IBIS PWM Submissions

# Regulatory Sequence Analysis Tools (RSAT)

## Team name

RSAT

## Primary Disciplines

A2G-PWM, G2A-PWM

## Models

k-mers; oligonucleotides; dyads; over-representation; position bias; PSSM; PCM; matrix clustering; optimisation; genetic algorithm

# Summary

We describe the procedure used to produce the Position-Weight Matrices (PSM) submitted by the RSAT team to the IBIS challenge 2024. Our approach combines the detection of exceptional k-mers in the training sequences, using these k-mers to build position-count matrices (PCM) based on the collection of binding sites in the training sequences, clustering of the resulting PCMs, and a new genetic algorithm-based optimisation approach to optimize the capability of these matrices to discriminate between a positive and a negative sequence set (measured by area under the ROC curve). .

# Implementation & Software

## RSAT software suite

* <https://github.com/rsa-tools/rsat-code>,
  + tag [2024-08-28c](https://github.com/rsa-tools/rsat-code/releases/tag/2024-08-28c)
  + installation (Ubuntu + Mac OS) : <https://github.com/rsa-tools/installing-RSAT>
* <https://hub.docker.com/r/eeadcsiccompbio/rsat/tags>
  + tag: 2024-08-28c
  + installation: docker pull eeadcsiccompbio/rsat:2024-08-28c

## optimize-matrix-GA

* <https://github.com/pvhelden/optimize-matrix-GA>
  + requires RSAT installation, either in the Unix shell or as a docker

## RSAT scripts and protocol for the IBIS challenge 2024

* <https://github.com/jvanheld/IBIS_2024>
* Includes detailed instructions to reproduce the software environment and all the steps of the analysis.

# Methods

## Motif discovery

We used the tool *peak-motifs* (Thomas-Chollier et al., 2011, Thomas-Chollier et al. 2912) from the software suite *Regulatory Sequence Analysis Tools (RSAT)* (Santana Garcia et al., 2022) to discover motifs in the train sequences using a Docker container (Ksouri et al., 2021). Motif discovery relies on the detection of exceptional over-represented k-mers (*oligo-analysis*, van Helden et al., 1998à or dyads (*dyad-analysis*, van Helden et al., 2000) as well as positionally biased k-mers (*position-analysis*, van Helden et al., 2001). We systematically analyzed 6-mers and 7-mers for *oligo-analysis* and *position-analysis*. Significant k-mers serve as seed to build a position-count matrix by collecting putative binding sites in the train sequences. For GCS, GHTS, HTS and SMS experiments, a Markov background model of order k-2 was estimated from the train sequences themselves, For PBM data, *peak-motifs* was used in differential mode to detect oligos and dyads over-represented in the top 500 spots (having the strongest signal) versus the 3500 bottom spots (considered as background sequences).

Note that *RSAT peak-motifs* includes an option to compare each discovered motif with one or several motif databases. For the sake of information, the peak-motifs reports systematically include a comparison with Hocomoco Human and Jaspar non-redundant vertebrate databases, but the correspondence between discovered motifs and annotated motifs was deliberately ignored for the selection of the submitted PWM, which was exclusively based on the analysis of AuROC metrics, as described below (see section “Matrix selection”).

## Motif clustering

Primary motifs returned by *peak-motifs* were fed to the RSAT tool *matrix-clustering* [4] in order to regroup similar motifs (e.g. discovered by the different algorithms or with different k-mer lengths).

## Motif optimisation

Both primary discovered and clustered motifs were used as seeds for a new software named *optimize-matrices-GA* [5] that runs a genetic algorithm to optimize motifs with respect to a given objective function. In our case, the objective function was the area under the ROC curve (AuROC) measured with a positive and a negative sequence set. For the negative sequence set, we alternatively considered two approaches:

1. Train versus random
   * positive set: the training sequence for the considered dataset
   * negative set: random genome fragments of the same sizes as the training sequences (picked up with the RSAT tool *random-genome-fragments*)
2. TF versus others
   * The positive sequence set was a collection of all the training sequences for the considered factor in the considered experiment type (replicates and multi-cycle sequence sets were thus merged).
   * The negative set was the concatenation of all the sequence sets of all the other factors in the same experiment type.

