090 TAGC, Marseille, F-13288, France.

2 AixMarseilleUniversity, U1090 TAGC, Marseille, F-13288, France.

3 Estación Experimental de Aula Dei-CSIC, Av. Montañana 1.005, 50059 Zaragoza, Spain.

4 Fundación ARAID, calle María de Luna 11, 50018 Zaragoza, Spain.

* The two first authors contributed equally to the manuscript.

Corresponding author: Jacques.van-Helden@univ-amu.fr

Running Head: Uncovering regulatory motifs in ChIP-seq peak sequences.

Summary/Abstract

In this protocol we explain how to run *ab initio* motif discovery in order to gather putative transcription factor binding motifs (*TFBMs*) from sets of genomic regions returned by ChIP-seq experiments. The protocol starts from a set of peak coordinates (genomic regions) which can be either downloaded from ChIP-seq databases, or produced by a peak-calling software tool. We provide a concise description of the successive steps to discover motifs, cluster the motifs returned by different motif discovery algorithms, and compare them with reference motif databases. The protocol is documented with detailed notes explaining the rationale underlying the choice of options. The interpretation of the results is illustrated with an example from the model plant *Arabidopsis thaliana*.

Key Words: Chromatin ImmunoPrecipitation DNA-Sequencing (ChIP-seq), transcription factor (TF), transcription factor binding motifs (TFBM), transcription factor binding site (TFBS), gene ontology (GO), functional enrichment

1. Introduction

1.1. The ChIP-seq Technology

The ChIP-seq method (1,2), which enables one to characterize transcription factor binding sites (TFBS) or chromatin marks in a whole genome, has gained a tremendous popularity to study genetic and epigenetic regulation. Although the main field of application so far has been Human and model organisms (Table 1), the ChIP-seq technology opens wide perspectives for the analysis of plant regulation.

Chromatin immunoprecipitation, followed by high-throughput sequencing and mapping on a reference genome, shows regions with high enrichment in reads. These regions, so-called *ChIP-seq peaks*, can be detected by using *peak-calling* algorithms. They typically encompass a few hundreds basepairs, and are centered on a binding site for the immunoprecipitated transcription factor (TF). They thus need to be further processed in order to discover transcription factor binding motifs (TFBM) and define the precise locations of the binding sites.

The characterization of TFBM from ChIP-seq experiments presents several advantages:

- 1 ChIP-seq peaks provide a relatively precise information about TF binding locations (~200bp precision). This makes a drastic difference with the approaches based on co-expression clusters (transcriptome arrays, RNA-seq), in particular for multicellular organisms (Metazoa, Plants), where regulatory regions can be found not only in the upstream promoter, but also in introns, downstream, and dispersed over wide distances.
- 2 The transition from ChIP-chip to ChIP-seq yet increased the precision of genome-wide location analyses.

- 3 Motifs discovered in ChIP-seq peaks are typically built from several hundreds or thousands of binding sites, and are thus much more robust than the previous-generation motifs built from a handful of sites that had been gathered one by one with Electrophoretic Mobility Shift Assays (EMSA) or footprint (low throughput) experiments.
- 4 Peak collections better reflect the *in vivo* diversity of binding sites for the TF of interest than *in vitro* methods such as Systematic Evolution of Ligands by EXponential Enrichment (SELEX).
- 5 Since peaks encompass a few hundred base pairs, they contain binding sites not only for the immunoprecipitated factor, but also for other interacting factors. *Ab initio* motif discovery thus enables us to detect additional motifs, and infer putative partners of the studied factor.

The knowledge gained from analyzing motifs and sites in ChIP-seq peaks may be used to enforce the design of synthetic promoters by predicting potentially important interactions between multiple TF (i.e. co-occurring motifs), synthetic promoters, and native promoters of the target species.

Since ChIP-seq peaks typically encompass several megabases or tens of megabases, specialized bioinformatics tools have been developed to discover motifs *ab initio* and scan the peaks for putative binding sites (3-6). In this chapter, we explain how to combine the motif discovery workflow *peak-motifs*(5,6) and some other tools of the Regulatory Sequence Analysis Tools (RSAT, <http://rsat.eu/>) (7) to discover and interpret TFBMs from plant ChIP-seq peaks.

1.2. Principle of the ChIP-seq Technology

The principle of Chromatin Immunoprecipitation sequencing (ChIP-seq) technology (1,2) is to cross-link a DNA-binding protein (TF, histone) with its bound DNA, shear the DNA by

ultrasonication, immunoprecipitate the protein of interest, release the cross-link, select DNA fragments of reasonable size (~300bp), and sequence their extremities (NGS sequencing is typically restricted to sequences smaller than the fragments). The primary result of a ChIP-seq experiment is a file with *raw short reads* (typically 36 to 75bp), which can be mapped onto a reference genome.

Fig. 1A shows the density profile of ChIP-seq reads for the transcription factor MYB3R3, mapped onto chromosomes 1 and 2 of the genome of *A. thaliana* (TAIR10 assembly version). This primary view of the data reveals a first difficulty for the interpretation of ChIP-seq data: some genomic regions are covered by a huge number of reads. These regions correspond to repetitive elements in centromeric and telomeric regions of the chromosomes. For the sake of comparison, Fig. 1B shows the density profile of a control experiment where the ChIP-seq protocol was run with an anti-GFP antibody, supposed to give an unspecific signal. This mock experiment reveals the same hyper-mapped regions, and can serve to estimate background and discard unspecific reads for the *peak-calling*. Note that mock experiments generally give reduced libraries. An alternative way to estimate unspecific background is to sequence genomic DNA without applying the immunoprecipitation procedure (*genomic input*).

1.3. Choice of a Peak-Caller and Tuning of its Parameters

One of the most crucial steps of the ChIP-seq analysis is the choice of a peak-calling program and the tuning of its parameters.

The *peak-calling* procedure consists in identifying genomic regions presenting a significant enrichment in reads in the ChIP-seq data, compared to some control set. The control set can

either be a mock experiment, as in Fig. 1B, or a full-genome sequencing. A large number of different programs exist for peak-calling(8,9).

Fig. 2 shows a detailed view of the peaks identified by some popular peak-callers on an arbitrary genomic region of the MYB3R3. Note the difference between the numbers and widths of the peaks, depending on the peak-calling tool. One of the most popular peak-calling programs, MACS, comes in two releases (10). The first version, MACS14, tends to return wide regions encompassing several topological peaks (compare the peaks with the MYB3R3 density profiles). MACS2, an upgraded version of MACS14, allows to specify parameters to obtain narrower peaks. Homer (11), based on the findPeaks algorithm, outputs very sharp peaks. The series of SWEMBL (12) peaks illustrates the impact of the parameters. This peak-caller proposes a "gradient" option (-R), which strongly affects the number of peaks and their width. SPP (13), using the FDR as a main parameter, is also to be carefully configured.

Most publications rely on the prior choice of a popular peak-caller, which is run with default parameters. Table 2 shows the wide range of peaks that can be found on a single dataset depending on the peak-calling algorithm and its configuration. However, the most appropriate algorithm and, even more, the fine-tuning of its parameters depend on the organism, data type, and even the purpose of the analysis (gathering high-confidence binding locations, identifying likely target genes, building a transcription factor binding motif, etc) (9). There is unfortunately no gold standard that would permit to assess the relative merits of peak-callers, and define their optimal parameters.

However, a variety of criteria can be used to evaluate the relevance of the returned peaks by various indirect indications, some of which will be illustrated in this protocol:

- enrichment of the reference motif (annotated motif for the immunoprecipitated factor) in the peak sequences (RSAT *matrix-quality*);
- concentration of the reference motif at peak centres;
- significance of the motifs discovered by *ab initio* approaches (RSAT *peak-motifs*);
- biological relevance of the transcription factors putatively bound to the discovered motifs (*FootprintDB* search);
- functional enrichment of the genes linked to the peaks (Gene ontology);
- concentration of the discovered motifs at the peak centers (RSAT *position-analysis*);

1.4. The Plant Regulatory Sequence Analysis Tools

Regulatory Sequence Analysis Tools (RSAT, <http://rsat.eu/>) is a specialized software suite for the analysis of cis-regulatory elements in genomic sequences (7). Since 2015, the services have been distributed on taxon-specific servers, including a Plant RSAT (<http://plants.rsat.eu/>). This address will redirect you to the host server <http://floresta.eead.csic.es/rsat>, which will be used for this protocol.

1.5. Functional Interpretation of ChIP-Seq Peaks

RSAT supports several approaches to interpret the peaks in functional terms:

- 1 **Motif enrichment.** In some cases, the immunoprecipitated factor is already known, and a reference motif exists in some database. It is generally a good practice to start by measuring the enrichment of the peak set for this reference motif, in order to check that the procedure went fine (from the wet lab to the bioinformatics workflow that produced the peaks).

2 **Motif discovery.** Several *ab initio* methods can be used to detect exceptional motifs in the peak sequences, based on different criteria: over-representation, biased positional distribution relative to the peak centers, etc.

1.6. Transcription factor binding motifs

Transcription Factor Binding Motifs (**TFBMs**) are generally represented as Position-Specific Scoring Matrices (PSSMs). They are built from an alignment of TF binding sites. Each cell of the matrix indicates the frequency of a given nucleotide (matrix rows) in a given column of the aligned sites (matrix columns). They can be depicted as sequence logos(**14**).

The widespread use of high-throughput technologies, for example ChIP-seq, allows to discover novel TFBMs or improve the quality of those existing (i.e. by increasing the number of sites to build the TFBMs). As more TFBMs are available, repertoires are required to give an easy access to these motifs. Currently there are many public and private motif databases, some of them specialized on few organisms (Athamap for *Arabidopsis thaliana*; Hocomoco for Human and Mouse, etc) and others have taxon-wide collections of TFBMs (Jaspar, TRANSFAC, CisBP) for plants, vertebrates, fungi, insects, etc. However, as these databases are growing, and since a single new study could produce an entire collection of motifs (**15**), efforts to collect, integrate and update many motif databases must be done. One option is FootprintDB(**16**) which is a meta-database encompassing 14 up-to-date motif databases (see chapter by Contreras-Moreira and Sebastián in this Volume).

In this protocol, we show how to run *ab initio* discovery on a set of ChIP-seq peak sequences, compare discovered motifs with a reference motif database, and cluster the discovered motifs to obtain a non-redundant collection.

2. Materials

2.1. Required Software

This protocol requires to dispose of

- a computer with any Web browser installed;
- a set of peak coordinates from a ChIP-seq or related experiment.

For visualization purposes (section 3.5), we also recommend to install the Integrative Genome Viewer (*17*).

2.2. Data sources

Peaks can be obtained either from NGS databases (*18,19*) or by running a peak-calling software tool on genome-mapped reads. This protocol starts from pre-computed peak coordinates, and does not cover the read mapping and peak calling procedures.

2.3. Data Formats

Peak coordinates should be provided in *bed* format (see the description of NGS file formats at the UCSC genome browser (*see Note 1*). Alternatively, this protocol can be run with peak sequences in *fasta* format (in which case the sequence retrieval steps can be skipped).

2.4. Study Case

As a study case we take a recent MYB3R3 study (*20*). We will use a BED file available at the Gene Expression Omnibus Database (GEO), under accession GSE60554 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60554>), which contains the results of a ChIP-seq experiment with the MYB3R3 transcription factor of *Arabidopsis thaliana*. The peaks

can be found at the bottom of the GEO Web page for the MYB3R3-ChIP-ped sample (GSM1482283, <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1482283>, peak file [GSM1482283_MYB3R3-GFP_ChIP_peaks.bed.txt.gz](#)) (see **Note 2**).

The reference motif for this case is that of c-Myb in tobacco (*21*), likely to be similar to MYB3R3 in *Arabidopsis thaliana*.

