GimmeMotifs: an analysis framework for transcription factor motif analysis

This manuscript (permalink) was automatically generated from simonvh/gimmemotifs-manuscript@5f7f461 on October 17, 2018.

Authors

- **Simon J. van Heeringen**
  0000-0002-0411-3219 · simonvh · svheeringen

  Radboud University, Faculty of Science, Department of Molecular Developmental Biology, Radboud Institute for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands · Funded by NWO-ALW, grant 863.12.002

Abstract

Background: Transcription factors (TFs) bind to specific DNA sequences, TF motifs, in cis-regulatory sequences and control the expression of the diverse transcriptional programs encoded in the genome. The concerted action of TFs within the chromatin context enables precise temporal and spatial expression patterns. To understand how TFs control gene expression it is essential to model TF binding. TF motif information can help to interpret the exact role of individual regulatory elements, for instance to predict the functional impact of non-coding variants.

Findings: Here we present GimmeMotifs, a comprehensive computational framework for TF motif analysis. It includes tools for *de novo* motif discovery, motif scanning and sequence analysis, clustering, calculation of performance metrics and visualization. Included with GimmeMotifs is a non-redundant database of clustered motifs. Compared to other motif databases, this collection of motifs shows competitive performance in discriminating bound from unbound sequences. Using our *de novo* motif discovery pipeline we find large differences in performance between *de novo* motif finders on ChIP-seq data. Finally, we demonstrate *maelstrom*, a new ensemble method that enables comparative analysis of TF motifs between multiple high-throughput sequencing experiments, such as ChIP-seq or ATAC-seq. Using a collection of ~300 H3K27ac ChIP-seq data sets we identify TFs that play a role in hematopoietic differentiation and lineage commitment.

Conclusion: GimmeMotifs is a fully-featured and flexible framework for TF motif analysis. It contains both command-line tools as well as a Python API and is freely available at: https://github.com/vanheeringen-lab/gimmemotifs.

Introduction

The regulatory networks that determine cell and tissue identity are robust, yet remarkably flexible. Transcription factors (TFs) control the expression of genes by binding to their cognate DNA sequences, TF motifs, in cis-regulatory elements [1]. To understand how genetic variation affects binding and to elucidate the role of TFs in regulatory networks we need to be able to accurately model binding of TFs to the DNA sequence.

The specificity of DNA-binding proteins can be modeled using various representations [2]. One of the most widely adopted is the position frequency matrix (PFM). This matrix, a TF motif, contains (normalized) frequencies of each nucleotide at each position in a collection of aligned binding sites. These PFMs can be derived from high-throughput experiments such as Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) [3,4,5,6], HT-SELEX [7] or Protein Binding Microarrays (PBMs) [8]. Through straightforward transformations a PFM can be expressed as a weight matrix, using log likelihoods, or information content, using the Kullback-Leibler divergence.

Even though the PFM is a convenient representation, it has certain limitations. A PFM cannot model inter-nucleotide dependencies, which are known to affect binding of certain TFs. Multiple different representations have been proposed [10,11,12,13,9], but no single one of these has gained much traction. Ultimately we will need these types of advanced models to accurately represent TF binding. However, PFMs still serve as a very useful abstraction that enables an intuitive understanding of TF binding.

Here, we present GimmeMotifs, a Python module and set of command-line tools to provide an comprehensive framework for transcription factor motif analysis. Amongst other possibilities it can be used to perfom *de novo* motif analysis, cluster and visualize motifs and to calculate enrichment statistics. A new ensemble method, *maelstrom*, can be used to determine differential motif activity between two or more experiments. We illustrate the functionality of GimmeMotifs using three different examples.

Findings

Benchmark of transcription factor motif databases

A variety of transcription factor (TF) motif databases have been published based on different data sources. One of the most established is JASPAR, which consists of a collection of non-redundant, curated binding profiles [14]. The JASPAR website contains many other tools and the underlying databases are also accessible via an API [15]. Other databases are based on protein binding micorarrays [8], HT-SELEX [7] or ChIP-seq profiles [3,4,5,6]. CIS-BP integrates many individual

motif databases, and includes assignments of TFs to motifs bases based on DNA binding domain homology [16].

For the purpose of motif analysis, it is beneficial to have a database that is non-redundant (i.e., similar motifs are grouped together), yet as complete as possible (i.e., covers a wide variety of TFs). To establish a quantitative measure of database quality, we evaluated how well motifs from different databases can classify ChIP-seq peaks from background sequences using the GimmeMotifs tool `gimme roc`. We downloaded ChIP-seq peaks from ReMap 2018 [17], and selected all TFs with at least 1,000 peaks. We then evaluated 8 motif databases to test how well they could distinguish peaks from random genomic sequences. When a data set contained more than 5,000 peaks we randomly selected 5,000 peaks for the analysis. We included the following databases: JASPAR 2018 vertebrate [14], SwissRegulon [18], Homer [6], Factorbook [3], the ENCODE motifs from Kheradpour et al. [4], HOCOMOCO [5], the RSAT clustered motifs [19] and the motif database created by Madsen et al. [20]. Figure 1A shows distribution of the ROC AUC (area under the curve for the Receiver Operator Curve) of the best motif per database for all 294 transcription factors in a box plot. There is generally a wide distribution of ROC AUCs. For some factors, such as ELK1, CTCF, CBFB and MYOD1, peaks are relatively easy to classify using a single PFM motif. Other factors don't have peaks with a consistently enriched motif, or do not contain a sequence-specific DNA-binding domain, such as EP300 or CD2 for example.