We used the following parameters for *optimize-matrices-GA* at each generation:

* 20 generations after (the initial matrices is referred to as generation 0)
* 10 children per parent matrix
* for each child, we mutated 1 position by changing the frequency of a randomly chosen residue by a factor sampled randomly (uniform distribution between 5 and 25 percent)
* we keep the 5 top-ranking individuals for the next generation
* note that the 5 parent matrices from the previous generation are included in the current generation (they are competing with their own children)

## Matrix selection

For the IBIS submission, we gathered all the motifs produced in these different steps (primary peak-motifs discovered, clustered matrices, optimized matrices), measured their performances in binary classification, sorted them altogether by decreasing AuROC, and selected the top-ranking matrices for each TF:experiment couple. This first selection criterion systematically returned motifs of the 20th or 19th generation.

However, these matrices were optimized based on the training set, and some of them may fail to generalize to new data. We thus also included some of the pre-optimization matrices (generation 0).

Note that before the submission deadline, we had no time to implement the classical methods to mitigate over-fitting (e.g. k-fold cross-validation on the training set). This functionality will be implemented in a future version of *optimize-matrix-GA*.

## PWM naming for the submissions

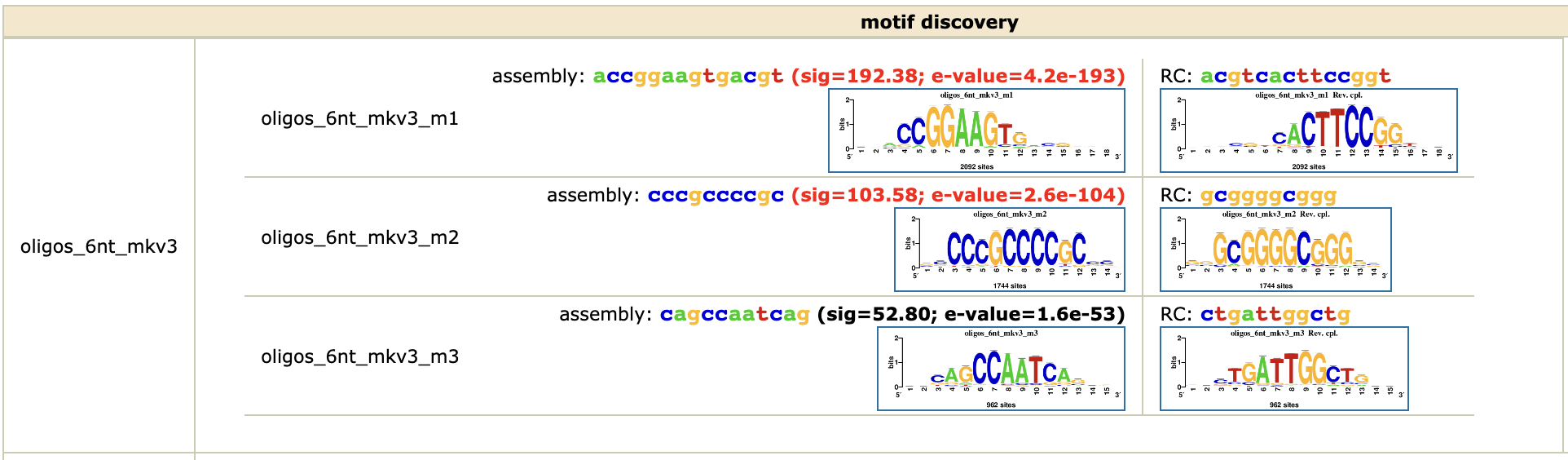
We adopted a systematic syntax for naming the PWMs before submission.