3. Methods

3.1. Retrieval of Peak Sequences from the Peak Coordinates

- 1 Obtain a bed-formatted list of peak coordinates (*see* **Note 3**).
- 2 Open a connection to the Plant Regulatory Sequence Analysis Tools server (<http://plants.rsat.eu/>).
- 3 On the left-side panel, open the toolbox "**Sequence tools**" and click "**sequences from bed/gff/vcf**".
- 4 Choose the appropriate genome in the **Organism** pop-up menu (*see* **Note 4**). For the study case, the reference organism is *Arabidopsis thaliana.TAIR10.29*, where the suffix *TAIR10* indicates the assembly, and the number *29* the EnsemblGenome version.
- 5 Enter the **Genomic coordinates** of your peaks (*see* **Note 5**). Coordinates can be entered in different ways: (i) directly pasted in the text area; (ii) large files can be uploaded from your computer to the server (option **Choose file**); (iii) enter the *URL of a coordinates file available on a Web server* (e.g. BEDfile on your account of a Galaxy server). For the study case you can enter the downloaded file [GSM1482283_MYB3R3-GFP_ChIP_peaks.bed.txt.gz](#).
- 6 Verify that the option **Mask repeats** is checked, as plant genomes are often repeat-rich (*see* **Note 6**).
- 7 For the **Output** option, choose *server*, and click **GO** to submit the job.
- 8 After a few seconds, a result page (shown in Fig. 3) should appear with the links to the FASTA file containing the peak sequences, plus some additional links to the inputBED file

and a log file. Note that the results are kept on the server for a restricted duration (72 hours).

If you want to keep track of the results, you can right-click on the fasta sequence file and download it to your computer.

At this stage of the protocol, you should have at your disposal a file containing peak sequences in fasta format. Typical peak sets include a few hundreds to tens of thousands of peaks, with lengths varying from tens to hundreds of base pairs each.

Note that the results page contains links to other RSAT tools. These enable you to transfer the obtained fasta file directly to the next step of the analysis.

3.2. *Ab Initio* Motif Discovery in ChIP-Seq Peak Sequences

We will now describe the way to discover motifs from ChIP-seq peak sequences. We obtained these sequences in the previous section, in the form of a fasta file, but it is also possible to upload your own fasta file from your computer directly in the *peak-motifs* section. We assume here that the sequences are transferred from the previous step.

- 1 At the bottom of the sequence retrieval result page, the **Next step** box presents a series of buttons to transfer the fasta sequences to another tool for further analyses (Fig. 3). Click on the *peak-motifs* button. This will display a new Web form shown in Fig. 4, pre-loaded with the URL of the peak sequences.
- 2 Before running *peak-motifs*, you are requested to type a **Title** for the job. For the study case, we can for example type "*A.thaliana MYB3R3 versus GFP - GSM1482283*".
- 3 The **Reduce peak sequences** frame allows you to trim the number and length of the peaks. By default all peaks are retained but those longer than 1Kb (500bp on either side of the peak

center) are shortened, because they are suspected to result from peak-calling artifacts rather than to represent trustable binding sites.

- 4 The **Motifs discovery** frame permits to choose the discovery algorithms and tune their parameters. By default only *oligo-analysis* and *position-analysis* are activated (*see***Note 7**).
- 5 Under **Motifs discovery** activate **oligomer lengths** 6 and 7 (*see***Note 8**).
- 6 Check that the **Markov order** is set to *automatic (adapted to sequence length)* (*see***Note 9**).
- 7 Check that the **Number of motifs per algorithm** is set to 5 (*see***Note 10**).
- 8 Under **Compare discovered motifs with databases**, you can select one or more motif collections in order to annotate any discovered motifs. For plant sequences we recommend *footprintDB-plants*, which integrates motifs from diverse public databases (see chapter by Contreras-Moreira and Sebastián in this Volume).
- 9 Optionally, the button below **Add your own motif database** allows you to upload a custom database of transcription factor binding motifs in a TRANSFAC-formatted file.
- 10 If there is a known motif for the immunoprecipitated factor, you can upload it with option **Add known reference motifs for this experiment** (*see***Note11**).
- 11 Click on the title **Locate motifs and export predicted sites**, check the option **Search putative binding sites in the peak sequences**, and activate the option **Peak coordinates specified in fasta headers in bedtools getfastaformat (also for retrieve-seq-bedoutput)**. Here we assume that the sequences were obtained from RSAT *retrieve-seq-bed* as indicated above (*see***Note 12**) but some alternative formats are also supported.

- 12 You can type in your **email address** to be notified of the job submission and completion, or you can choose **display**, and click **GO**.

After a few seconds, the server displays a confirmation of the job submission, with a link to the result Web page. Clicking on this link will open the result page on a separate tab of your Web browser. This page will be progressively updated to show the results of the analysis. A typical analysis should take from a few minutes to one hour, depending on the sequence size and the selected options (motif discovery algorithms, motif databases, sequence scanning).

- 13 Results will progressively be displayed on this page. Once the job is completed, a summary of all results will appear in a box at the top of the results page. After completion of the *peak-motifs* workflow, we recommend to **download the results** on your computer for further analyses, since they are kept on the server for a restricted time.

- a. Clicking on the link **Download all results**, in the header box of the result Web page, will allow you to save a zipped file containing the whole HTML report. You will thus be able to visualize these pages locally on your computer.
- b. Right-clicking on the link **Download all matrices (TRANSFAC format)** and saving it as *peak-motifs_motifs_discovered.tf* will allow you to keep a file containing all the motifs matrices. This file contains all discovered motifs, in the flat-file motif description format designed for the TRANSFAC database (this format is convenient because it allows to associate annotations to each motif). We will use it below in the section about matrix clustering (section 3.3).
- c. In the **Sequence composition (test sequences)** section, right-click on the link "[coordinates: UCSC BED track]" (right panel) and save the BED file as *peak-motifs_test_seqcoord.bed*. This file contains the peaks used for the peak-motifs analysis.

- d. At the bottom of the Web page, look for section **Motif locations (sites)**, then **Predicted sites on test peaks (all motifs)**. Right-click on the "[bed]" link to download the corresponding file *peak-motifs_all_motifs_seqcoord.bed*. This file can be loaded in a genome browser such as IGV (17).

3.2.1. Interpretation of the *Peak-Motifs* Results

The *peak-motifs* results are displayed in a Web form giving access to all the files generated during the analysis.

Fig. 5 shows a partial snapshot of the *peak-motifs* results with the study case. Since the workflow covers many types of analyses and results, here we attempted to present a human-readable report, organized according to the successive steps of the workflow: sequence composition (Fig. 5A), motif discovery (Fig. 5B), comparison of discovered motifs with known motifs (Fig. 5C).

Sequence composition

This section, described in Fig. 5A, shows some properties of the peak sequences.

- The top panel of the synthetic table shows the distribution of sequence lengths. In this study case, we can observe that most sequences have a length around 200 bp, which is a good indication for transcription factor ChIP-seq peaks (histone peaks are generally longer).
- The second panel shows the nucleotide composition of the sequences, with a heatmap indicating the frequencies of each nucleotide, and a plot displaying the profile of frequencies for each nucleotide along the peaks. In this example we can see that *G* and *C* are less frequent than *A* and *T* over the whole peak width. Interestingly, we also notice a

nucleotidiskew, with an enrichment of *As* and *Gs* upstream peak centers, and a symmetrical enrichment of *Ts* and *Cs* downstream.

- The third panel shows the dinucleotide composition of the sequences. The *transition table* indicates the probabilities of each nucleotide (column) depending on the preceding nucleotide ("prefix", rows). Gray shades denote the relative frequencies, and highlight dependencies between adjacent nucleotides. For example, in the study case, we observe that the frequency of *As* varies from 0.36 after another *A* (*AA* dinucleotide) to 0.21 after a *T* (*TA* dinucleotide). The *dinucleotide profiles* provide a visual representation of the positional distribution for each dinucleotide. On the study case we note an upstream-downstream skew for *AA*, *TT*, *CC* and *GG*, and a local depletion of *TA* and *AT* in the peak centers.

Discovered motifs (by algorithm)

This section (Fig. 5B) shows the full list of discovered motifs, organized by motif discovery algorithm (*oligo-analysis*, *position-analysis*) and by k-mer size.

The name of each motif (e.g. *oligos_6nt_mkv3_m1*) indicates:

- The algorithm used (*oligos* for *oligo-analysis*, *positions* for *position-analysis*).
- The k-mer length used to build the motif (6nt, 7nt).
- The order of the Markov model (mkv).
- The rank of the motif (m1 to m5).

In addition, the motif logo is displayed in both orientations.

In this section, an important information is that each discovered motif is associated with an e-value and a derived significance score: $sig = -\log_{10}(E\text{-value})$. The e-value indicates the expected

number of false positives. E-values much lower than 1 (corresponding to highly positive *sig* scores) indicate a very significant over-representation (oligo-analysis) or positional bias (position-analysis) of the motif. The most significant motifs are highlighted in red and bold. In our study case, the motif CACGTG is over-represented with a significance of 187, which corresponds to an e-value (expected number of false positives) of $\sim 10^{-187}$. The same motif is found by *position-analysis*, yet with a much lower significance ($s=3.34$, e-value 0.00046). It is thus the most significant motif in terms of over-representation, but other motifs are much more significant in terms of positional bias, in particular **rwttGGCGGGAaaat** (`positions_6nt_m1`), which achieves a significance of 34.61. This example shows the interest of combining two independent criteria to discover exceptional motifs.

Discovered motifs (with motif comparison)

Illustrated in Fig. 5C, this section displays each motif individually with matches found in collections of known TFBMs (e.g. FootprintDB plants, Jaspar plants, etc).

Additionally, for each motif, two other plots are shown:

- The positional distribution of predicted sites relative to peak centers (e.g. showing that most matches are located around the center of the peaks).
- The distribution of the number of binding sites per sequence. For the CACGTG motif, occurrences per peak show a particular teeth-shaped distribution due to the reverse complementary palindromic nature of the motif (occurrences are systematically found on both strands).

Note that the algorithms produce redundant motifs. For example a motif with the core GGCGGG is found by both *oligo-analysis* and *position-analysis*, with different k-mer lengths; thus, the next step in the analysis is to reduce the redundancy of the motifs.

3.3. Motif Clustering

Using different motif discovery algorithms to analyze the same sequences is useful and recommended to increase the sensitivity (some algorithms discover motifs that others do not) or to corroborate the results (e.g. gain confidence by observing that the same motif is both over-represented and concentrated on the peak centers). However in some cases the redundancy between motifs returned by different algorithms and with different parameters makes it difficult to interpret the results as a whole.

The RSAT web site includes a new specialized tool called *matrix-clustering*, which identifies groups of similar motifs, generates consensus matrices, and provides a dynamical visual interface to browse and inspect the relationships between multiple motifs. We will use this tool to obtain a non-redundant collection of motifs from the motifs discovered with *peak-motifs*.

- 1 Open a connection to the Plants Regulatory Sequence Analysis Tools server (<http://plants.rsat.eu/>).
- 2 On the left-side panel, open the toolbox "**Matrix tools**" and click "**matrix-clustering**".
- 3 On the *title* box you can give a title to the analysis for example *Myb3R3 discovered motifs*.
- 4 **Upload** the motif file obtained from *peak-motifs* and select the **TRANSFAC format**.
- 5 In the **Motif comparison options** section, you can fine-tune the thresholds that will be used to split the tree with all the motifs in a collection of trees (forest). The default cutoffs are

relatively lenient, but for this application more conservative values can be chosen. In the column *lower threshold*, set **w** to 5, **cor** to 0.75 and **Ncor** to 0.55.

- 6 In the **Clustering options** section, select *Ncor* (Normalized Pearson Correlation) as a **Metric to build the trees** and *average* as the **Agglomeration rule**.
- 7 You can either select **emailoutput** and fill up your address, or **display**, and click **GO**.