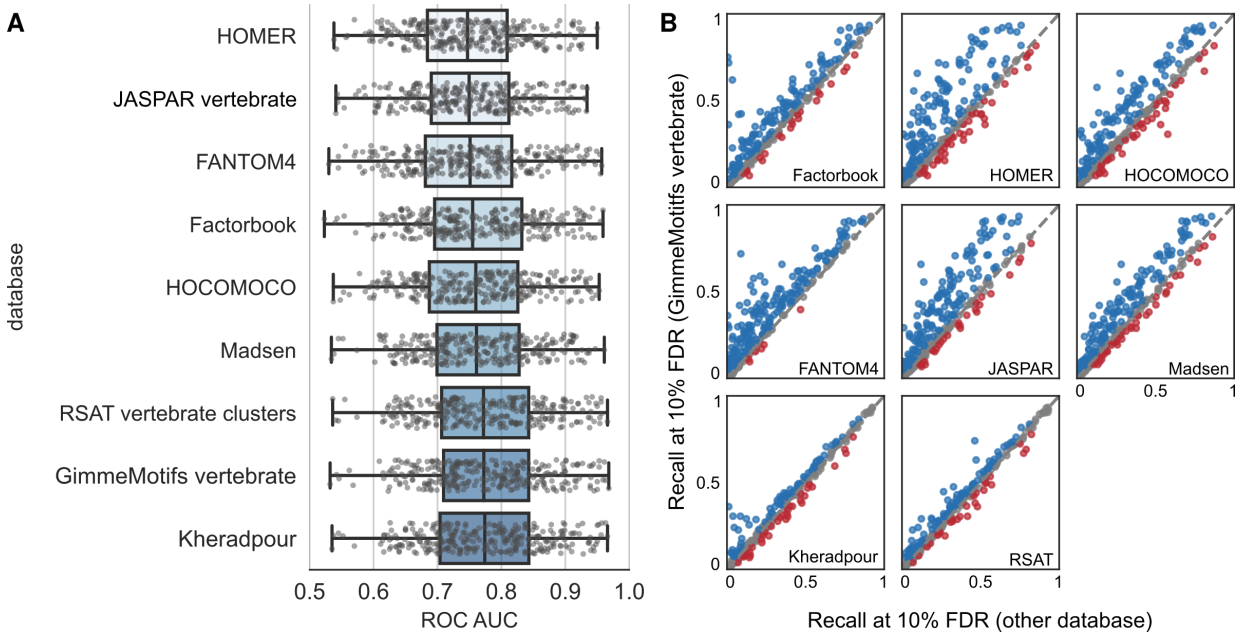

Figure 1: Benchmark of transcription factor motif databases. **A)** Motif-based classification of binding sites for 294 TFs from the ReMap ChIP-seq database. For all TFs 5000 peaks were compared to background regions using each motif database. The boxplot shows the The ROC AUC of the best motif per database for all TFs. **B)** Recall at 10% FDR of motif databases compared to the GimmeMotifs vertebrate motif database (v4.0). The same data is used as in **A)**. The X-axis represents the recall for the different databases, the Y-axis represents the recall for the GimmeMotifs vertebrate database. Differences of more than 0.025 are marked blue, and less then -0.025 red.

The difference in maximum ROC AUC between databases is on average not very large, with a mean maximum difference of 0.05. The largest difference (~0.24) is found for factors that were not assayed by ENCODE, such as ONECUT1, SIX2 and TP73, and are therefore not present in the Factorbook motif database. Unsurprisingly, the databases that were based on motif collections of different sources (Kheradpour, Madsen, RSAT and Gimme) generally perform best. It should be noted that, for this task, using motif databases based on motif identification from ChIP-seq peaks is in some sense “overfitting”, as the motifs in these databases were inferred from highly similar data.

While the ROC AUC is often used to compare the trade-off between sensitivity versus specificity, in this context it is not the best metric from a biological point of view. An alternative way of measuring performance is evaluating the recall (ie. how many true peaks do we recover) at a specific false discovery rate. This is one of the criteria that has been used by the ENCODE DREAM challenge for evaluation [21]. Figure 1B shows scatterplots for the recall at 10% FDR for all motif databases compared to the clustered, non-redundant databases that is included with GimmeMotifs. This database shows better performance than most other databases using this benchmark. The non-redundant RSAT database, which was created in a very similar manner [19], scores comparably.

These results illustrate how `gimme roc` can be used for evaluation of motifs. The choice of a motif database can greatly influence the results of an analysis. The default database included with GimmeMotifs shows good performance on the metric evaluated here. However, this analysis illustrates only one specific use case of application of a motif database. In other cases well-curated databases such as JASPAR can be beneficial, for instance when linking motifs to binding proteins.

Large-scale benchmark of *de novo* motif finder performance on ChIP-seq peaks

It has been noted that there is no *de novo* motif prediction algorithm that consistently performs well across different data sets [22]. New approaches and algorithms for *de novo* motif discovery continue to be published, however, many of them are not tested on more than a few datasets. Benchmarks that have been published since Tompa et al. [23,24] typically have tested only a few motif finders or used only a few datasets.

Here, we used the GimmeMotifs framework as implemented in `gimme motifs` to benchmark 14 different *de novo* motif finders. To evaluate the different approaches, we downloaded 495 peak files for 270 proteins from ENCODE [25] and selected the 100bp sequence centered on the summit of top 5000 peaks. Of those peaks, half were randomly selected as a prediction set and the other half was used for evaluation. As a background set we selected regions of the same length flanking the original peaks. To assess the performance, we calculated two metrics, the ROC AUC and the recall at 10% FDR. Figure 2A shows the distribution of the ROC AUC scores over all ENCODE peaks in a boxplot, ordered by the mean ROC AUC. The ROC AUC is distributed between 0.58 and 0.98, with a mean of 0.75. All proteins that have low ROC AUC are not sequence-specific transcription factors such as POL2, TAF7 and GTF2B, the PRC2-subunit SUZ12 and the H3K9 methyltransferase SETDB1. The factors with the highest ROC AUC are CTCF and members of the cohesin complex, SMC3 and RAD21, that bind at CTCF sites.