(TF)\_([CGHSP])\_((o6\_7)(p6\_7)|d|(c\d\_n\d\_\dm))(\_G\d+\_M\d+\_C\d+){0,1}

| (TF) | transcription factor name |
| --- | --- |
| [CGHSP] | initial of the experiment (CHS, GHTS, HTS, SMS, PBM |
| ([opdc]) | algorithm having produced the motif: oligo-analysis (o), dyad-analysis (d), position-analysis (p), matrix-clustering (c), followed by the motif number (\d) |
| o6\_7  p6\_7 | For oligo-analysis and position-analysis, the lengths of the seeding oligonucleotides are indicated (6\_7) |
| c\d\_n\d\_\dm | Cluster motifs are followed by an indication of the tree node number and the number of descendants (e.g. c1\_n4\_5m), except for single-motif clusters which are simply numbered (e.g. c3) |
| \_G\d+\_M\d+\_C\d+ | Suffix present in motifs resulting from optimize-matrices-GA: generation, parent number, child number |

# Results & Discussion

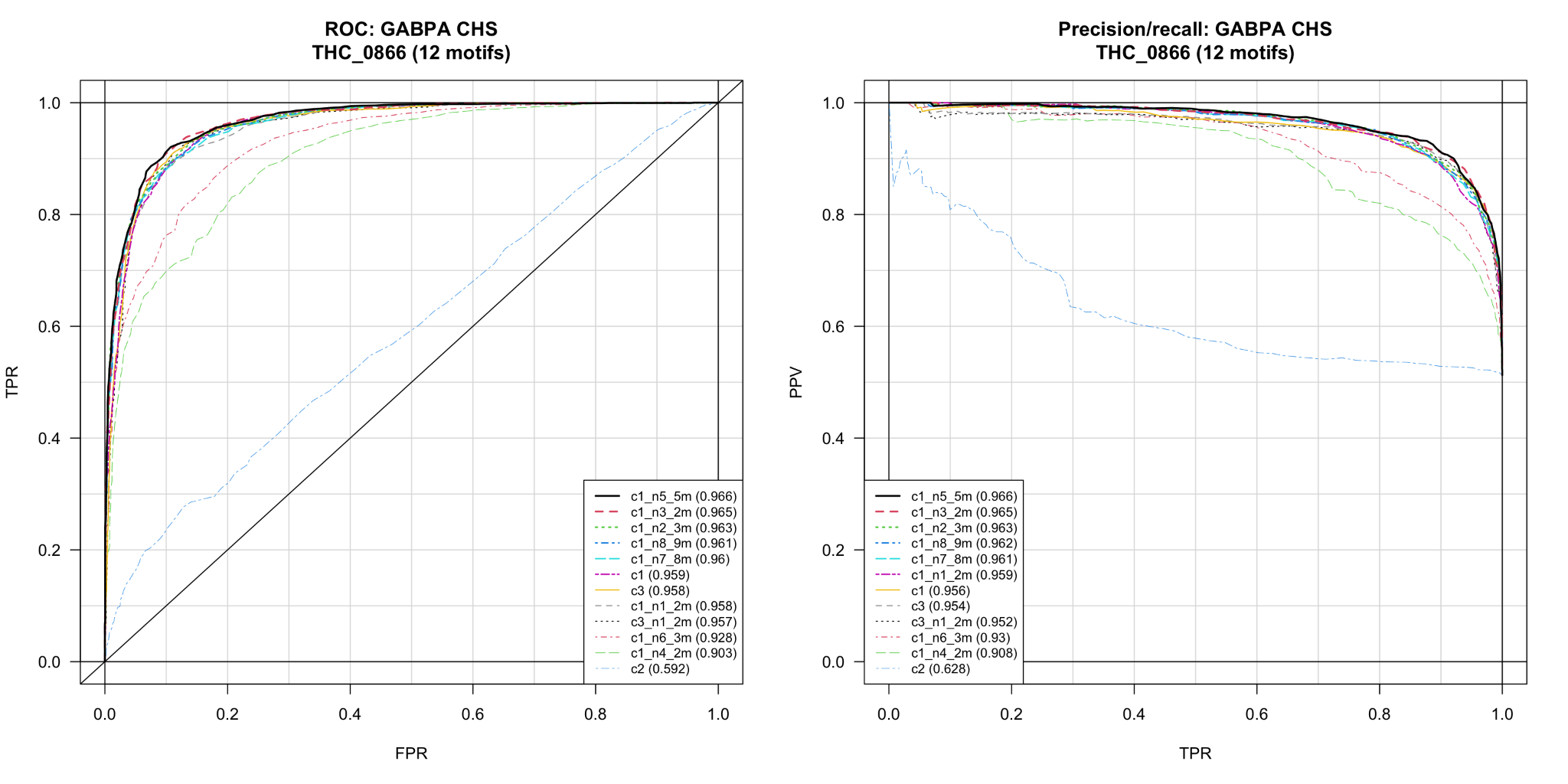
## Motif discovery



* Show a snapshot of peak-motifs discovered motifs
  + Head of a k-mer table
  + Logos and sig of the discovered motifs
* Collect stats about k-mer significance with peak-motifs