After a few seconds the Web site displays a link to the result page. You can already open this page as soon as the link appears. Even though the program may take a few minutes to accomplish the clustering, the result page will be updated periodically.

3.3.1. Interpretation of the *Matrix-Clustering* Results

The *matrix-clustering* results are organized in different sections (Fig. 6). You can display/hide each one by clicking on the buttons.

- The *Results summary* section shows a table indicating the number of input motifs, the number of clusters and the parameter used to cluster the motifs, additionally a link to download all the results in zip. In this case the 20 motifs discovered with *peak-motifs* were regrouped in 8 distinct clusters.
- The *Clusters summary* section shows a table with the motifs belonging to each cluster and the logos in both orientations representing the *root motifs* of each cluster (i.e. a motif formed by summing or averaging the counts of all the motifs belonging to the cluster).
- The *Logo Forest* section points to a link where the clusters are displayed as a set of trees, each corresponding to a cluster. In this link you can dynamically expand/collapse the tree, each time a branch is collapsed, it shows the *branch-motif* which represents all the

descendant motifs of the collapsed branch. Fig. 7 shows the first three clusters of the logo forest produced by *matrix-clustering* from the motifs discovered by *peak-motifs* in MYB3R3 peaks.

- The *Individual Motif View* section shows a table with all the input motifs and some of their attributes (assigned cluster, aligned and colored consensus, small logos).
- The *Individual Cluster View* section shows some properties of each cluster individually. You can select a specific numbered node of tree to select its corresponding *branch-motif*.
- The *Heatmap view* section shows a heatmap of the motifs grouped in clusters.
- The *Additional Files* section shows a table with additional files (motif comparison results, the motifs associated to each cluster, etc) including the *Root motifs* file, which contains the collection of non-redundant motifs. This file will be used for the following part of the analysis.

8 Right-click the "**Root motifs**" link and save file as *matrix-clustering_cluster_root_motifs.tf* on your computer.

The 20 motifs discovered with *peak-motifs* were separated in 8 clusters of variable size (Fig. 7).

For example, cluster 1 contains 8 motifs corresponding to the EF2 family while the motif for cluster 6 (singleton) corresponds to the MSA motif reported in the published work selected as our case study (15).

3.4. Negative controls with random genomic regions

RSAT motif discovery tools compute the significance of the motifs based on theoretical models (Markov chains, which take into account the dependencies between adjacent nucleotides).

However, it is not obvious *a priori* that these models perfectly suit the properties of biological sequences. A pragmatic way to check the correctness of the models is to measure the empirical rate of false positives with a *negative control set*, i.e. set of sequences supposedly not enriched for any particular TFBM. In principle, motif discovery programs should be able to return a negative answer (no result) when such datasets are submitted.

When analyzing genomic regions such as ChIP-seq peaks, the recommended negative control consists in analyzing regions of the same sizes as the peaks picked up at random in the reference genome.

- 1 Open a connection to the Plant Regulatory Sequence Analysis Tools server (<http://plants.rsat.eu/>).
- 2 In the left-side panel, open the toolbox "NGS ChIP-seq" and click "**random genome fragments**".
- 3 Under **Random fragments**, click the "**Browse...**" button and locate the peak sequences file on your computer (the fasta file downloaded at step 8 in section 3.1.).
- 4 Under **Organisms**, select the reference organism. For the study case, this is *Arabidopsis thaliana.TAIR10.29*.
- 5 In the **Output** section, select *Sequences in fasta format (only for RSAT organisms)* and check the *Mask repeats* option.
- 6 Select the *server* output and click **GO**. The selection of random genomic regions should take a few seconds.
- 7 On the result page, you can access the randomly picked up genomic sequences by clicking on the link to the fasta file (*Genomic fragments (fasta)*). You can optionally save this result to keep a copy of these random genomic fragments.

- 8 In the **Next Step** section of the result page, click on the **peak-motifs** button. This will display a *peak-motifs* form pre-filled with the URL of the random genomic sequences. Set the title to “A.thaliana random fragments”. Check that all the other parameters have the same parameters as for the analysis of the actual ChIP-seq peaks in the previous sections, and click **GO** (*see Note 13*).
- 9 Once the job is completed, open the results page, and click the link **Download all matrices (TRANSFAC format)** in the summary, to store the matrices on your computer.
- 10 Repeat the **matrix-clustering** analysis (steps 22 to 28) using the matrices obtained with *Random fragments* (TRANSFAC file).

3.4.1. Interpretation of the Negative Control

The goal of this negative control is to obtain an empirical estimation of the rate of false positives. In some cases, these controls reveal that the actual rate of false positive exceeds the theoretical expectation (indicated by the e-value of the motif discovery programs).

When the sequences of interests are genomic regions such as ChIP-seq peaks, the most relevant negative control consists on selecting random genomic regions of the same sizes. For the study case, we analyzed a dataset made of 2,931 random regions from *Arabidopsis thaliana*. The sequence length distribution is, as expected, exactly the same as for the actual peaks analyzed above. However the mono- and di-nucleotide composition may differ, because they reflect a random sampling of any type of genomic regions rather than regulatory regions.

Peak-motifs

The analysis of random genomic regions returned 17 motifs (Fig. 8A), most of which are of low complexity (e.g. atAAaATAaata, aaaAACAAAA, or motifs showing repeated sequences, e.g.

TATATATA). Some of these motifs show a high similarity with some reference motifs stored in FootprintDB, suggesting that they might correspond to some actual transcription factor.

The most important criterion in this control is to inspect the significance of the discovered motifs in the section **Discovered motifs (by algorithm)** (Fig. 8A). In our experience, programs based on a global over-representation (*oligo-analysis*, *dyad-analysis*) tend to return results even with random genomic regions, although with significance hopefully lower than with the real peaks: in the study case, *oligo-analysis* returns significance scores of 188 with the actual peaks, and 13.6 with random genomic regions. These motifs are actually correctly qualified of over-represented, but their over-representation is general in the genome rather than specific to the peaks. These motifs can correspond to low complexity regions or to functional elements found in abundance throughout the genome.

In contrast, programs relying on positional distributions (*position-analysis*, *local-word-analysis*) generally perform very well in negative controls (Fig. 8A), in the sense that they return motifs of poor significance (lower than 3) or no motif at all. This emphasizes once again the importance of evaluating multiple criteria before considering a motif as relevant.

In the section **Discovered motifs (with motif comparison)**, the positional distribution of predicted sites is not as concentrated around the centers of random fragments as they were for actual MYB3R peaks (Fig. 5C). Also, the number of matches is generally lower than the real peaks.

Matrix-clustering

With our random trial, the clustering separated the 17 significant motifs into 14 clusters (Fig. 8B,C), where only three clusters contain at least two motifs (the rest are singletons). This lack of

consistency between the discovered motifs is also an indication of the poorer relevance of the motifs discovered in random regions, relative to those found in actual peaks.

4. Notes

1. Format descriptions at UCSC: <https://genome.ucsc.edu/FAQ/FAQformat.html>

2. Direct access to the peak coordinates of the study case:

ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM1482nnn/GSM1482283/suppl/GSM1482283_MYB3R3-GFP_ChIP_peaks.bed.txt.gz

3. When working with lab data, peaks are obtained by running peak-calling programs on the aligned reads. Alternatively they can be downloaded from specialized databases such as GEO (18, <http://www.ncbi.nlm.nih.gov/geo/>) or ArrayExpress (19, <https://www.ebi.ac.uk/arrayexpress/>).

4. It is very important to specify the same assembly as used for the read mapping, since otherwise the coordinates on the BED file might not match the correct genomic sequences. Please contact the administrator of the RSAT Plant site if the required assembly does not appear in the list.

5. A common difficulty with BED files is that the chromosome naming convention differs between genome databases. In particular, some databases systematically use a "chr" prefix (chr1, chr2, chr3, ..., chrMt, chrPt) whereas some others simply use the chromosome number (1, 2, 3, ...) or name (Mt, Pt). To circumvent this problem, the sequence retrieval tool automatically checks the consistency of chromosome names between the query BED file and the genome sequence file installed on RSAT, and prepends or removes the chr prefix if required.

6. In plant genomes, repeated elements may result from various sources: transposons, polyploidy, etc (see chapter by Contreras-Moreira, Castro-Mondragon et al. in this Volume). Repetitive elements cause particular problems for motif discovery, because the statistics of over-representation rely on an assumption of independence between the sequences. It is thus

recommended to mask repeated elements during the motif discovery step of a ChIP-seq analysis workflow. Note that in some other contexts (for example scanning sequences with a TF binding motif), it might be relevant to keep the repetitive elements in order to detect all the putative binding sites.

7. Two other algorithms can be selected for finding motifs: *dyad-analysis* detects over-represented dyads (spaced pairs of trinucleotides), which are typically bound by dimeric transcription factors; *local-words* detects k-mers with local overrepresentation, i.e. having a higher number of occurrences in a particular positional window, relative to the rest of the peaks. Selecting more algorithms is sometimes helpful to gather a wider set of discovered motifs, as some algorithms can discover motifs that other would not. However, in many cases the different algorithms return very similar motifs, thus producing redundancy in the result. We thus activated by default the two algorithms offering a good trade-off between computing time and sensitivity, and which rely on two complementary criteria (over-representation and positional distribution relative to peak centers).

8. Beware, oligomer-length is not the same as motif length. Indeed, the significant k-mers and dyads are assembled and used as seeds to collect sites, which are in turn aligned to build the final motifs (position-specific scoring matrices). The resulting matrices are thus generally wider than the oligomer length. The default lengths were chosen because they generally provide a good tradeoff between sensitivity and specificity, and were shown to return the most relevant motifs (22).

9. The program *oligo-analysis* relies on Markov models to compute the prior probability of each k-mer, i.e. its probability to be found at a given position in the sequence. In peak-motifs, the prior probability of each oligonucleotide (k-mer) is estimated on the basis of the frequencies of smaller

k-mers in the sequence. The Markov order specifies the stringency of the background model. Increasing the order improves the specificity at the cost of sensitivity. This automatic option applies an ad-hoc rule to choose a Markov order ensuring a balance between sensitivity and specificity, depending on the total size of the peak set.

10. By default the program restricts the results to 5 motifs (assembled matrices) per algorithm. This number could be increased if you have some particular reason to think that the peak set contains a wider variety of motifs, with a proportional increase in the computing time. This can be useful for example for peaks from particular histone modification marks corresponding to enhancer regions supposedly bound by multiple factors.

11. Beware, there is a distinction between the options *reference motifs* and *custom database*. Reference motifs should be one or a few motifs expected to be found in the ChIP-seq peaks, whereas the custom database may be a large collection encompassing all the known motifs for the organism or taxon of interest.

12. By default, sequence scanning returns the putative binding site coordinates relative to the peak sequences. If appropriately formatted, the sequence headers of the peak file can indicate the coordinates of each peak relative to the chromosomes. The program can then convert each binding site coordinate from peak-relative to chromosome coordinates. The resulting files can then be loaded in a genome viewer (e.g. IGV).

13. The *peak-motifs* analysis will take approximately the same time as for the actual peaks, between a few minutes and several tens of minutes depending on the sequence size.