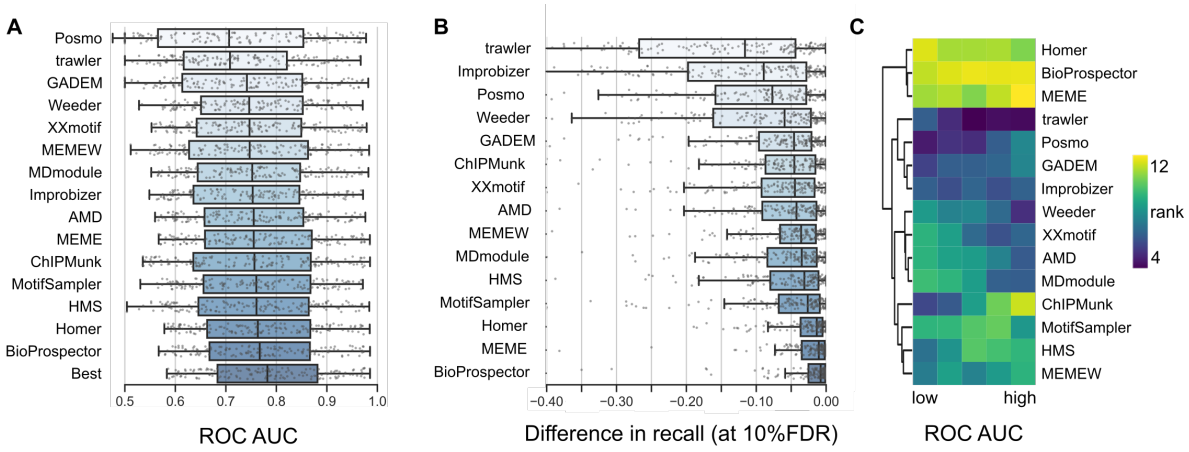

Figure 2: Benchmark of *de novo* motif finders. **A)** Comparison of the ROC AUC of the best motif of each motif finder. The boxplot shows the best motif per peak set of 495 peaks for 270 proteins from ENCODE. The best motif from all motif finders is indicated as 'Best'. **B)** Comparison of the best motif per motif finder compared to the best overall motif for each data set. Plotted is the difference in recall compared to the best motif. Recall is calculated at 10% FDR. **C)** The relative motif rank as a function of the motif quality. Rank is the mean overall rank of three metrics (ROC AUC, recall at 10% FDR and MNCP).

Generally, the ROC AUC distribution of all evaluated motif finders is very similar. However, a few outliers can be observed. Trawler and Posmo show an overall lower distribution of ROC AUC scores. Compared to the ROC AUC scores of the next best program, GADEM, this is significant ($p < 0.01$, Wilcoxon signed-rank). Selecting the best motif for each experiment results in a ROC AUC distribution that is significantly higher than the best single method, BioProspector ($p < 1e\text{-}21$, Wilcoxon signed-rank).

As stated in the previous section, the ROC AUC is not the best measure to evaluate motif quality. Therefore, we selected for every peak set the best motif from all motifs predicted by the different motif finders on the basis of the recall at 10% FDR. We then plotted the difference between the best motif from each individual *de novo* approach with this best overall motif (Fig. 2B). For this figure, we used only the data sets where at least one motif had a recall higher than 0 at 10% FDR.

In line with previous results [22], there is no single tool that consistently predicts the best motif for each transcription factor. However, the motifs predicted by BioProspector, MEME and Homer are, on basis of this metric, consistently better than motifs predicted by other methods. In 75% of the cases, the motif predicted by BioProspector has a difference in recall smaller than 0.026 compared to the best overall motif. In this benchmark, four programs (Trawler, Improbizer, Posmo and Weeder) generally perform worse than average, with a mean decrease in recall of 0.11 to 0.17, as compared to the best motif. In addition, these programs tend to have a much more variable performance overall.

Predicted motifs identified using MEME with different motif widths show better performance than running MEME with the `minw` and `maxw` options (MEME vs. MEMEW in Fig. 2B). Of the best

performing algorithms, both MEME and BioProspector were not specifically developed for ChIP-seq data, however, they consistently outperform most methods created for ChIP-seq data. Of the ChIP-seq motif finders Homer consistently shows good performance.

Finally, to gain further insight into *de novo* motif finder performance, we stratified the ChIP-seq datasets by motif "quality". We divided the transcription factors into five bins on basis of the ROC AUC score of the best motif. For each bin we ranked the tools on basis of the average of three metrics (ROC AUC, recall at 10% FDR and MNCP [26]). The results are visualized as a heatmap in Figure 2C. From this visualization, it is again clear that BioProspector, MEME and Homer produce consistently high-ranking motifs, while the motifs identified by Trawler, Posmo, GADEM and Improbizer generally have the lowest rank. Interestingly, for some motif finders, there is a relation between motif presence and the relative rank. Weeder, XXMotif and MDmodule yield relatively high-ranking motifs when the ROC AUC of the best motif for the data set is low. On the other hand, ChIPMunk shows the opposite pattern. Apparently this algorithm works well when a motif is present in a significant fraction of the data set.

These results illustrate that motif finders need to be evaluated along a broad range of data sets with different motif presence and quality. Another interesting observation is that this ChIP-seq benchmark shows a lower-than-average performance for Weeder, which actually was one of the highest scoring in the Tompa et al. benchmark. It should be noted that our metric specifically evaluates how well *de novo* motif finders identify the primary motif in the context of ChIP-seq peaks. It does not evaluate other aspects that might be important, such as the ability to identify many low-abundant motifs. Furthermore, with ChIP-seq data there are usually thousands of peaks available. This allows for other algorithms than those that work well on a few sequences. Interestingly, the original MEME shows consistently good performance, although the running time is longer than most other tools. On the basis of this analysis, BioProspector should be the top pick for a program to identify primary motifs in ChIP-seq data. However, an ensemble program such as GimmeMotifs will report high-quality motifs more consistently than any single tool.

Differential motif analysis of hematopoietic enhancers identifies cell type-specific regulators

While many motif scanners and methods to calculate enrichment exist, there are few methods to compare motif enrichment or activity between two or more data sets. The CentriMo algorithm from the MEME suite implements a differential enrichment method to compare two samples [27]. Other approaches, such as MARA [28,29] and IMAGE [20], are based on linear regression. Here we present the *maelstrom* algorithm that integrates different methods to determine motif relevance or activity in an ensemble approach (Fig. 3A).