## Performances of the motifs for binary classification

Below, some examples of ROC and PR curves illustrating the diversity of the quality.



ROC (left) and PR (right) curves for the motifs discovered with peak-motifs and clustered with matrix-clustering

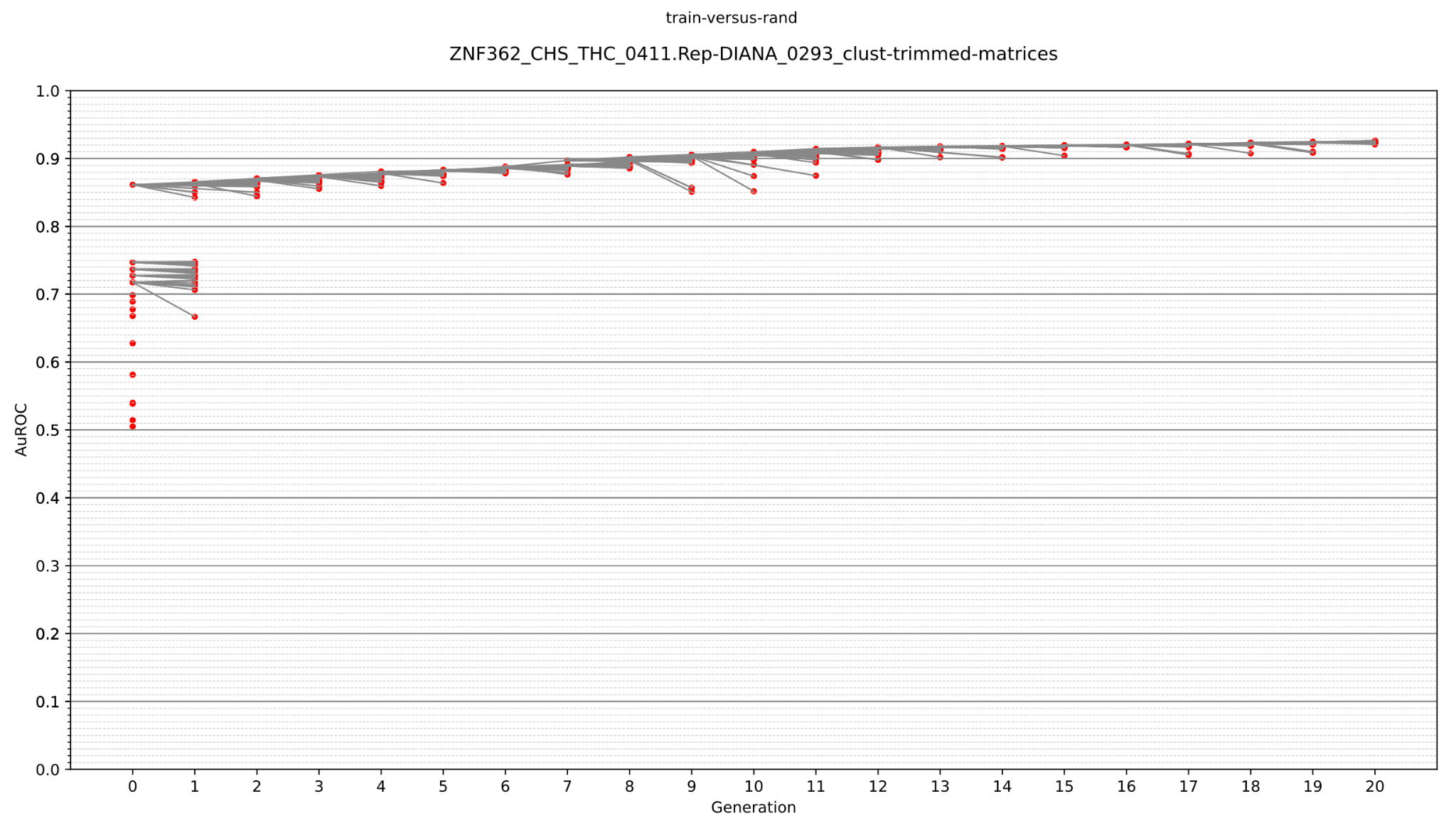
* show some commented examples of ROC curves
* provide a link to the html report on github

