IMPROVED ENSPART FOR DNA MOTIF PREDICTION

Allen Chieng-Hoon Choong
Universiti Malaysia Sarawak

Nung-Kion Lee♣
Universiti Malaysia Sarawak

Chih-How Bong
Universiti Malaysia Sarawak

Norshafrina Omar
Universiti Malaysia Sarawak

ABSTRACT

In our previous work we proposed ENSPART-an ensemble method for DNA motif discovery which partitions input dataset into several equal size subsets runs by several distinct tools for candidate motif prediction. The candidate motifs obtained from different data subsets are merged to obtain the final motifs. Nevertheless, the original ENSPART has several limitations: (1) the same background sequences are used for the calculation of Receiver Operating Cost (ROC) of motifs obtained from different datasets. This causes bias because different datasets might have different background distribution; (2) it does not consider the duplication of a motif and its reverse complement. This causes many redundant motifs in the result set which requires filtering. In this work, we extended the original ENSPART to solve those two issues. For the first issue, we employed background sequences that is based on the distribution of bases in the input sequences. As for the second issue, we employ a "triple" merging strategy to reduce redundant motifs. The evaluation results indicate that the two improvements obtain better AUC values in comparison to the original implementation.

Keywords: DNA Motif Discovery; Machine Learning; Ensemble.

1. INTRODUCTION

ENSPART (Lee, Choong, & Omar, 2016) is an ensemble approach which utilizes an ensemble of 7 motif discovery tools for motif prediction. It is designed for tackling large-scale ChIP dataset for the discovery of primary motifs in a DNA dataset enriched with motifs. The idea of ENSPART is to partition a large-scale ChIP dataset into small subsets and use an ensemble of motif discovery tools for motif prediction in each subset. The assumption is the binding sites of a primary transcription factor protein is abundance in each of the partitioned subset and thus can be predicted by motif discovery tools independently. Furthermore, utilizing many tools for prediction would increase the chances of obtaining true motifs. The tools run on each partitioned dataset for motif discovery and predicted motifs from individual tool are merged to produce the final motifs. An

♣ Corresponding author: Nung-Kion Lee, Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia. Email: nklee@unimas.my

alignment free method is employed to merge motifs obtained from different data subsets to reduce redundancy and groups similar motifs. The merging managed to reduce about 49 to 55% of the motifs produced for all the evaluated datasets. The receiver operating curve (ROC) is used to rank the candidate motifs before the final motifs selection. Our previous simulation results demonstrated ENSPART good performance in comparison to MEME.

Nevertheless, the original implementation of ENSPART has several noticeable weaknesses:

- The calculation of ROC used for ranking of candidate motifs require a set of background sequences which does not contain the motifs. In our implementation, we employed the same background sequences for the computation of ROC for the ranking of final motifs from different datasets. This could be biased since it is not guaranteed the background sequences do not contain motifs.
- The existing merging method does not consider the similarity between the motifs in the forward and reverse complement. There could be many redundant motifs due to that.

In this paper, the improvements over the original ENSPART by addressing the two issues above will be presented.

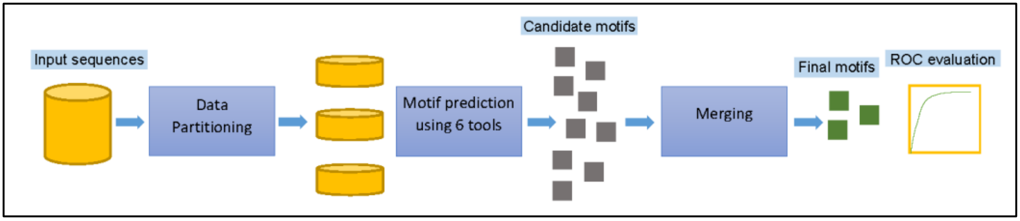
This paper is organized as follows. Section 2 provides background of DNA motif prediction problem and ENSPART algorithm. Section 3 presents the modifications proposed on the original ENSPART algorithm. Section 4 gives the evaluation results of ENSPART using real datasets. The last section is discussion and conclusion of this study.