To demonstrate the utility of maelstrom we identified motif activity based on enhancers in hematopoietic cells. We downloaded 69 human hematopoietic DNaseI experiments

(Supplementary Table S1), called peaks, and created a combined peak set as a collection of putative enhancers. In addition we downloaded 193 hematopoietic H3K27ac ChIP-seq experiments, mainly from BLUEPRINT [30] (Supplementary Table S1). We determined the number of H3K27ac reads per enhancer (Supplementary Table S2). After log2 transformation and scaling, we selected the 50,000 most dynamic peaks. Figure 3B shows the correlation of the H3K27ac enrichment in these 50,000 enhancers between cell types. For this plot, replicates were combined by taking the mean value and all experiments corresponding to treated cells were removed. We can observe five main clusters 1) non-hematopoietic cells, megakaryocyte and erothrocytes 2), lymphoid cells, 3) neutrophilic cells, 4) macrophages and dendritic cells and 5) monocytes. The lymphoid cluster furthermore seperates between B-cells and T- and NK cells and non-hematopoietic cells are distinct from the megakaryocytes and erythroblasts. We can conclude that the H3K27ac profile within this enhancers set recapitulates a cell type-spefific regulatory signal.

Figure 3: Predicting TF motif activity using maelstrom. **A)** An overview of the *maelstrom* ensemble method. **B)** Heatmap of the correlation of H3K27ac signal in hematopoietic enhancers. We counted H3K27ac ChIP-seq reads in 2kb sequences centered at DNase I peaks. Counts were log2-

transformed and scaled and replicates were combined by taking the mean value. This heatmap shows the Pearson r, calculated using the 50,000 most dynamic peaks. **C)** Results of running `gimme maelstrom` on the 50,000 most dynamic hematopoietic enhancers. The reported motifs activity represents log10-transformed p-value of the rank aggregation. For high-ranking motifs -log10(p-value) is shown.

To determine differential motif activity from these dynamic enhancers we used maelstrom. We combined Bayesian ridge regression, multi-class regression using coordinate descent [31] and regression with boosted trees [32]. The coefficients or feature importances were ranked and combined using rank aggregation [33]. A p-value was calculated for consistently high ranking and consistently low ranking motifs. The results are visualized in Figure 3C.

Two of the most signicant motifs are SPI1 (PU.1) and CEBP. The motif activity for SPI1 is high in monocytes and macrophages, consistent with its role in myeloid lineage commitment [34]. The CEBP family members are important for monocytes and granulocitic cells [35], and show a high motif activity in neutrophils and monocytes. Other strong motifs include RUNX for T cells and NK cells, GATA1 for erythroid cells.

We identified a strong activity for motifs representing the ZEB1 and Snail transcription factors. The Snail transcription factors play an important role in the epithelial-to-mesenchymal transition (EMT), and their role in hematopoietic cells is less well-described. However, recently Snai2 and Snai3 were found to be required to generate mature T and B cells [36,37] in mice. ZEB1 is expressed in T cells and represses expression of IL-2 [38], as well as other immune genes such as CD4 [39] and GATA3 [40]. ZEB1 knockout mice exhibit a defect in thymocyte development [41]. Together, this suggests that these TFs play an important role in lymphocyte development.

Finally, an interesting observation is the predicted motif activity of NANOG in endothelial cells. NANOG is expressed in embryonic stem cells and is essential for maintenance of pluripotency [42]. However, NANOG is indeed also expressed in endothelial cells and has been shown to play a role in endothelial proliferation and angiogenesis [43].

Methods

GimmeMotifs

Implementation

GimmeMotifs is implemented in Python, with the motif scanning incorporated as a C module. The software is developed on GitHub (https://github.com/simonvh/gimmemotifs/) and documentation is available at https://gimmemotifs.readthedocs.io. Functionality is covered by unit tests, which are run through continuous integration. GimmeMotifs can be installed via bioconda [44], see https://

bioconda.github.io/ for details. All releases are also distributed through PyPi [45] and stably archived using Zenodo [46]. For *de novo* motif search, 14 different external tools are supported (Table 1). All of these are installed when conda is used for installation. By default, genomepy is used for genome management [47]. In addition, GimmeMotifs uses the following Python modules: numpy [48], scipy [49], scikit-learn, scikit-contrib-lightning [31], seaborn [50], pysam [51,52], xgboost [32] and pandas. In addition to the command line tools, all GimmeMotifs functionality is available through a Python API.

De novo motif prediction pipeline

Originally, GimmeMotifs was developed to predict *de novo* motifs from ChIP-seq data using an ensemble of motif predictors [53]. The tools currently supported are listed in Table 1. An input file (BED, FASTA or narrowPeak format) is split into a prediction and validation set. The prediction set is used to predict motifs, and the validation set is used to filter for significant motifs. All significant motifs are clustered to provide a collection of non-redundant *de novo* motifs. Finally, significant clustered motifs are reported, along with several statistics to evaluate motif quality, calculated using the validation set. These evaluation metrics include ROC AUC, distribution of the motif location relative to the center of the input (i.e., the ChIP-seq peak summit) and the best match in a database of known motifs.

Table 1: External *de novo* motif prediction tools supported by GimmeMotifs.

Name	Citation
AMD	[54]
BioProspector	[55]
ChIPMunk	[56]
GADEM	[57]
HMS	[58]
Homer	[6]
Improbizer	[59]
MDmodule	[60]
MEME	[61]
MotifSampler	[62]
Posmo	[63]
Trawler	[64]
Weeder	[65]
XXmotif	[66]

Motif activity by ensemble learning: maelstrom

GimmeMotifs implements eight different methods to determine differential motif enrichment between two or more conditions. In addition, these methods can be combined in a single measure of *motif activity* using rank aggregation. Four methods work with discrete sets, such as different peak sets or clusters from a K-means clustering. The hypergeometric test uses motif counts with an empirical motif-specific FPR of 5%. All other implemented methods use the PFM log-odds score of the best match.

The hypergeometric test is commonly used to calculate motif enrichment, for instance by Homer [6]. In GimmeMotifs, motifs in each cluster are tested against the union of all other clusters. The reported value is -log10(p-value) where the p-value is adjusted by the Benjamini-Hochberg procedure [67].

Using the non-parametric Mann-Whitney U test, GimmeMotifs tests the null hypothesis that that the motif log-odds score distributions of two classes are equal. For each discrete class in the data, such as a cluster, it compares the score distributions of the class to the score distribution of all other classes. The value used as activity is the -log10 of the Benjamini-Hochberg adjusted p-value.