## Motif optimization

The 3 figures below show the profiles of Area under the Receuving Operating Characteristic curve (auROC) with 3 illustrative datasets.

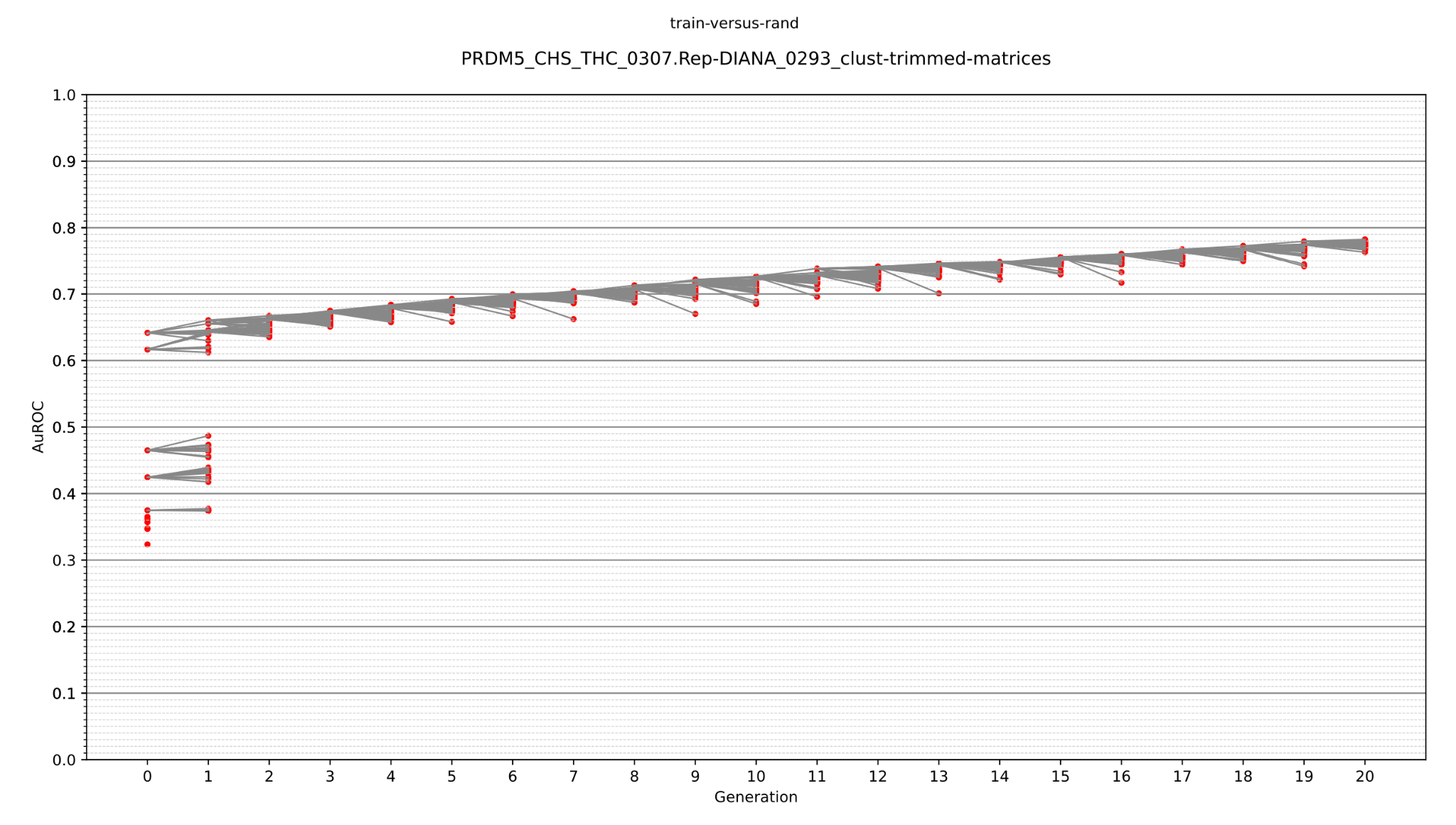
### Case 1: ZNF362 ChIP-seq data

The discovered motifs produced by RSAT achieve AuROC scores ranging from 0.5 to 0.87. During 20 generations, the genetic algorithm progressively increases the AuROC and seems to converge around AuROC=094. .