14. Example of structured query to gather ChIP-seq series (GSE) for a given taxon in GEO datasets (<http://www.ncbi.nlm.nih.gov/gds/>): ("gse"[Entry Type] AND "genome

binding/occupancy profiling by high throughput sequencing"[DataSet Type] AND
"Viridiplantae"[Organism])

5. References

- 1 Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4:651–657
- 2 Mardis ER (2007) ChIP-seq: welcome to the new frontier. *Nat Methods* 4:613–614
- 3 Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* 26:2622–2623
- 4 P Machanick, TL Bailey (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27:1696-1697
- 5 Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat Protoc* 7:1551–1568
- 6 Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res* 40:e31
- 7 Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM, Contreras-Moreira B, Artufel M, Charbonnier-Khamvongsa L, Hernandez C, Thieffry D, Thomas-Chollier M, van Helden J (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res* 43:W50–56
- 8 Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 6:S22–32

- 9 Steinhauser S, Kurzawa N, Eils R, Herrmann C (2016) A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief Bioinform* doi: 10.1093/bib/bbv110
- 10 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;9(9):R137. doi: 10.1186/gb-2008-9-9-r137
- 11 Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* 38(4):576-589
- 12 Wilder S (2009) SWEMBL: a generic peak-calling program. Unpublished.
<http://www.ebi.ac.uk/~swilder/SWEMBL/>
- 13 Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 26:1351 - 1359
doi:10.1038/nbt.1508
- 14 Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18:6097–6100
- 15 Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J (2013) DNA-binding specificities of human transcription factors. *Cell* 152:327–339

- 16 Sebastian A, Contreras-Moreira B (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics* 30:258–265
- 17 Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14:178–192
- 18 Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41:D991–995
- 19 Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A (2015) ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 43:D1113–1116
- 20 Kobayashi K, Suzuki T, Iwata E, et al. (2015) Transcriptional repression by MYB3R proteins regulates plant organ growth. *EMBO J* 34:1992–2007
- 21 Ito M, Araki S, Matsunaga S, Itoh T, Nishihama R, Machida Y, Doonan JH, Watanabe A (2001) G2/M-phase-specific transcription during the plant cell cycle is mediated by c-Myb-like transcription factors. *Plant Cell* 13:1891–1905
- 22 van Helden J, André B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 281:827–842

Acknowledgements

We thank C. Dubos for feedback on MYBR3 proteins. This work was funded in part by Fundación ARAID and by the Enseignants-Chercheursinvités program of Aix-Marseille Université (to BCM). CR was supported by the *France Génomique* National infrastructure, funded as part of the *Investissementsd’Avenir*, program managed by the *AgenceNationale pour la Recherche*(contract ANR-10-INBS-09).

Figure captions

Figure 1. Density profiles of reads mapped on chromosomes 1 and 2.

A. MYB3R3-bound immunoprecipitated chromatin. Reads were mapped on the TAIR10 assembly of *Arabidopsis thaliana* genome.

B. Control experiment (mock with anti-GFP antibody). Reads from the control experiment are used as "input" for the peak-calling, which enables peak-callers to avoid reporting peaks in the repetitive regions.

In both, ChIP and control tracks, note the striking concentration of reads in particular genomic locations, corresponding to repetitive regions. The map was generated with the Interactive Genome Viewer (*12*).

Figure 2. Peak profiles obtained with a variety of peak-calling algorithms and parameters.

Zoom of the reads and peaks in an illustrative region of chromosome 1 (coordinates 570,625 to 580,625). Each peak-caller is denoted by a specific color: MACS (pink) (*10*), Homer (green) (*11*), SWEMBL (orange) (*12*) and SPP (cyan) (*13*). Two peaks are detected by most peak-callers, although with different widths. One of these peaks is located in a gene promoter (at 572 kb), and another one within an intron (576 kb). The sensitivity of each peak-caller can be tuned with some specific parameters, as illustrated with the SWEMBL series (sensitivity increases from top to bottom) or SPP (false discovery rate set to 0.001 or 0.01, resp.). Relatively stringent settings are recommended to obtain a good tradeoff between sensitivity and relevance of the peaks.

Figure 3. Results of the sequences retrieval procedure.

View of the result page from the sequence retrieval step, made using aBED file (15) and the tool “sequences from bed/gff/vcf”. Next analysis steps can be processed with by simply clicking the corresponding buttons.

Figure 4. View of the *peak-motifs* form.

Two fields are required in order to proceed with the analysis: “title” and “peak sequence”. Here, the sequence file was automatically uploaded from the previous step.

Figure 5. Peak-motifs results.

A. General information about the peak sequences of our study case, including their composition in nucleotides and dinucleotides, and the corresponding profiles.

B. Discovered motifs (by algorithm). This example shows the 6-nucleotide motifs found with the *oligo-analysis* algorithm, using a Markov model of order 3.

C. Discovered motifs (with motif comparison). Shows the comparison of the discovered motifs versus a collection of TFBMs databases, and the distribution profile of the motifs in the peaks.

Figure 6. Matrix-clustering results.

General view of matrix-clustering report. Each button can be clicked to show/hide details.

Figure 7. Matrix-clustering results.

The 20 motifs discovered by peak-motifs in the MYR3R3 ChIP-seq peaks were separated in eight clusters. Each tree shows the alignments of a cluster of similar motifs. The leaves indicate the motif discovery algorithm with which each motif was found. Note that the similar motifs are

discovered independently by different algorithms (*oligo-analysis*, *position-analysis*), or are found with different parameters (e.g. k-mer length) of the same algorithm.

Figure 8. Negative controls with random genomic regions.

A. Partial results of the *peak-motifs* motif discovery result in random genomic regions. Note that the most significant motifs are poor-complexity motifs corresponding to repetitive elements.

B. Overview of the *matrix-clustering* results for these motifs. Note the high number of clusters, indicating that most motifs are detected by only one motif discovery method.

C. Clustering of the motifs discovered in the random peaks.

Figure 9. Heatmap of mutual coverage of peak-calling results.

The second column indicates the number of peaks depending on the peak-calling program and the main parameters affecting the stringency of the result. Further columns indicate the proportion of peaks of one peak-calling result (row) covered by peaks of another peak-calling result (columns).

Table captions

Table 1. ChIP-seq samples per taxa.

Number of ChIP-seq samples available in the Gene Expression Omnibus database (**13**) (Dec 18, 2015) per taxonomic group (*see* **Note 14**).

Table 2. Contingency tables comparing peak-caller results.

Each cell indicates the number of peaks of one peak-calling result covered by peaks of another peak-calling result. The diagonal indicates the number of peaks detected by each one of them.

See also Fig. 9 for a heatmap of the relative frequencies.





Tables

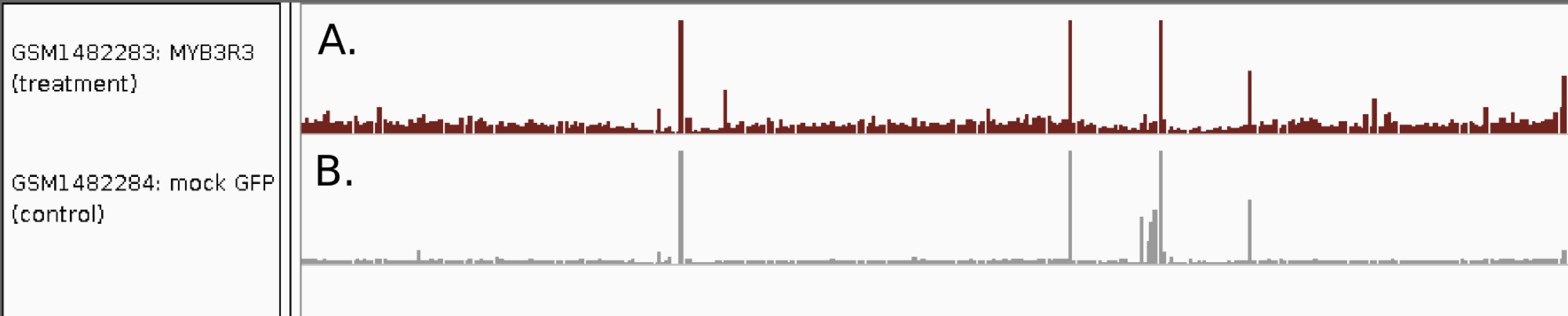
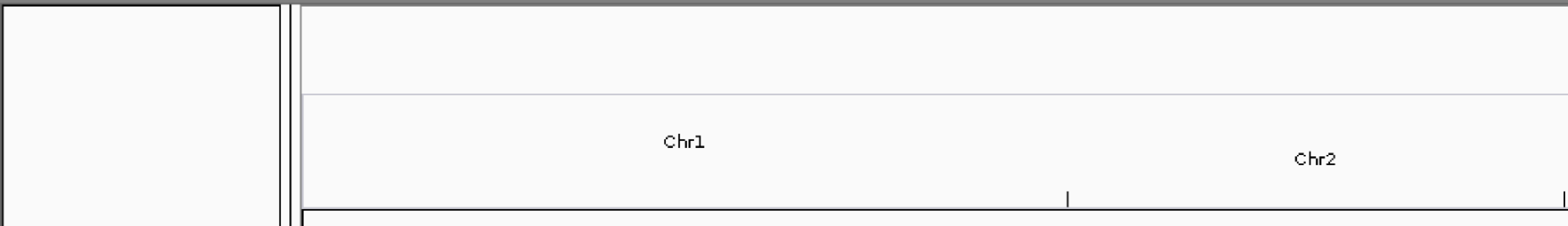
Table 1

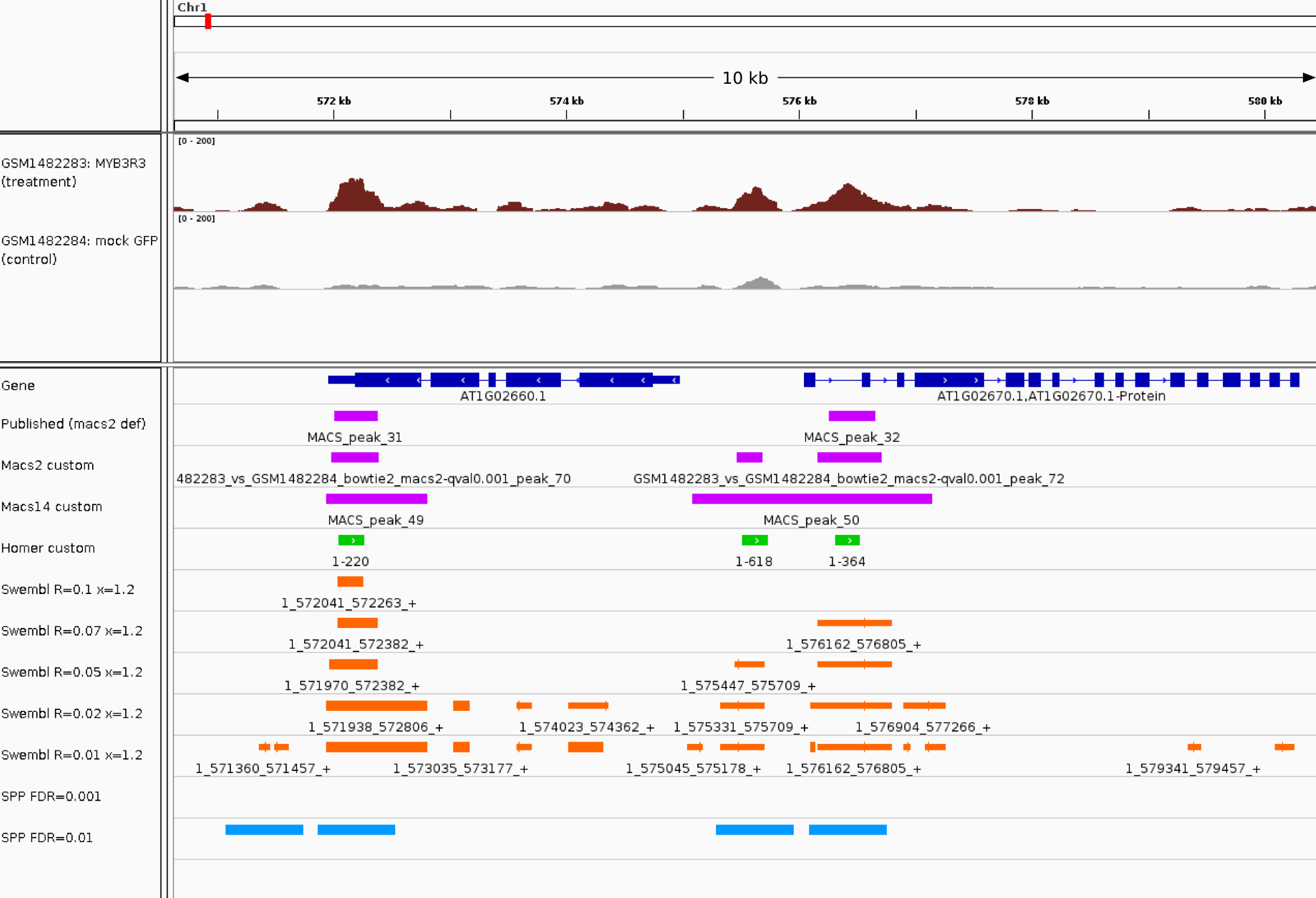
Taxon	GEO ChIP-seq series
<i>No taxon specified (any taxon)</i>	4722
Metazoa	4255
Homo sapiens	1542
Musmusculus	1793
Caenorhabditiselegans	410
Drosophila melanogaster	542
Fungi	238
Saccharomyces cerevisiae	163
Viridiplantae	157
Bacteria	64
Escherichia coli	24
Alveolata	14
Archaea	1