The two other methods are classification algorithms: random forest using scikit-learn and a large-scale multiclass classifier using block coordinate descent [31] as implemented in the scikit-contrib-lightning module. The classifier in GimmeMotifs uses a l1/l2 penalty with squared hinge loss where the alpha and C parameters are set using grid search in 10 fold cross-validation procedure.

The other four methods that are implemented relate motif score to an experimental measure such as ChIP-seq or ATAC-seq signal or expression level. These are all different forms of regression. In addition to ridge regression, which is similar to Motif Activity Response Analysis (MARA) [28,29], these methods include regression using boosted trees (XGBoost [32]), multiclass regression [31] and L1 regularized regression (LASSO).

To combine different measures of motif significance or activity into a single score, ranks are assigned for each individual method and combined using rank aggregation based on order statistics [33]. This results in a probability of finding a motif at all observed positions. We use a Python implemention based on the method used in the R package RobustRankAggreg [68]. The rank aggregation is performed twice, once with the ranks reversed to generate both positively and negatively associated motifs.

Clustering

Transcription factor motif database benchmark

We downloaded all ChIP-seq peaks from Remap 2018 v1.2 [17] (http://tagc.univ-mrs.fr/remap/index.php?page=download). We removed all factors with fewer than 1000 peaks and created regions of 100 bp centered at the peak summit. Background files were created for each peak set using bedtools shuffle [69], excluding the hg19 gaps and the peak regions. The ROC AUC and Recall at 10% FDR statistics were calculated using `gimme roc`. The motif databases included in the comparison are listed in Table 2. We only included public databases that can be freely accessed and downloaded.

Table 2: Motif databases.

Name	Version	Citation
Factorbook	Sep. 2012	[3]
SwissRegulon	Nov. 2006	[18]
GimmeMotifs vertebrate	v4.0	
HOCOMOCO	v11	[5]
Homer	v4.10	[6]
JASPAR	2018	[14]
Kheradpour	Dec. 2013	[4]
Madsen	1.1	[20]
RSAT vertebrate clusters	Sep. 2017	[19]

The workflow is implemented in snakemake [70] and is available at https://github.com/simonvh/gimme_analysis.

`De novo` motif prediction benchmark

We downloaded all spp ENCODE peaks (January 2011 data freeze) from the EBI FTP (http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/peaks/jan2011/spp/optimal/). We selected the top 5000 peaks and created 100bp regions centered on the peak summit. As background we selected 100 bp regions flanking the original peaks. For the `de novo` motif search default settings for `gimme motifs` were used. The workflow is implemented in snakemake [70] and is available at https://github.com/simonvh/gimme_analysis.

Motif analysis of hematopoietic enhancers

To illustrate the functionality of `gimme maelstrom` we analyzed an integrated collection of hematopoietic enhancers. We downloaded all H3K27ac ChIP-seq and DNase I data from BLUEPRINT and hematopoietic DNase I data from ROADMAP (Supplementary Table S1). All DNase I data were processed using the Kundaje lab DNase pipeline version 0.3.0 https://github.com/kundajelab/atac_dnase_pipelines [71]. The ChIP-seq samples were processed using the Kundaje lab AQUAS TF and histone ChIP-seq pipeline https://github.com/kundajelab/chipseq_pipeline. For all experiments from BLUEPRINT we used the aligned reads provided by EBI. All ROADMAP samples were aligned using bowtie2 [72] to the hg38 genome. DNase I peaks

were called using MACS2 [73]. We merge all DNase I peak files and centered each merged peak on the summit of the strongest individual peak. H3K27ac reads were counted in a region of 2kb centered at the summit (Supplementary Table S2) and read counts were log2-transformed and scaled. We removed all samples that were treated and averaged all samples from the same cell type. We then selected all enhancers with at least one sample with a scaled log2 read count of 2, sorted by the maximum difference in normalized signal between samples and selected the 50,000 enhancers with the largest difference. Using this enhancer collection as input, we ran `gimme maelstrom` using default settings. The motif analysis workflow is implemented in a Jupyter notebook and is available at https://github.com/simonvh/gimme_analysis.

Conclusions

We demonstrated the functionality of GimmeMotifs with three examples. First, to evaluate different public motif databases, we quantified their performance on distinguishing ChIP-seq peaks from background sequences. The databases that perform best on this benchmark are collections of motifs from different sources. Of the individual databases HOCOMOCO and Factorbook rank highest using this collection of human ChIP-seq peaks. Based on our results it is recommended to use a composite database, such as the RSAT clustered motifs or the GimmeMotifs database (v4.0), for the best vertebrate motif coverage. However, these motifs are less well annotated. For instance, motifs based on ChIP-seq peaks from some sources might be from co-factors or cell type-specific regulators instead of the factor that was assayed. An example are motifs that associated with the histone acetyl tranferase EP300. This transcriptional co-activator lacks a DNA binding domain, and associated motifs depend on the cell type. For instance, in a lymphoblastoid cell line such as GM12878 these include PU.1 and AP1. The lack of high-quality annotation makes it more difficult to reliably link motifs to transcription factors. This can be an advantage of using JASPAR. Although the motifs might not be optimal, JASPAR contains high-quality metadata that is manually curated.

In the second example, we benchmarked 14 different *de novo* motif finders using a large compendium of ChIP-seq data. While performance can vary between different data sets, there are several *de novo* motif finders that consistently perform well, with BioProspector, MEME and Homer as top performers. Interestingly, only Homer was specifically developed for ChIP-seq data. An ensemble approach such as GimmeMotifs still improves on the use of individual tools. This example also illustrates that newly developed *de novo* motif finders should be evaluated on many different data sets, as this is necessary to accurately judge the performance.

Finally, we presented a new ensemble approach, *maelstrom*, to determine motif activity in two or more epigenomic or transcriptomic data sets. Using H3K27ac ChIP-seq signal as a measure for enhancer activity, we analyzed cell-type specific motif activity in a large collection of hematopoietic cell types. We identified known lineage regulators, as well as motifs for factors that are less well

studied in a hematopoietic context. This illustrates how `gimme maelstrom` can serve to identify cell type-specific transcription factors.