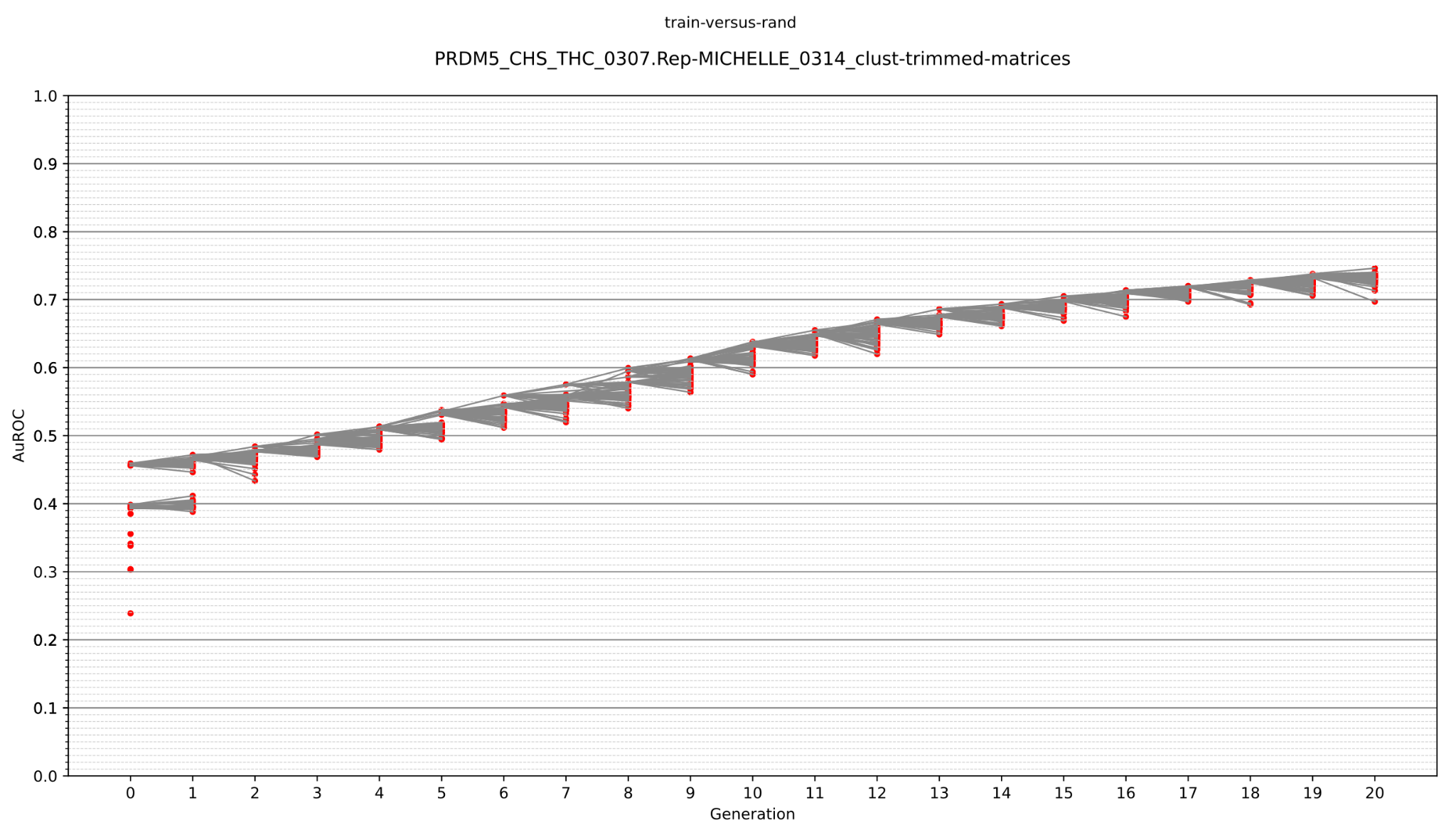
### Case 2: PRDM5 ChIP-seq data

Surprisingly, 8 of the motifs returned by peak-motifs achieve scores inferior to 0., which is the random expectation. The best motif, which achieves an AuROC of 0.64 (generation 0), evolves towards steadily increasing AuROC scores during the 20 generations,