2. BACKGROUND

DNA motif discovery can be formulated as a multiple-local-alignment or consensus pattern enumeration problem. Given an input DNA dataset which is enriched with motifs, the multiple-local-alignment is to find optimal multiple alignment of fixed length short sequence segments, typically assuming zero, one, or multiple from each input sequences, that optimizes an objective function. While consensus pattern enumeration method finds over-represented consensus patterns in an input dataset. It does that by enumerating all possible consensuses of a certain length and determine which consensus patterns are over-represented in the input dataset as contrast to the background sequences. The enumeration can be exhaustive or heuristic in nature.

Ensemble approach for DNA motif discovery involves utilization of several motif prediction tools for motif discovery and combine their results by filtering or merging. For a comprehensive review of existing ensemble approach for DNA motif discovery readers can refer to (Lihu & Holban, 2015). Most of the existing ensemble methods use the whole input dataset for prediction by multiple motif discovery tools and then merged similar motifs using clustering or heuristics. Likewise, ENSPART is an ensemble motif prediction algorithm which employs several motif discovery tools for prediction. To enable the tools executes on large-scale dataset, the dataset is first partitioned into non-overlapped subsets before tackling each separately by different tools. An important step then is to merge similar motifs (possible partially) predicted by different tools from different subsets. Figure 1 shows the ENSPART computational pipeline.

Figure 1: ENSPART framework



The alignment-free method which uses the k-mer frequency vector (KFV) (Xu & Su, 2010) to represent each motif is used for determining similar motifs that can be merged. The merging is the key-step in ENSPART since it determines the number of final motifs and the motif quality. Key decisions that need to be made in the merging step is how to determine which pair of motifs to merge, the similarity merging threshold, and the order they are merged. Likewise, it is essential to decide how many iterations the motifs should be merged. Once the final motifs are obtained from the merging process, the ROC measure is used to rank them for final selection.

3. METHOD

3.1. Datasets

Ten ChIP datasets which previously employed in (Lee et al., 2016) is used in this study. Table 1 shows the information on the datasets.

Table 1: Information on Datasets Used for the Benchmark

Dataset	No of sequences	Average length of sequence	File size
E2F4	128	543	76K
OCT4 Ntera	154	553	92K
p53	542	1186	669.9k
NRSF	1657	283	528K
FoxA1	2119	357	828K
CREB	2342	1141	3.3M
FoxA2	4051	218	1.3M
OCT4	7776	461	4.7M
CTCF	13804	816	12.7M
STAT1	27470	246	8.1M

3.2. Motif Discovery

The ENSPART algorithm is similar to our early work (Lee et al., 2016) and it is described briefly here. The aim of ENSPART algorithm is to predict the primary motifs (top three) in an input dataset. An input dataset is partitioned into three subsets which will be fed to the seven motif discovery tools. Each tool runs **three (3) times** with different parameters to increase the possibility of discovering true motifs since they have different characteristics such as lengths, conservation, or abundance.

Seven popular motif discovery tools were selected to perform candidate motif prediction.

AlignACE (Roth, Hughes, Estep, & Church, 1998), BioProspector (Liu, Brutlag, & Liu, 2001), and MEME-ChIP (Machanick & Bailey, 2011) are local search algorithms, MDscan (Liu, Brutlag, & Liu, 2002) is enumeration and heuristic search. Weeder (Pavesi, Mauri, & Pesole, 2001) and AMD (Shi et al., 2011) are pattern enumeration algorithms. With the collection of tools having different strengths, it increases the chances of predicting distinct motif characteristics. Each tool ran three times with the parameters specified in our previous work (Lee et al., 2016). For comparing ENSPART and non-partitioned method, we also ran MEME-ChIP, ChIPMunk (Kulakovskiy, Boeva, Favorov, & Makeev, 2010) and RSAT peak-motif (Thomas-Chollier et al., 2012) on the whole dataset.

In this study, each input dataset was partitioned into 3 non-overlaps partitions with each partition size is 10% of the whole dataset.

3.3. Calculation of ROC

To evaluate the quality of the merged motifs, ROC of the motifs is plotted and its Area Under Curve (AUC) value is computed. ROC is a standard method for evaluating a motif's quality. The ROC is a popular evaluation method which has been used for evaluation of tools such as MEME (Bailey & Elkan, 1995), GAPWM(Li, Liang, & Bass, 2007), and GimmeMotifs (van Heeringen & Veenstra, 2011).