Table 2

Peak-caller	Macs2 (qval=0.05)	Macs2 (qval=0.001)	Macs14 (pval=0.00001)	Homer (fdr=0.01)	SPP (fdr=0.01)	SPP (fdr=0.001)	SWEMBL (R=0.1)	SWEMBL (R=0.07)	SWEMBL (R=0.05)	SWEMBL (R=0.02)	SWEMBL (R=0.01)
Macs2 (qval=0.05)	2931	3335	2699	2854	3408	532	1298	2224	2704	2848	3263
Macs2 (qval=0.001)	2930	9711	5494	8136	9576	544	1340	2767	5020	8840	11973
Macs14 (pval=0.00001)	2895	7360	6242	7359	9767	535	1325	2659	4510	9518	17225
Homer (fdr=0.01)	2851	8884	6114	18812	16898	534	1328	2743	5091	17125	24503
SPP (fdr=0.01)	2920	9291	6166	15364	24781	544	1333	2751	5104	20399	39018
SPP (fdr=0.001)	534	640	533	534	680	544	532	533	536	557	654
SWEMBL (R=0.1)	1374	1764	1294	1377	1786	534	1352	1343	1340	1368	1502
SWEMBL (R=0.07)	2355	3606	2561	2861	3679	535	1352	2788	2765	2864	3424
SWEMBL (R=0.05)	2852	6302	4224	5316	6697	538	1352	2787	5256	5541	7277
SWEMBL (R=0.02)	2931	9692	6236	16734	20518	544	1352	2788	5256	31867	41904
SWEMBL (R=0.01)	2931	9710	6242	18611	24365	544	1352	2788	5256	31864	92695

A. thaliana (TAIR 10) All Go        |





Result

Result file(s)

Content	URL
Coordinate file (bed)	http://floresta.eead.csic.es/rsat/tmp/apache/2016/02/22/retrieve-seq-bed_2016-02-22.110814_7yvS2g_coordinates.txt
Genome fragment lengths	http://floresta.eead.csic.es/rsat/tmp/apache/2016/02/22/retrieve-seq-bed_2016-02-22.110814_7yvS2g_lengths.tab
Result sequences (fasta)	http://floresta.eead.csic.es/rsat/tmp/apache/2016/02/22/retrieve-seq-bed_2016-02-22.110814_7yvS2g.fasta
Command log (text)	http://floresta.eead.csic.es/rsat/tmp/apache/2016/02/22/retrieve-seq-bed_2016-02-22.110814_7yvS2g_log.txt
Error log (text)	http://floresta.eead.csic.es/rsat/tmp/apache/2016/02/22/retrieve-seq-bed_2016-02-22.110814_7yvS2g_error_log.txt

next step

Motif Discovery (*ab initio*)

oligo-analysis

Over- or under-represented words

info-gibbs

Gibbs sampling
(DeFrance, 2009)

Pattern matching (known patterns)

dna-pattern

Regular expressions and IUPAC search.

Utilities

convert sequence

Format inter-conversions + mask short fragments.

dyad analysis

Overrepresented spaced pairs

peak-motifs

Full work flow
for chip-seq peaks
and other seq types

matrix-scan-quick (matrices)

Position-specific scoring matrices

purge sequence

Mask redundant fragments.

position analysis

Positionally biased words

matrix-scan (full options)

Position-specific scoring matrices

local-word-analysis

Windows of word over-representation

RSAT - peak-motifs

Pipeline for discovering motifs in massive ChIP-seq peak sequences.

References

1. Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D. and van Helden, J. (2011). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets Nucleic Acids Research doi:10.1093/nar/gkr1104, 9. [[Open access](#)]
2. Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J. (2012). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat Protoc 7(8): 1551-1568. [[PMID 22836136](#)]

► Information on the methods used in peak-motifs

Peak Sequences
Title (mandatory)
Peak sequences (mandatory) Paste your sequence (fasta format)

Or select a file to upload (.gz compressed files supported)
 No file chosen
URL of a sequence file available on a Web server (e.g. Galaxy).

Mask
[\(I only have coordinates in a BED file, how to get sequences ?\)](#)

Optional: control dataset for differential analysis (test vs control)
Control sequences Paste your sequence (fasta format)

Or select a file to upload (.gz compressed files supported)
 No file chosen
URL of a sequence file available on a Web server (e.g. Galaxy).

Mask

► Reduce peak sequences

► Motif discovery parameters

► Compare discovered motifs with databases (e.g. against Jaspar) or custom reference motifs

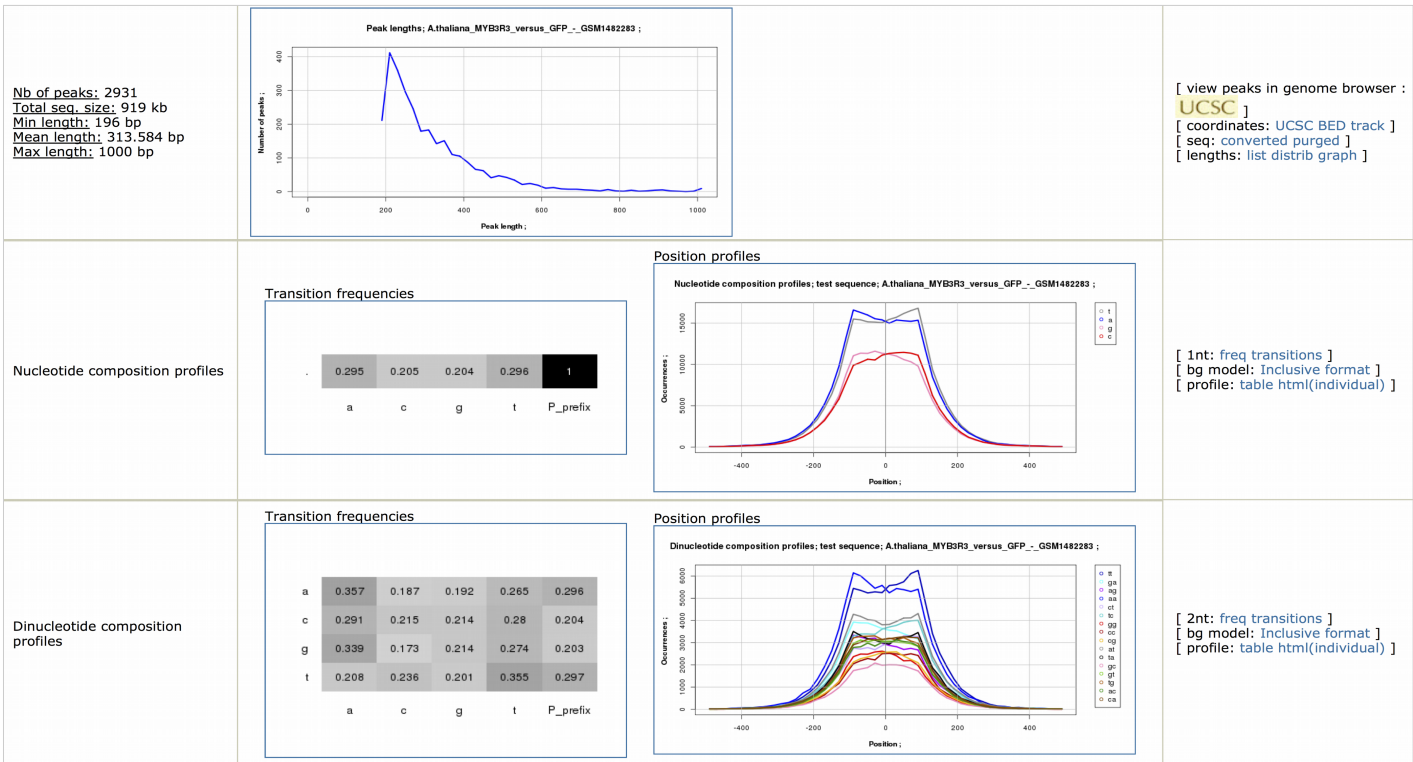
► Locate motifs and export predicted sites as custom UCSC tracks

► Reporting options

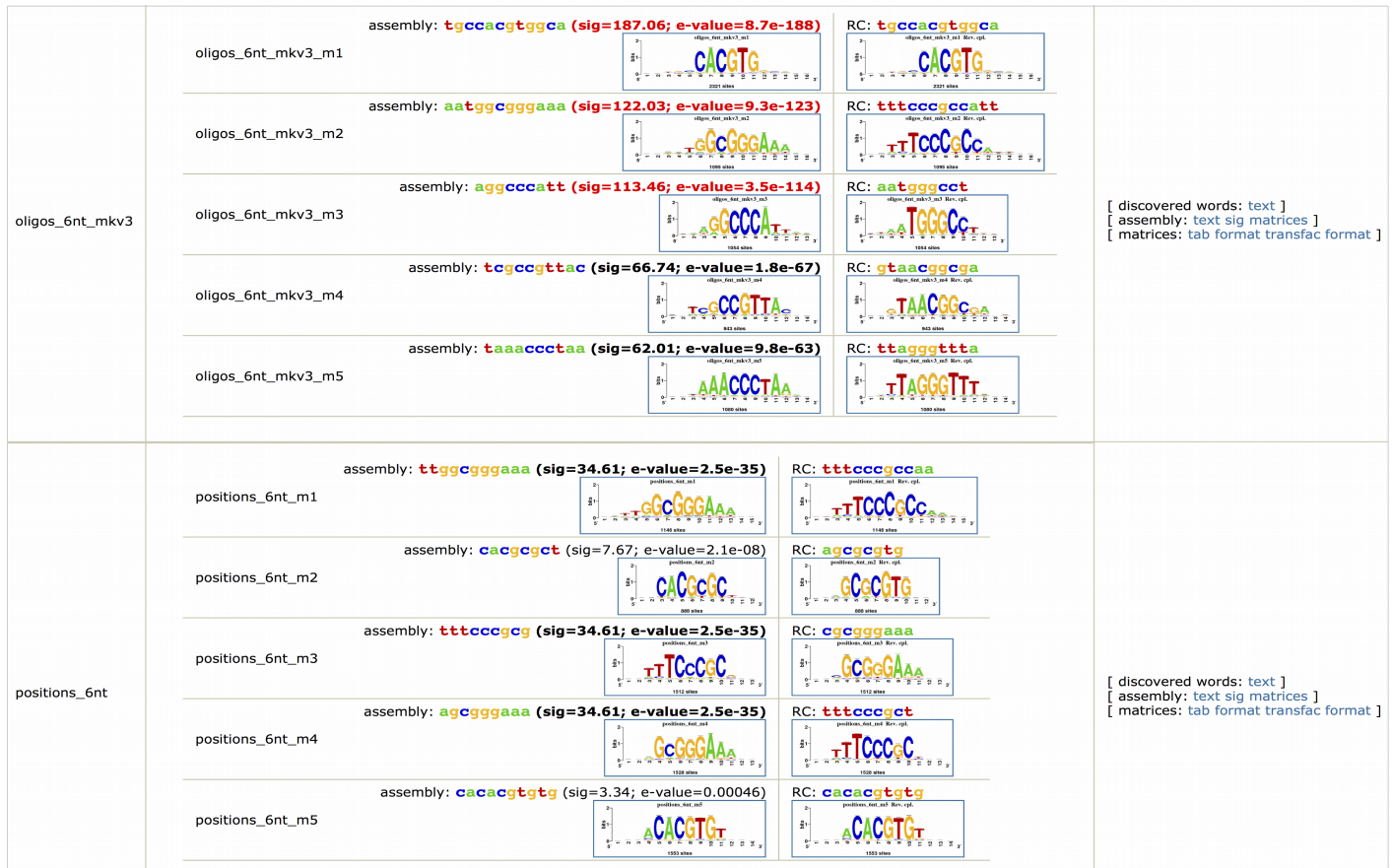
Output display email

Note: email output is preferred for very large datasets or many comparisons with motifs collections