In conclusion, GimmeMotifs is a flexible and highly versatile framework for transcription factor motif analysis. Both command line and programmatic use in Python are supported. One planned future improvement to GimmeMotifs is the support of more sophisticated motif models. Although PFMs are very informative, it is clear that they represent an oversimplification of TF binding preferences. While several approaches that incorporate positional dependencies have been developed, it is still unclear how well these models perform and their use depends on specific tools. Supporting these different models and benchmarking their performance relative to high-quality PFMs will simplify their use and give insight into their benefits and disadvantages. Second, there is significant progress recently in modeling TF binding using deep neural networks (DNNs) [74,75]. These DNNs can learn sequence motifs, as well as complex inter-dependencies, directly from the data. However, while biological interpretation is possible [76], it becomes less straightforward. We expect that analyzing and understanding a trained DNN can benefit from high-quality motif databases and comparative tools such as GimmeMotifs.

Availability and requirements

- Project name: GimmeMotifs
- Project home page: https://github.com/simonvh/gimmemotifs
- Operating system(s): Linux, Mac OSX
- Programming language: Python 3
- Other requirements: *de novo* motif finders
- License: MIT

Availability of supporting data

Additional files

Competing interests

The authors declare that they have no competing interests.

Funding

SJvH was supported by the Netherlands Organization for Scientific research (NWO-ALW, grant 863.12.002). Part of this work was carried out on the Dutch national e-infrastructure with the

support of SURF Foundation. This work was sponsored by NWO Exact and Natural Sciences for the use of supercomputer facilities.

Acknowledgements

References

1. **The Human Transcription Factors**
Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, Matthew T. Weirauch
Cell (2018-02) https://doi.org/10.1016/j.cell.2018.01.029

2. **Modeling the specificity of protein-DNA interactions**
Gary D. Stormo
Quantitative Biology (2013-04-02) https://doi.org/10.1007/s40484-013-0012-4

3. **Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium**
J. Wang, J. Zhuang, S. Iyer, X.-Y. Lin, M. C. Greven, B.-H. Kim, J. Moore, B. G. Pierce, X. Dong, D. Virgil, … Z. Weng
Nucleic Acids Research (2012-11-29) https://doi.org/10.1093/nar/gks1221

4. **Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments**
P. Kheradpour, M. Kellis
Nucleic Acids Research (2013-12-13) https://doi.org/10.1093/nar/gkt1249

5. **HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis**
Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, … Vsevolod J Makeev
Nucleic Acids Research (2017-11-11) https://doi.org/10.1093/nar/gkx1106

6. **Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities**
Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, Christopher K. Glass
Molecular Cell (2010-05) https://doi.org/10.1016/j.molcel.2010.05.004

7. **Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities**
A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpaa, … J. Taipale
Genome Research (2010-04-08) https://doi.org/10.1101/gr.100552.109

8. **UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions**
Maxwell A. Hume, Luis A. Barrera, Stephen S. Gisselbrecht, Martha L. Bulyk
Nucleic Acids Research (2014-11-05) https://doi.org/10.1093/nar/gku1045

9. **FROM BINDING MOTIFS IN CHIP-SEQ DATA TO IMPROVED MODELS OF TRANSCRIPTION FACTOR BINDING SITES**
IVAN KULAKOVSKIY, VICTOR LEVITSKY, DMITRY OSHCHEPKOV, LEONID BRYZGALOV, ILYA VORONTSOV, VSEVOLOD MAKEEV
Journal of Bioinformatics and Computational Biology (2013-02) https://doi.org/10.1142/s0219720013400040

10. **The Next Generation of Transcription Factor Binding Site Prediction**
Anthony Mathelier, Wyeth W. Wasserman
PLoS Computational Biology (2013-09-05) https://doi.org/10.1371/journal.pcbi.1003214

11. **Varying levels of complexity in transcription factor binding motifs**
Jens Keilwagen, Jan Grau
Nucleic Acids Research (2015-06-26) https://doi.org/10.1093/nar/gkv577

12. **InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites**
Ralf Eggeling, Ivo Grosse, Jan Grau
Bioinformatics (2016-12-28) https://doi.org/10.1093/bioinformatics/btw689

13. **Automated incorporation of pairwise dependency in transcription factor binding site prediction using dinucleotide weight tensors**
Saeed Omidi, Mihaela Zavolan, Mikhail Pachkov, Jeremie Breda, Severin Berger, Erik van Nimwegen
PLOS Computational Biology (2017-07-28) https://doi.org/10.1371/journal.pcbi.1005176

14. **JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework**
Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R Kulkarni, Ge Tan, … Anthony Mathelier
Nucleic Acids Research (2017-11-13) https://doi.org/10.1093/nar/gkx1126

15. **JASPAR RESTful API: accessing JASPAR data from any programming language**
Aziz Khan, Anthony Mathelier
Bioinformatics (2017-12-15) https://doi.org/10.1093/bioinformatics/btx804

16. **Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity**
Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, … Timothy R.

Hughes
Cell (2014-09) https://doi.org/10.1016/j.cell.2014.08.009

17. **ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments**
Jeanne Chèneby, Marius Gheorghe, Marie Artufel, Anthony Mathelier, Benoit Ballester
Nucleic Acids Research (2017-11-08) https://doi.org/10.1093/nar/gkx1092

18. **SwissRegulon: a database of genome-wide annotations of regulatory sites**
M. Pachkov, I. Erb, N. Molina, E. van Nimwegen
Nucleic Acids Research (2007-01-03) https://doi.org/10.1093/nar/gkl857

19. **RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections**
Jaime Abraham Castro-Mondragon, Sébastien Jaeger, Denis Thieffry, Morgane Thomas-Chollier, Jacques van Helden
Nucleic Acids Research (2017-06-07) https://doi.org/10.1093/nar/gkx314

20. **Integrated analysis of motif activity and gene expression changes of transcription factors**
Jesper Grud Skat Madsen, Alexander Rauch, Elvira Laila Van Hauwaert, Søren Fisker Schmidt, Marc Winnefeld, Susanne Mandrup
Genome Research (2017-12-12) https://doi.org/10.1101/gr.227231.117