The plotting of the ROC in this study was done by using the **rocpwm** program bundled with the GAPWM tool (Li et al., 2007). **rocpwm** by default can only accept a maximum of 500 input sequences. To cater for the larger sizes datasets, it is modified to allow input of 30k DNA sequences. The **rocpwm** tool requires a dataset that contains the true motifs and a dataset that contains the background sequences.

To prepare the background sequences, the tool "fasta-dinucleotide-shuffle" from the MEME Suite (Bailey et al., 2009) was used. According to (Zeng, Edwards, Liu, & Gifford, 2016) the tool offers more accurate background sequences for evaluation purpose. For each input dataset, its corresponding background sequences were generated using the tool. In the original implementation (Lee et al., 2016), a background dataset as used by Amadeus(Linhart, Halperin, & Shamir, 2008) was used for all the evaluation datasets for the computations of ROC. However, such negative dataset is not guarantee to exclude sequences in the positive datasets.

With "fasta-dinucleotide-shuffle" tool, the background sequences are generated from each input dataset which assures they are statistically negative to the target dataset. As a result, the ROC will be more reliable than using the same background across the different input datasets.

3.4. Merging

The original merging algorithm as implemented in ENSPART was used. Nevertheless, in this study, the merging is repeated three times consecutively. The justification is that multiple merging would find more similar motifs that can be grouped. In this improved version, we also consider the forward and reverse complement of the motifs during merging.

4. RESULTS WITH DISCUSSION

We ran the seven motif prediction tools on three partitioned datasets. The number of motifs predicted by each tool on each dataset is shown in Table 2.

Table 2: Number of Motifs Discovered from the Partitioned Datasets (3 Subsets)

Datasets	MDscan	BioProspector	Weeder2	MEME-ChIP	AMD	AlignACE	W-AlignACE
CREB	30	45	172	45	55	0	0
CTCF	30	45	174	45	55	0	0
E2F4	30	45	161	45	55	111	262
FOXA1	30	45	152	45	66	0	496
FOXA2	30	45	165	45	42	0	325
OCT4 Ntera	30	45	156	45	81	111	165
NRSF	30	45	169	45	25	3	290
OCT4	30	45	168	45	61	0	43
P53	30	45	171	45	50	28	534
STAT1	30	45	165	45	60	0	0

Note: *This table is reproduced from (Lee et al., 2016)

Table 3 shows the number of motifs after triple-merging. Table 3 shows that, after multiple merging, the numbers of candidate motifs have reduced significantly of approximately 79 to 83% for all datasets. This indicates that large numbers of the candidate motifs are redundant. The triple merging has greatly reduced the redundancies. We found repeating three merging steps is sufficient to obtain improved results.

To investigate how the multiple merging would affect the motif quality, the ROCs of the best three motifs from four of the datasets are shown in Figure 2. It is noticed that the curves are near to each other. It implied that the best three motifs are very similar, which potentially can be merged further. The top three motifs from the E2F4 and CREB dataset have rather weak discriminative hits against positive and negative input sequences. While for the CTCF and FOXA1 dataset, the top three motifs produced have good discrimination between the positive and negative dataset. Figure 3 shows the sequence logos of the top three motifs obtained from the CREB dataset. They appear to be variations of the same motif.

Table 3: Number of Motifs after Each Successive Merging.

Dataset	Before merging	1st	2nd	3rd	Reduced (%)
CREB	347	190	113	73	79.0
CTCF	349	181	109	68	80.5
E2F4	709	375	216	138	80.5
FOXA1	834	450	255	139	83.03
FOXA2	652	348	193	116	82.2
NRSF	607	322	181	107	82.4
OCT4 NTERA	633	336	194	114	82.0
OCT4	392	191	109	71	81.9
P53	903	483	273	161	82.2
STAT1	345	170	102	64	81.4

Figure 2: Best Three ROCs from the Ensemble Approach.