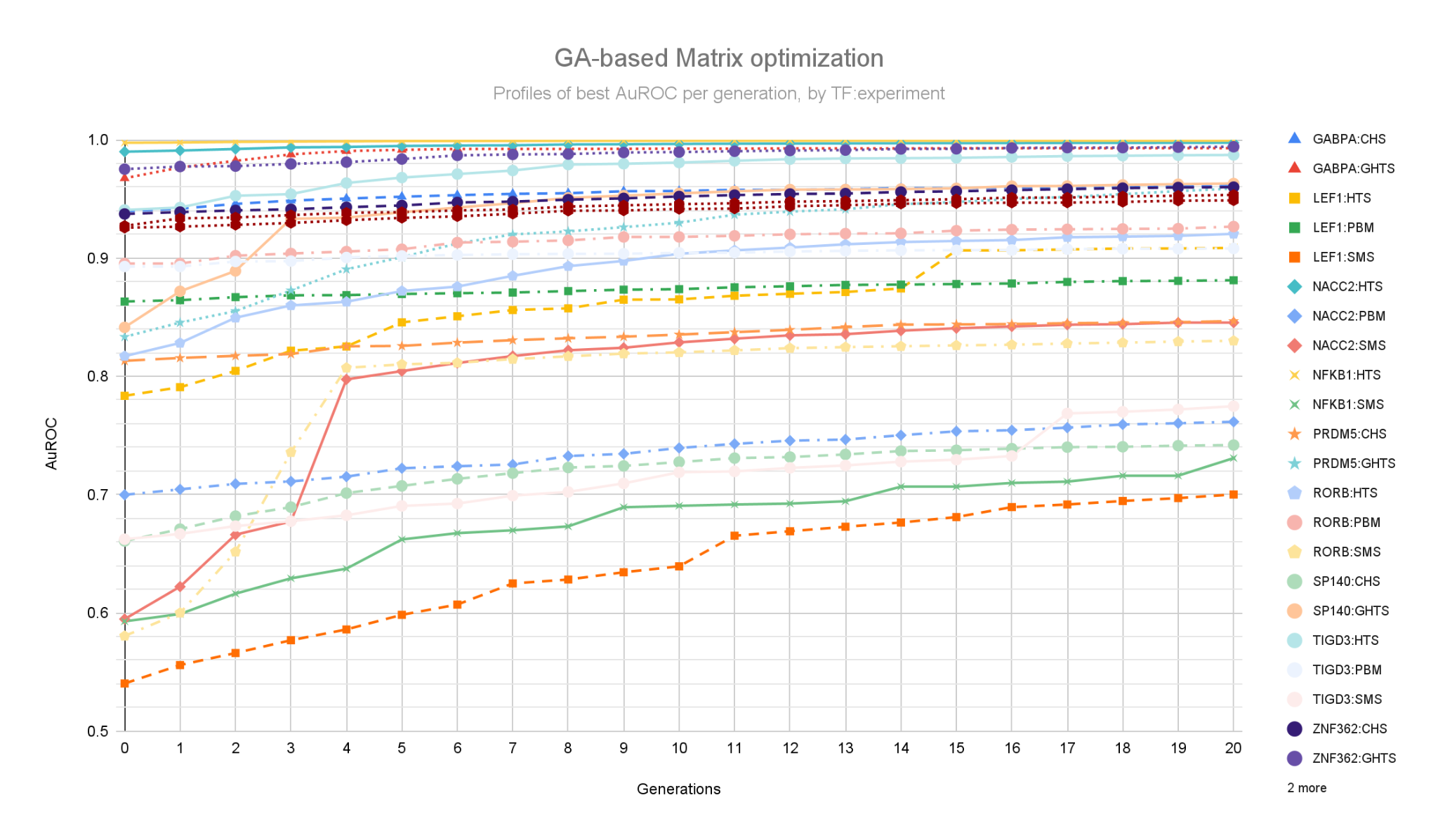
### Case 3. PRDM5 ChIP-seq

In this case all the motifs initially return d by peak-motifs achieve particularly weak AuROC scores, ranging between 0.24 and 0.46. Thetool optimize-matrices-GA evolves the top-ranking of these (poor) motifs up to an AuROC of ~75. This suggests that this algorithm might be able to discover motifs starting from some random seed matrix (e.g. a matrix with the same counts for all the residues in each column).



### General performances of the Genetic algorithm

The figure below summarizes the increase in performances as a function of the GA generations (from 0 to 20).



**Figure: profiles of best AuROC per generation, datasets grouped by TF x experiment.** Each dot indicates the maximal AuROC (ordinate) achieved among all the individual matrices of a given generation (abscissa) across all the datasets for a given TF and a given type of experiment.

## Leaderboard tests

The leaderboard tests were informative to choose the general motif selection strategy.

In particular, we observed a significant improvement of RSAT team ranking when we submitted the GA-optimized matrices, but we also noticed that for some combinations of TF:experiment the pre-optimisation matrices (generation 0) showed better performances. This effect likely results from some overfitting of the optimized matrices, which is expected especially for small datasets. We thus decided to submit a combination of the original and optimized matrices.

# References

Castro-Mondragon JA, Jaeger S, Thieffry D, Thomas-Chollier M, van Helden J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. 2017 Jul 27;45(13):e119. doi: 10.1093/nar/gkx314. PMID: 28591841; PMCID: PMC5737723.