21. **ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge - Dream Challenges***Dream Challenges* http://dreamchallenges.org/project/encode-dream-in-vivo-transcription-factor-binding-site-prediction-challenge/

22. **Assessing computational tools for the discovery of transcription factor binding sites**
Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, … Zhou Zhu
Nature Biotechnology (2005-01) https://doi.org/10.1038/nbt1053

23. **Evaluating tools for transcription factor binding site prediction**
Narayan Jayaram, Daniel Usvyat, Andrew C. R. Martin
BMC Bioinformatics (2016-11-02) https://doi.org/10.1186/s12859-016-1298-9

24. **Limitations and potentials of current motif discovery algorithms**
J. Hu, B. Li, D. Kihara
Nucleic Acids Research (2005-09-02) https://doi.org/10.1093/nar/gki791

25. **Index of /pub/databases/ensembl/encode/integration_data_jan2011/byDataType/peaks/jan2011/spp/optimal/hub**http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/peaks/jan2011/spp/optimal/hub/

26. **Rank order metrics for quantifying the association of sequence features with gene regulation.**
Neil D Clarke, Joshua A Granek
Bioinformatics (Oxford, England) (2003-01-22) https://www.ncbi.nlm.nih.gov/pubmed/12538241

27. **Differential motif enrichment analysis of paired ChIP-seq experiments**
Tom Lesluyes, James Johnson, Philip Machanick, Timothy L Bailey
BMC Genomics (2014) https://doi.org/10.1186/1471-2164-15-752

28. **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line**
Harukazu SuzukiAlistair RR Forrest, Erik van Nimwegen, Carsten O Daub, Piotr J Balwierz, Katharine M Irvine, Timo Lassmann, Timothy Ravasi, Yuki Hasegawa, …
Nature Genetics (2009-04-19) https://doi.org/10.1038/ng.375

29. **ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs**
P. J. Balwierz, M. Pachkov, P. Arnold, A. J. Gruber, M. Zavolan, E. van Nimwegen
Genome Research (2014-02-10) https://doi.org/10.1101/gr.169508.113

30. **BLUEPRINT: mapping human blood cell epigenomes**
J. H. A. Martens, H. G. Stunnenberg
Haematologica (2013-10-01) https://doi.org/10.3324/haematol.2013.094243

31. **Block coordinate descent algorithms for large-scale sparse multiclass classification**
Mathieu Blondel, Kazuhiro Seki, Kuniaki Uehara
Machine Learning (2013-05-08) https://doi.org/10.1007/s10994-013-5367-2

32. **XGBoost**
Tianqi Chen, Carlos Guestrin
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 (2016) https://doi.org/10.1145/2939672.2939785

33. **Gene prioritization through genomic data fusion**
Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, … Yves Moreau
Nature Biotechnology (2006-05) https://doi.org/10.1038/nbt1203

34. **PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors.**
C Nerlov, T Graf
Genes & development (1998-08-01) https://www.ncbi.nlm.nih.gov/pubmed/9694804

35. **Transcriptional control of granulocyte and monocyte development**
AD Friedman
Oncogene (2007-10) https://doi.org/10.1038/sj.onc.1210764

36. **Fatal autoimmunity results from the conditional deletion of Snai2 and Snai3**
Peter D. Pioli, Xinjian Chen, Janis J. Weis, John H. Weis
Cellular Immunology (2015-05) https://doi.org/10.1016/j.cellimm.2015.02.009

37. **Snai2 and Snai3 transcriptionally regulate cellular fitness and functionality of T cell lineages through distinct gene programs**
Peter D. Pioli, Sarah K. Whiteside, Janis J. Weis, John H. Weis
Immunobiology (2016-05) https://doi.org/10.1016/j.imbio.2016.01.007

38. **The transcription repressor, ZEB1, cooperates with CtBP2 and HDAC1 to suppress IL-2 gene activation in T cells**
J. Wang, S. Lee, C. E.-Y. Teh, K. Bunting, L. Ma, M. F. Shannon
International Immunology (2009-01-15) https://doi.org/10.1093/intimm/dxn143

39. **Negative regulation of CD4 expression in T cells by the transcriptional repressor ZEB.**
T Brabletz, A Jung, F Hlubek, C Löhberg, J Meiler, U Suchy, T Kirchner
International immunology (1999-10) https://www.ncbi.nlm.nih.gov/pubmed/10508188

40. **T-cell expression of the human GATA-3 gene is regulated by a non-lineage-specific silencer.**
JM Grégoire, PH Roméo
The Journal of biological chemistry (1999-03-05) https://www.ncbi.nlm.nih.gov/pubmed/10037751

41. **DeltaEF1, a zinc finger and homeodomain transcription factor, is required for skeleton patterning in multiple lineages.**
T Takagi, H Moribe, H Kondoh, Y Higashi
Development (Cambridge, England) (1998-01) https://www.ncbi.nlm.nih.gov/pubmed/9389660

42. **Nanog and transcriptional networks in embryonic stem cell pluripotency**
Guangjin Pan, James A Thomson
Cell Research (2007-01) https://doi.org/10.1038/sj.cr.7310125

43. **NANOG induction of fetal liver kinase-1 (FLK1) transcription regulates endothelial cell proliferation and angiogenesis**
E. E. Kohler, C. E. Cowan, I. Chatterjee, A. B. Malik, K. K. Wary
Blood (2010-11-30) https://doi.org/10.1182/blood-2010-07-295261

44. **Bioconda: sustainable and comprehensive software distribution for the life sciences**
Björn GrüningRyan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-

Tinch, Renan Valieris, Johannes Köster
Nature Methods (2018-07) https://doi.org/10.1038/s41592-018-0046-7

45. **PyPI – the Python Package Index***PyPI* https://pypi.org/

46. **Zenodo - Research. Shared.**https://zenodo.org/

47. **genomepy: download genomes the easy way**
Simon J. van Heeringen
The Journal of Open Source Software (2017-08-15) https://doi.org/10.21105/joss.00320