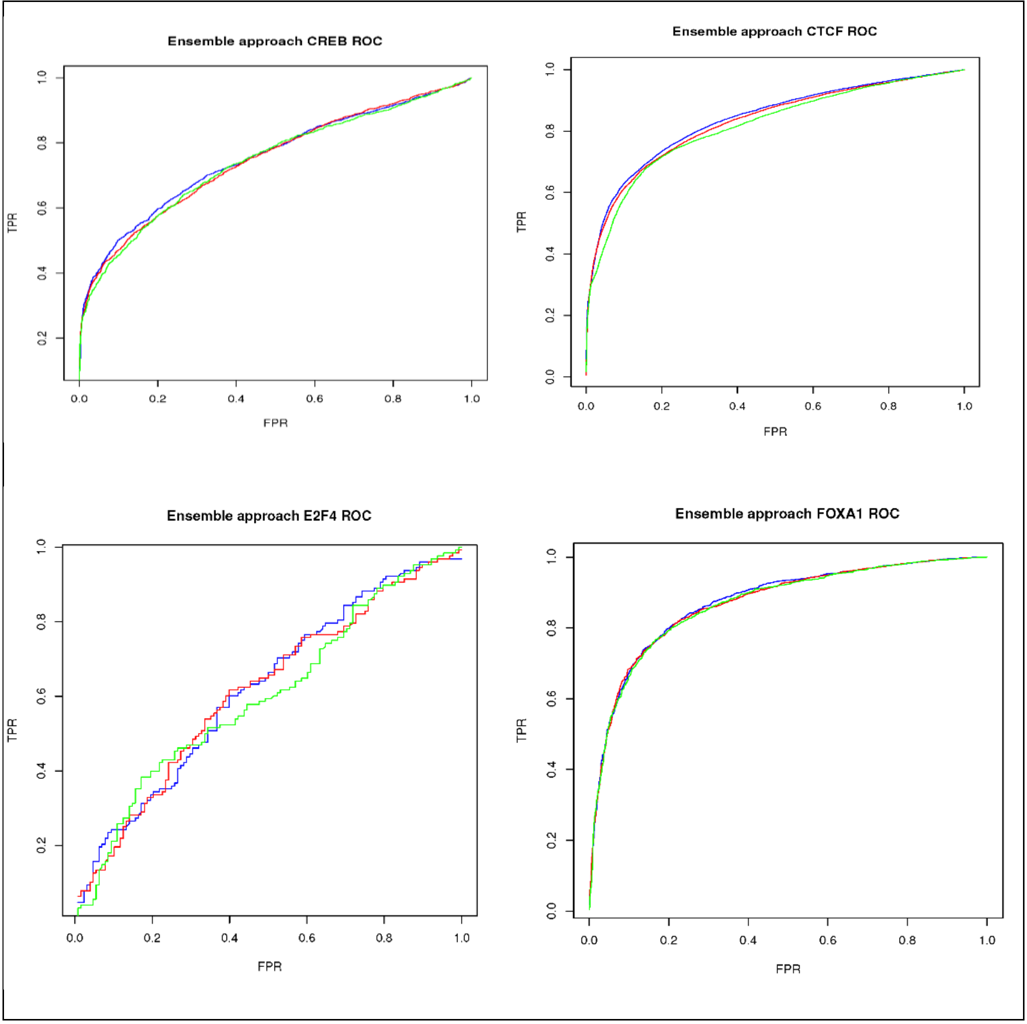
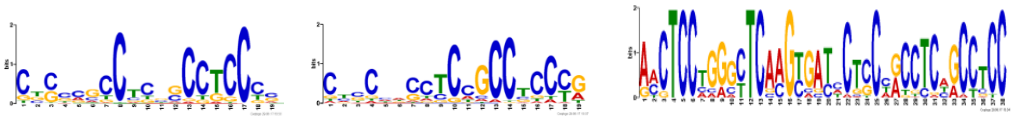


Table 4 shows the best Area Under Curve (AUC), worst AUCs, median, average of AUCs, and standard deviation from the proposed ensemble approach. The AUC values are calculated based on the number of motifs predicted after merging (see Table 3).

Figure 3: The Sequence Logos of the Top 3 Motifs Predicted from the CREB Dataset.