Ksouri N, Castro-Mondragón JA, Montardit-Tarda F, van Helden J, Contreras-Moreira B, Gogorcena Y. Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants: the peach example. Plant Physiol. 2021 Apr 2;185(3):1242-1258. doi: 10.1093/plphys/kiaa091. PMID: 33744946; PMCID: PMC8133646.

Santana-Garcia W, Castro-Mondragon JA, Padilla-Gálvez M, Nguyen NTT, Elizondo-Salas A, Ksouri N, Gerbes F, Thieffry D, Vincens P, Contreras-Moreira B, van Helden J, Thomas-Chollier M, Medina-Rivera A. RSAT 2022: regulatory sequence analysis tools. Nucleic Acids Res. 2022 Jul 5;50(W1):W670-W676. doi: 10.1093/nar/gkac312. PMID: 35544234; PMCID: PMC9252783.

Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat Protoc. 2012 Jul 26;7(8):1551-68. doi: 10.1038/nprot.2012.088. PMID: 22836136.

Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res. 2012 Feb;40(4):e31. doi: 10.1093/nar/gkr1104. Epub 2011 Dec 8. PMID: 22156162; PMCID: PMC3287167.

van Helden J, André B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol. 1998 Sep 4;281(5):827-42. doi: 10.1006/jmbi.1998.1947. PMID: 9719638.

van Helden J, del Olmo M, Pérez-Ortín JE. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res. 2000 Feb 15;28(4):1000-10. doi: 10.1093/nar/28.4.1000. PMID: 10648794; PMCID: PMC102588.

van Helden J, Rios AF, Collado-Vides J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. 2000 Apr 15;28(8):1808-18. doi: 10.1093/nar/28.8.1808. PMID: 10734201; PMCID: PMC102821.

van Helden P, Contreras-Moreira B, van Helden P (*in prep*). Machine learning strategies to optimize motifs discovered in transcription factor binding sequences.