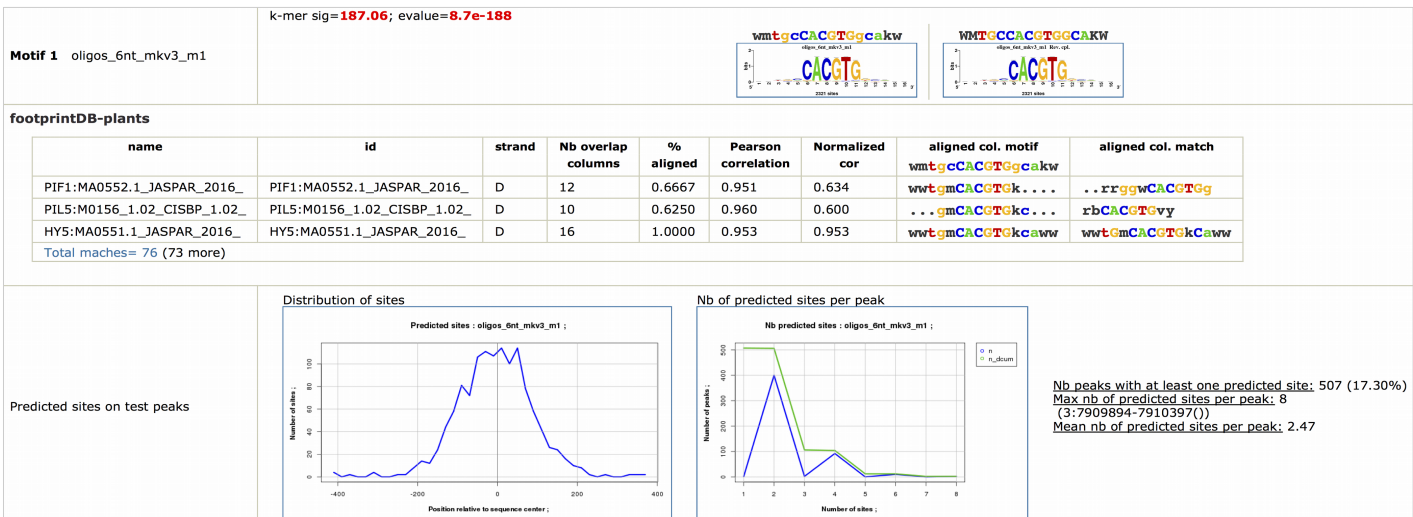
A



B



C



Results Summary

Results Summary

Nb Input motifs	Nb Clusters Found	Linkage method ?	Similarity metric ?	Thresholds to partition the tree ?	Complete results [zip]
20	8	average	Ncor	Ncor = 0.55 cor = 0.75 w = 5	Download

[Clusters Summary](#)

[Logo Forest](#)

[Individual Motif View](#)

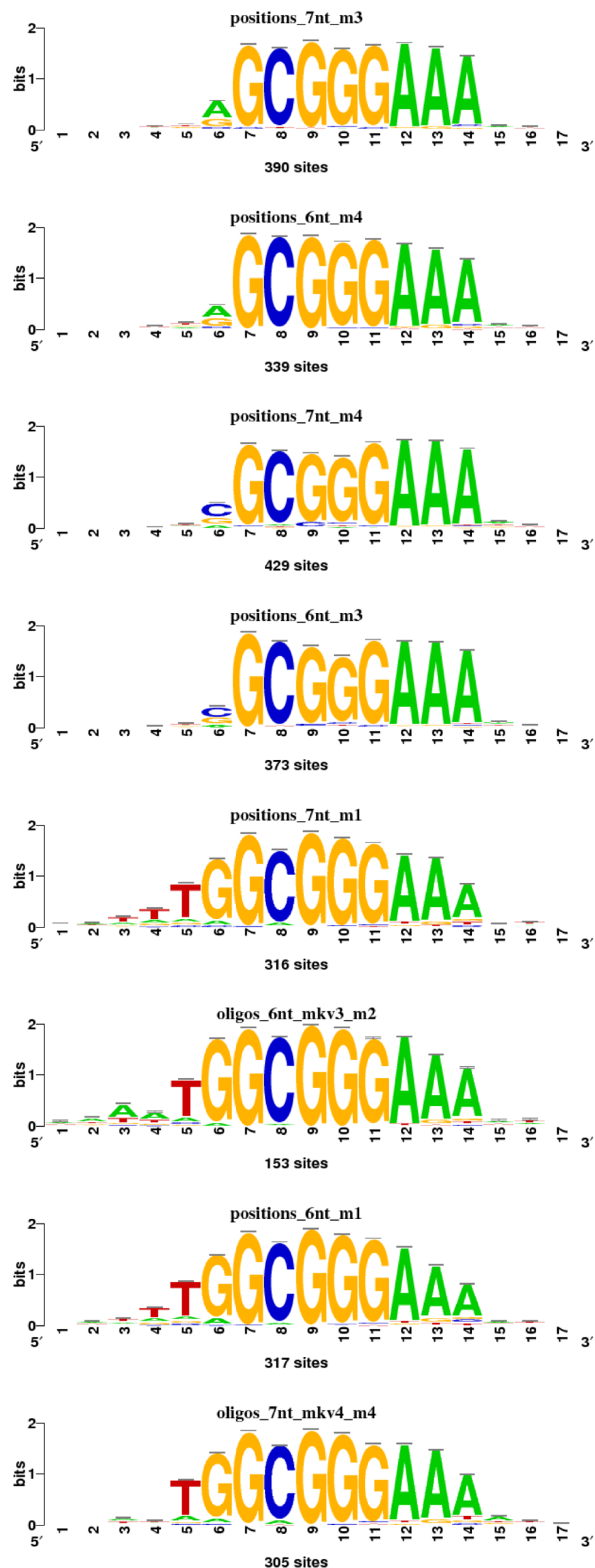
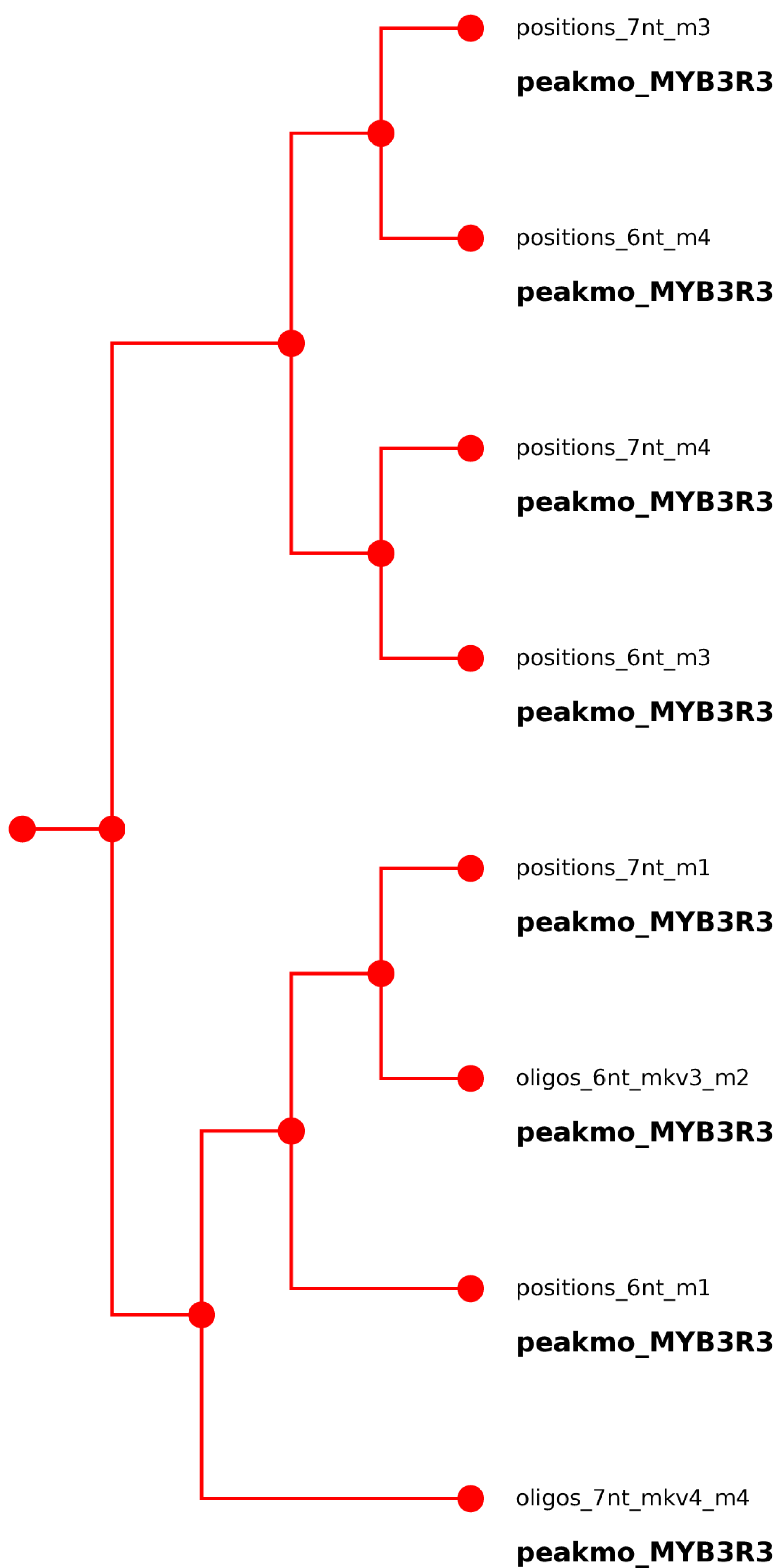
[Individual Cluster View](#)

[Heatmap View](#)