(a) (b) (c)

Table 4: Best AUCs, Worst AUCs, Median, Average AUC, and Standard Deviation of the Discovered Motifs Using the Ensemble Approach with Partitions.

TF	No of motifs	Best AUC	Worst AUC	Median	Average AUC*	Std Dev
CREB	73	0.7537	0.3704	0.6069	0.5941	0.0966
CTCF	68	0.8384	0.3975	0.5612	0.5642	0.1065
E2F4	138	0.6191	0.4126	0.5261	0.5223	0.0421
FOXA1	139	0.8721	0.4934	0.5815	0.6127	0.0877
FOXA2	116	0.8643	0.4727	0.5685	0.6119	0.1038
NRSF	107	0.7532	0.4816	0.4459	0.6040	0.0802
OCT4 NTERA	114	0.6670	0.4459	0.5664	0.5651	0.0508
OCT4	71	0.6847	0.3695	0.5294	0.5290	0.0590
P53	161	0.9499	0.4137	0.6698	0.6564	0.1063
STAT1	64	0.6607	0.4871	0.5558	0.5654	0.0446

Note: *the average AUC is obtained from the average of all motifs produced after merging.

Table 4 shows that the average AUC values are comparatively much lower than the best AUCs. Comparing to the original ENSPART, the average AUC values obtained in this study are lower. This is because the original ENSPART performed the merging process only once, while in this study the candidate motifs are merged three times. Moreover, the merging algorithm used in ENSPART does not consider merging of a motif with its reverse complement. As a result, many redundant or similar candidate motifs are used for the AUC calculation.

MEME-ChIP outputs three motifs by default for each input dataset. For this experiment, the whole set of input sequences is used. In addition, the 30% of each dataset that was used for the ensemble approach was also used for motif prediction by MEME-ChIP, ChIPMunk, and RSAT peak-motifs. The results are shown in Table 5.

From Table 5, it demonstrates that the best AUC values of ENSPART are better than MEME-ChIP (whole) on 9 out of 10 of the datasets, except for the CREB dataset. In addition, MEME-ChIP that uses 30% of the datasets has lower AUC values in comparison to both ENSPART and MEME-ChIP using the whole set. By comparing ENSPART with ChIPMunk, the latter has better AUC values on CREB, E2F4, P53, and STAT1 datasets. On the other hand, ENSPART obtained better AUC values in comparison to RSAT peak-motifs in all the datasets except NRSF and STAT1.

The findings of this study imply three important points:

- The performances of motif discovery using the partitioning and ensemble approach is a feasible way to tackle large-scale datasets. Its performance (in terms of the quality of motifs produced) is comparable to using the whole set for motif discovery task;
- The use of multiple-tools for DNA motif discovery increased the chances of discovery more true motifs considering the different characteristics of the motifs enriched in the input datasets; and
- The merging is a key step in determining the quality of the motifs produced by the partitioning and ensemble method.

Table 5: Comparisons of the Best AUC and Average AUC between ENSPART, MEME-ChIP, ChIPMunk, and RSAT-peak-motifs.

	ENSPART^{+}		**MEME-ChIP (whole)**		**MEME-ChIP^{+}**		**ChIPMunk$^{+,*}$**	**RSAT-peak-motifs^{+}**		
TF	Best	Avg.	Best	Avg.	Best	Avg	Best	Best	Avg	No of motifs
CREB	0.7537	0.5941	0.7671	0.7042	0.6943	0.6656	**0.8054**	0.7306	0.6464	20
CTCF	**0.8384**	0.5642	0.8195	0.6804	0.6067	0.5694	0.8381	0.6720	0.6059	20
E2F4	0.6191	0.5223	0.6018	0.5627	0.5245	0.4813	**0.6525**	0.5954	0.5006	11
FOXA1	**0.8721**	0.6127	0.8717	0.6554	0.5966	0.5639	0.8665	0.8345	0.7390	20
FOXA2	**0.8643**	0.6119	0.8446	0.6666	0.5819	0.5438	0.8469	0.8308	0.7587	20
NRSF	**0.7532**	0.6040	0.7334	0.6610	0.5351	0.4988	0.7452	0.7653	0.6428	20
NTERA	**0.6670**	0.5651	0.6424	0.5791	0.6116	0.5717	0.6351	0.6026	0.5342	20
OCT4	**0.6847**	0.5290	0.5541	0.5181	0.6799	0.5762	0.6493	0.6011	0.5506	20
P53	0.9499	0.6564	0.9473	0.7747	0.7202	0.5639	**0.9515**	0.8358	0.6214	20
STAT1	0.6607	0.5654	0.6675	0.5580	0.5650	0.5420	**0.6732**	0.6653	0.5993	20