48. **NumPy — NumPy**(2018-04-26) http://www.numpy.org/

49. **SciPy.org — SciPy.org**(2018-08-15) http://www.scipy.org/

50. **Mwaskom/Seaborn: V0.9.0 (July 2018)**
Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Joel Ostblom, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, … Adel Qalieh
Zenodo (2018-07-16) https://doi.org/10.5281/zenodo.1313201

51. **pysam-developers/pysam**
pysam-developers
GitHub https://github.com/pysam-developers/pysam

52. **The Sequence Alignment/Map format and SAMtools.**
Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin,
Bioinformatics (Oxford, England) (2009-06-08) https://www.ncbi.nlm.nih.gov/pubmed/19505943

53. **GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments**
Simon J. van Heeringen, Gert Jan C. Veenstra
Bioinformatics (2010-11-15) https://doi.org/10.1093/bioinformatics/btq636

54. **AMD, an Automated Motif Discovery Tool Using Stepwise Refinement of Gapped Consensuses**
Jiantao Shi, Wentao Yang, Mingjie Chen, Yanzhi Du, Ji Zhang, Kankan Wang
PLoS ONE (2011-09-12) https://doi.org/10.1371/journal.pone.0024576

55. **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.**
X Liu, DL Brutlag, JS Liu
Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (2001) https://www.ncbi.nlm.nih.gov/pubmed/11262934

56. **Deep and wide digging for binding motifs in ChIP-Seq data**
I. V. Kulakovskiy, V. A. Boeva, A. V. Favorov, V. J. Makeev
Bioinformatics (2010-10-15) https://doi.org/10.1093/bioinformatics/btq488

57. **GADEM: A Genetic Algorithm Guided Formation of Spaced Dyads Coupled with an EM Algorithm for Motif Discovery**
Leping Li
Journal of Computational Biology (2009-02) https://doi.org/10.1089/cmb.2008.16tt

58. **On the detection and refinement of transcription factor binding sites using ChIP-Seq data**
Ming Hu, Jindan Yu, Jeremy M. G. Taylor, Arul M. Chinnaiyan, Zhaohui S. Qin
Nucleic Acids Research (2010-01-07) https://doi.org/10.1093/nar/gkp1180

59. **cis-Site Seeker**https://users.soe.ucsc.edu/~kent/improbizer/index.html

60. **An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments**
X. Shirley Liu, Douglas L. Brutlag, Jun S. Liu
Nature Biotechnology (2002-07-08) https://doi.org/10.1038/nbt717

61. **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.**
TL Bailey, C Elkan
Proceedings. International Conference on Intelligent Systems for Molecular Biology (1994) https://www.ncbi.nlm.nih.gov/pubmed/7584402

62. **A Gibbs Sampling Method to Detect Overrepresented Motifs in the Upstream Regions of Coexpressed Genes**
Gert Thijs, Kathleen Marchal, Magali Lescot, Stephane Rombauts, Bart De Moor, Pierre Rouzé, Yves Moreau
Journal of Computational Biology (2002-04) https://doi.org/10.1089/10665270252935566

63. **A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information**
Xiaotu Ma, Ashwinikumar Kulkarni, Zhihua Zhang, Zhenyu Xuan, Robert Serfling, Michael Q. Zhang
Nucleic Acids Research (2011-01-06) https://doi.org/10.1093/nar/gkr1135

64. **Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation**
Laurence Ettwiller, Benedict Paten, Mirana Ramialison, Ewan Birney, Joachim Wittbrodt
Nature Methods (2007-06-24) https://doi.org/10.1038/nmeth1061

65. **Using Weeder, Pscan, and PscanChIP for the Discovery of Enriched Transcription Factor Binding Site Motifs in Nucleotide Sequences**
Federico Zambelli, Graziano Pesole, Giulio Pavesi
Current Protocols in Bioinformatics (2014-09) https://doi.org/10.1002/0471250953.bi0211s47

66. **P-value-based regulatory motif discovery using positional weight matrices**
H. Hartmann, E. W. Guthohrlein, M. Siebert, S. Luehr, J. Soding
Genome Research (2012-09-18) https://doi.org/10.1101/gr.139881.112

67. https://www.jstor.org/stable/2346101

68. **Robust rank aggregation for gene list integration and meta-analysis**
Raivo Kolde, Sven Laur, Priit Adler, Jaak Vilo
Bioinformatics (2012-01-12) https://doi.org/10.1093/bioinformatics/btr709

69. **BEDTools: a flexible suite of utilities for comparing genomic features**
Aaron R. Quinlan, Ira M. Hall
Bioinformatics (2010-01-28) https://doi.org/10.1093/bioinformatics/btq033

70. **Snakemake–a scalable bioinformatics workflow engine**
J. Koster, S. Rahmann
Bioinformatics (2012-08-20) https://doi.org/10.1093/bioinformatics/bts480

71. **Kundajelab/Atac_Dnase_Pipelines: 0.3.0**
Jin Lee, Grey Christoforo, Grey Christoforo, CS Foo, Chris Probert, Anshul Kundaje, Nathan Boley, Kohpangwei, Daniel Kim, Mike Dacre
Zenodo (2016-09-27) https://doi.org/10.5281/zenodo.156534

72. **Fast gapped-read alignment with Bowtie 2**
Ben Langmead, Steven L Salzberg
Nature Methods (2012-03-04) https://doi.org/10.1038/nmeth.1923

73. **Model-based Analysis of ChIP-Seq (MACS)**
Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nussbaum, Richard M Myers, Myles Brown, Wei Li, X Shirley Liu
Genome Biology (2008) https://doi.org/10.1186/gb-2008-9-9-r137

74. **Predicting effects of noncoding variants with deep learning–based sequence model**
Jian Zhou, Olga G Troyanskaya
Nature Methods (2015-08-24) https://doi.org/10.1038/nmeth.3547

75. **Sequential regulatory activity prediction across chromosomes with convolutional neural networks**

David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, Jasper Snoek
Genome Research (2018-03-27) https://doi.org/10.1101/gr.227819.117

76. **Discovering epistatic feature interactions from neural network models of regulatory DNA sequences**
Peyton Greenside, Tyler Shimko, Polly Fordyce, Anshul Kundaje
Bioinformatics (2018-09-01) https://doi.org/10.1093/bioinformatics/bty575