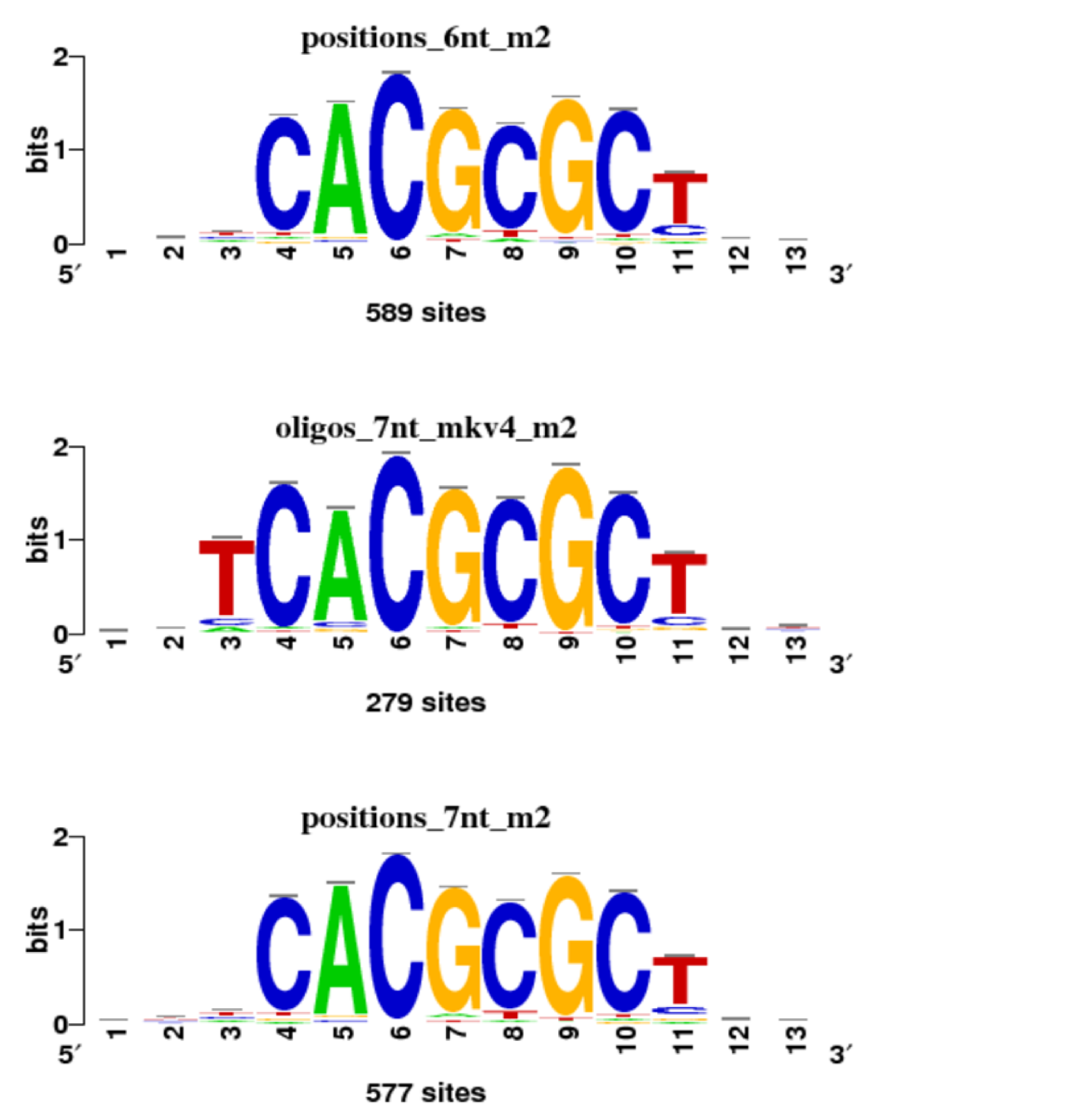
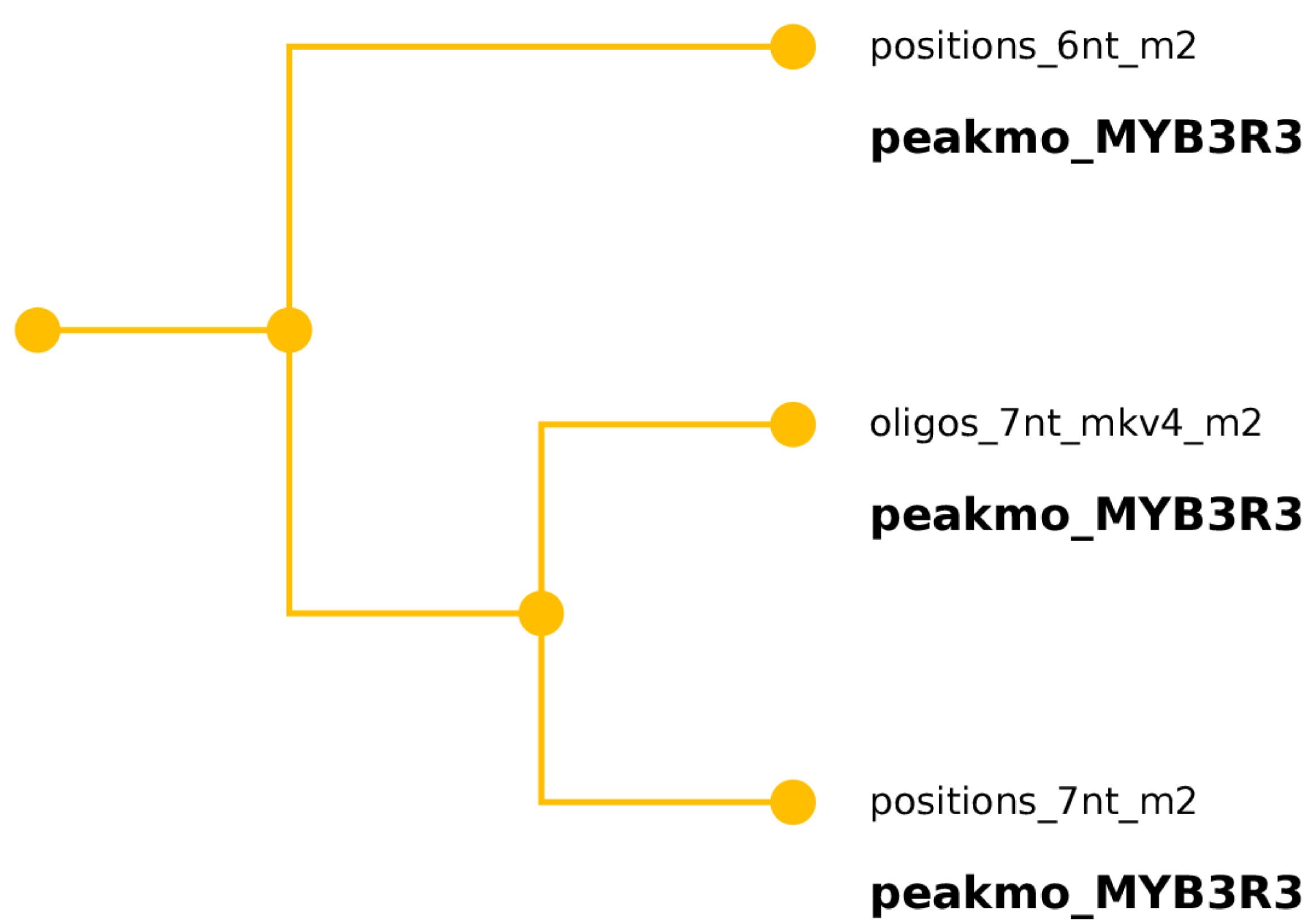
[Additional Files](#)

[References](#)

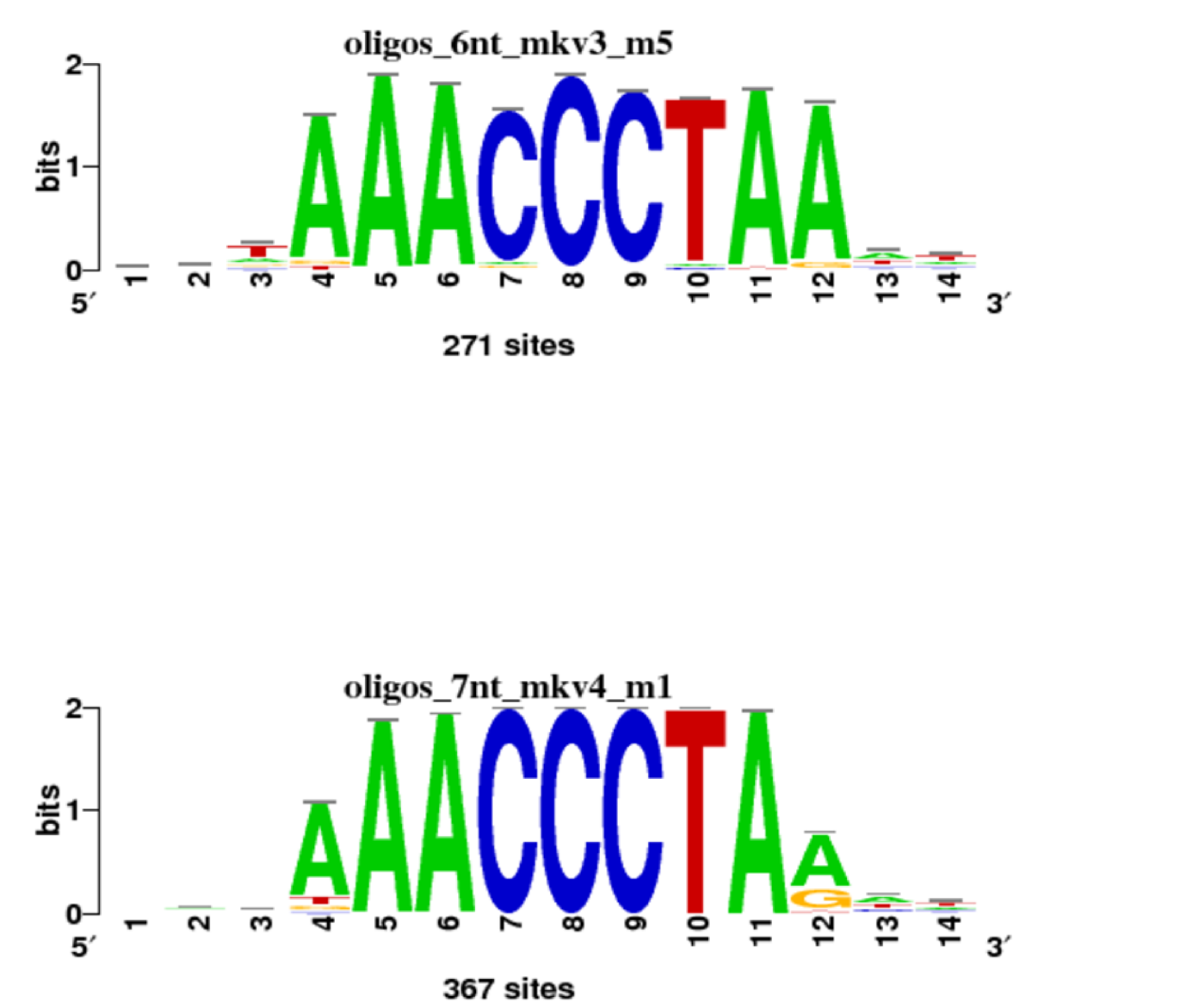
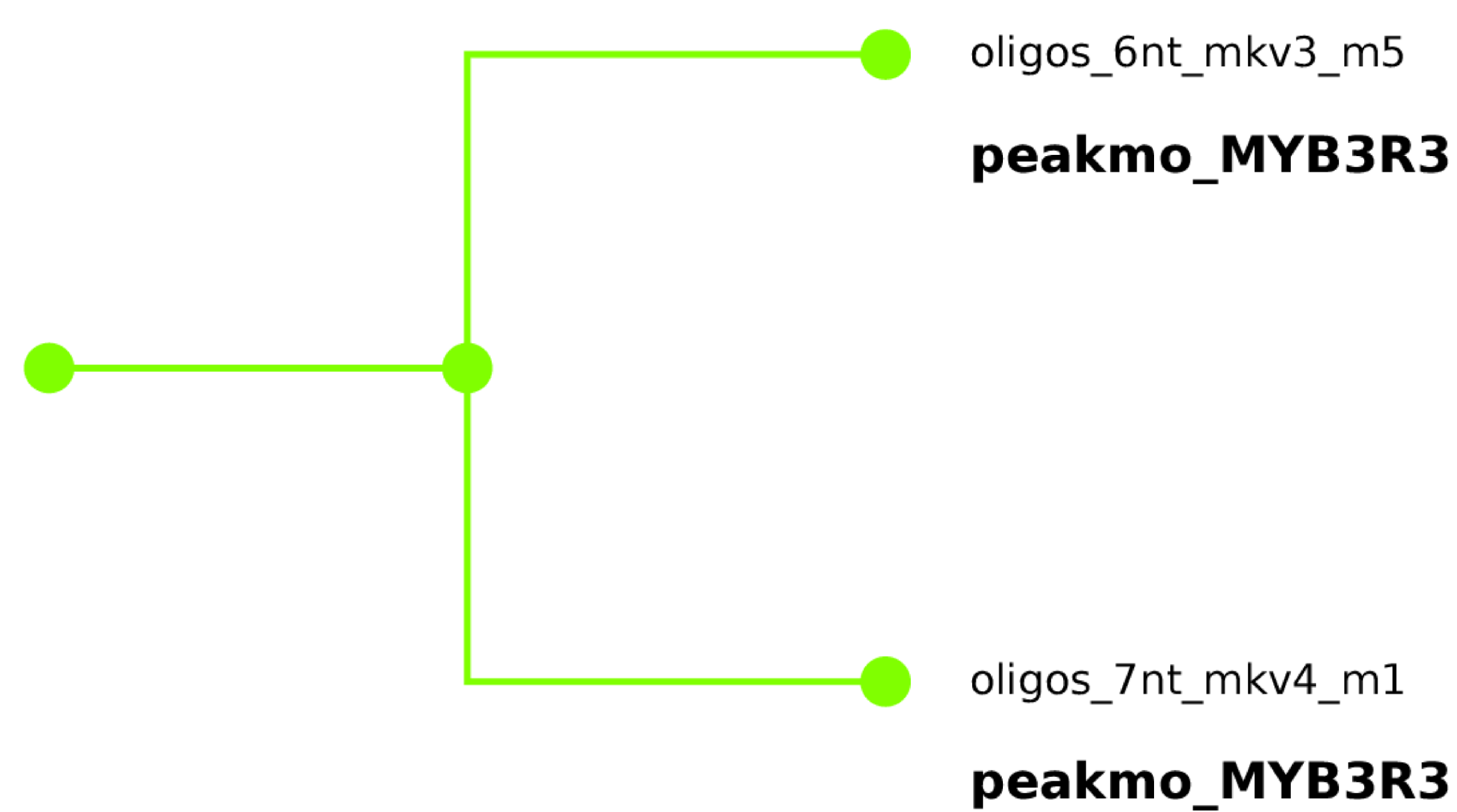
cluster_1