Notes: *ChIPMunk only produced a motif for every dataset, therefore there is no average. $^{+}$employed only 30% of the dataset used by ENSPART. The "**no of motifs**" is the total number of motifs produced by RSAT-peak-motifs.

5. CONCLUSION

In this study, we proposed two important modifications on the original ENSPART. The first is to use distinct background sequences set, generated based on the input dataset, for the computation of ROC points. The second is employed triple merging on the intermediate motif prediction results produced by multiple motif discovery tools. The benefit of these two modifications are striking. The results show that the AUC values are much improved in comparison to the original implementation benchmarked using ten datasets. The lesson learned from this study is that the merging step is a key consideration when designing ensemble algorithm for DNA motif discovery. In addition, ensemble technique with data partitioning is a feasible way for effective motif prediction and enabled the use of classic motif discovery tools (i.e. which was not designed for large-dataset), for large-scale motif prediction.

ACKNOWLEDGEMENT

This research is funded by the Ministry of Higher Education Malaysia, Fundamental Research Grant Scheme FRGS/SG03(01)/1134/2014(01).

REFERENCES

Bailey, T. L., & Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. In *International Conference on Intelligent Systems for Molecular Biology (ISMB)* (pp. 21–29).

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., … Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server), W202–W208. https://doi.org/10.1093/nar/gkp335

Kulakovskiy, I. V, Boeva, V. A., Favorov, A. V, & Makeev, V. J. (2010). Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, *26*(20), 2622–2623. https://doi.org/10.1093/bioinformatics/btq488

Lee, N. K., Choong, A. C. H., & Omar, N. (2016). ENSPART: An ensemble framework based on data partitioning for DNA Motif analysis. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 87–94). https://doi.org/10.1109/BIBE.2016.68

Li, L., Liang, Y., & Bass, R. L. (2007). GAPWM: a genetic algorithm method for optimizing a position weight matrix. *Bioinformatics*, *23*(10), 1188–1194. https://doi.org/10.1093/bioinformatics/btm080

Lihu, A., & Holban, Ş. (2015). A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Briefings in Bioinformatics*, bbv022--. https://doi.org/10.1093/bib/bbv022

Linhart, C., Halperin, Y., & Shamir, R. (2008). Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. *Genome Research*, *18*(7), 1180–1189. https://doi.org/10.1101/gr.076117.108

Liu, X. S., Brutlag, D. L., & Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Pacific Symposium on Biocomputing* (Vol. 6, pp. 127–138).

Liu, X. S., Brutlag, D. L., & Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, *20*(8), 835–839.

Machanick, P., & Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, *27*(12), 1696–1697.

Pavesi, G., Mauri, G., & Pesole, G. (2001). An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, *17*(Suppl 1), S207-214.

Roth, F. P., Hughes, J. D., Estep, P. W., & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, *16*(10), 939–945.

Shi, J., Yang, W., Chen, M., Du, Y., Zhang, J., & Wang, K. (2011). AMD, an automated motif discovery tool using stepwise refinement of gapped consensuses. *PLoS ONE*, *6*(9), e24576. https://doi.org/10.1371/journal.pone.0024576

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 Research*, *40*(4), e31–e31. https://doi.org/10.1093/nar/gkr1104

van Heeringen, S. J., & Veenstra, G. J. C. (2011). GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics*, *27*(2), 270–271. https://doi.org/10.1093/bioinformatics/btq636

Xu, M., & Su, Z. (2010). A novel alignment-free method for comparing transcription factor binding site motifs. *PLoS ONE*, *5*(1), e8797. https://doi.org/10.1371/journal.pone.0008797

Zeng, H., Edwards, M. D., Liu, G., & Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics*, *32*(12), i121--i127. https://doi.org/10.1093/bioinformatics/btw255