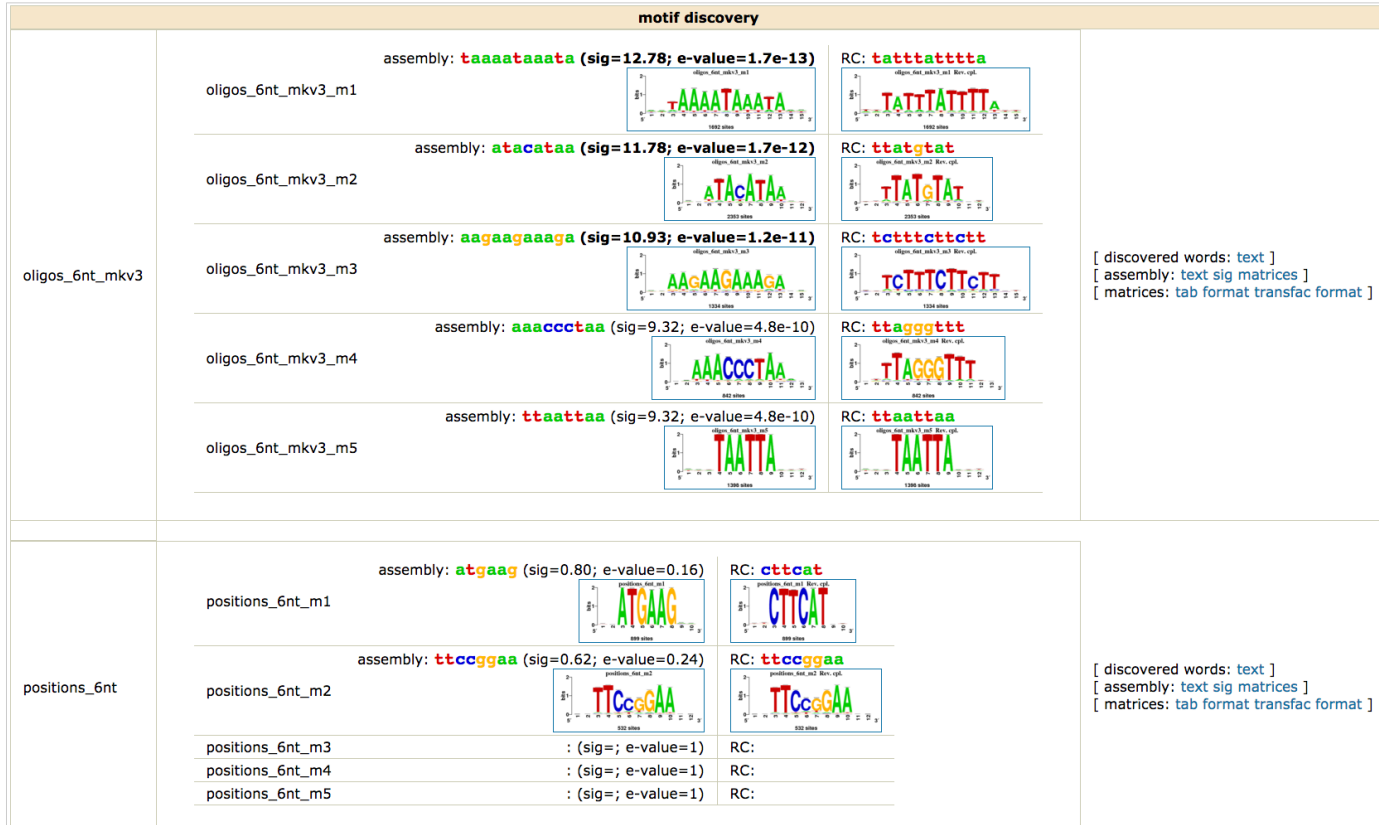
cluster_2



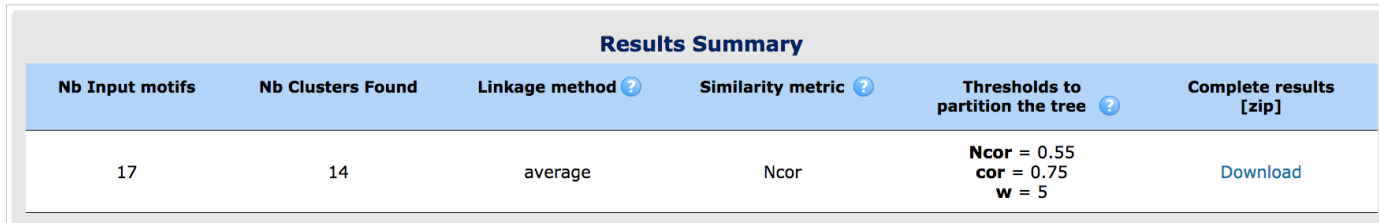
cluster_3



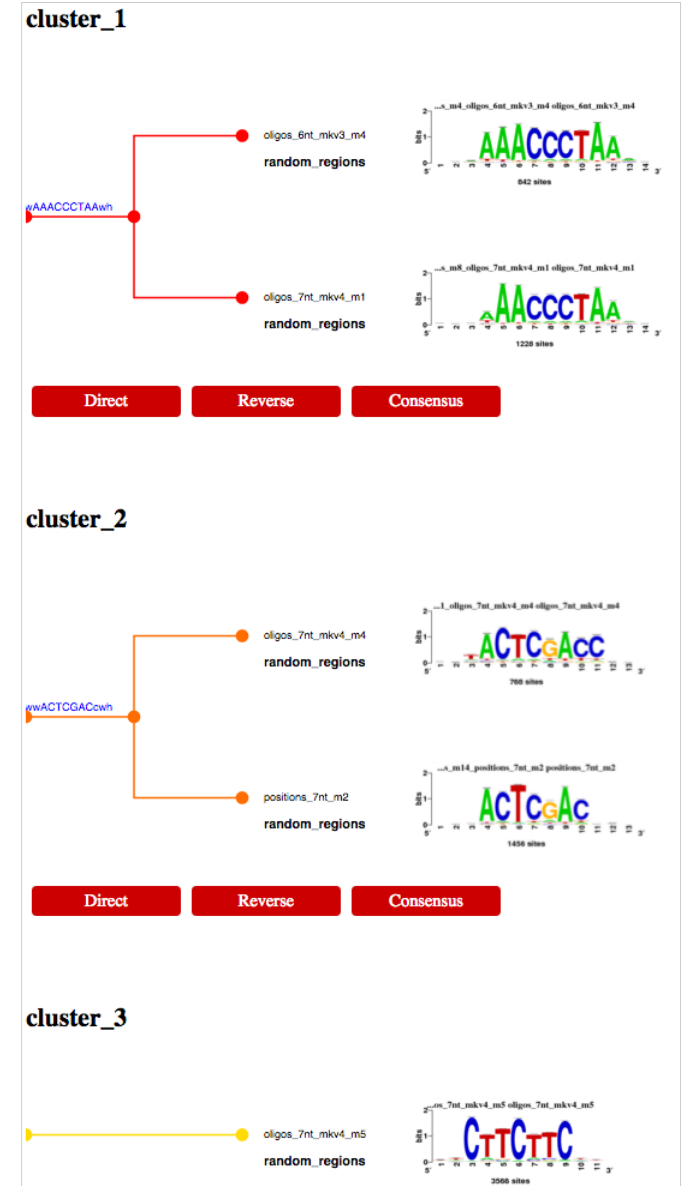
A



B



C



Macs2_qval0.05	2931	1	0.34	0.43	0.15	0.14	0.98	0.96	0.8	0.51	0.09	0.04
Macs2_qval0.001	9711	1	1	0.88	0.43	0.39	1	0.99	0.99	0.96	0.28	0.13
Macs14_pval0.00001	6242	0.99	0.76	1	0.39	0.39	0.98	0.98	0.95	0.86	0.3	0.19
Homer_fdr0.01	18812	0.97	0.91	0.98	1	0.68	0.98	0.98	0.98	0.97	0.54	0.26
SPP_fdr0.01	24781	1	0.96	0.99	0.82	1	1	0.99	0.99	0.97	0.64	0.42
SPP_fdr0.001	544	0.18	0.07	0.09	0.03	0.03	1	0.39	0.19	0.1	0.02	0.01
SWEMBL_R0.1	1352	0.47	0.18	0.21	0.07	0.07	0.98	1	0.48	0.25	0.04	0.02
SWEMBL_R0.07	2788	0.8	0.37	0.41	0.15	0.15	0.98	1	1	0.53	0.09	0.04
SWEMBL_R0.05	5256	0.97	0.65	0.68	0.28	0.27	0.99	1	1	1	0.17	0.08
SWEMBL_R0.02	31867	1	1	1	0.89	0.83	1	1	1	1	1	0.45
SWEMBL_R0.01	92695	1	1	1	0.99	0.98	1	1	1	1	1	1

Nb_peaks

Macs2_qval0.05

Macs2_qval0.001

Macs14_pval0.00001

Homer_fdr0.01

SPP_fdr0.01

SPP_fdr0.001

SWEMBL_R0.1

SWEMBL_R0.07

SWEMBL_R0.05

SWEMBL_R0.02

SWEMBL_R0.01