

Documentation

ChIPSummitDB is a collection of processed ChIP-seq data. It provides information about the possible direct or indirect interactions between transcription regulatory proteins and their positions on the genome.

About ChIPSummitDB:

The main goal of analysing ChIP-seq experiments is to identify regions in the genome where we find more sequencing reads (tags) than we would expect to see by chance. These regions are called peak regions due to the appearance of the visualized distribution of mapped tags (Albert, Wachi, Jiang, & Pugh, 2008).

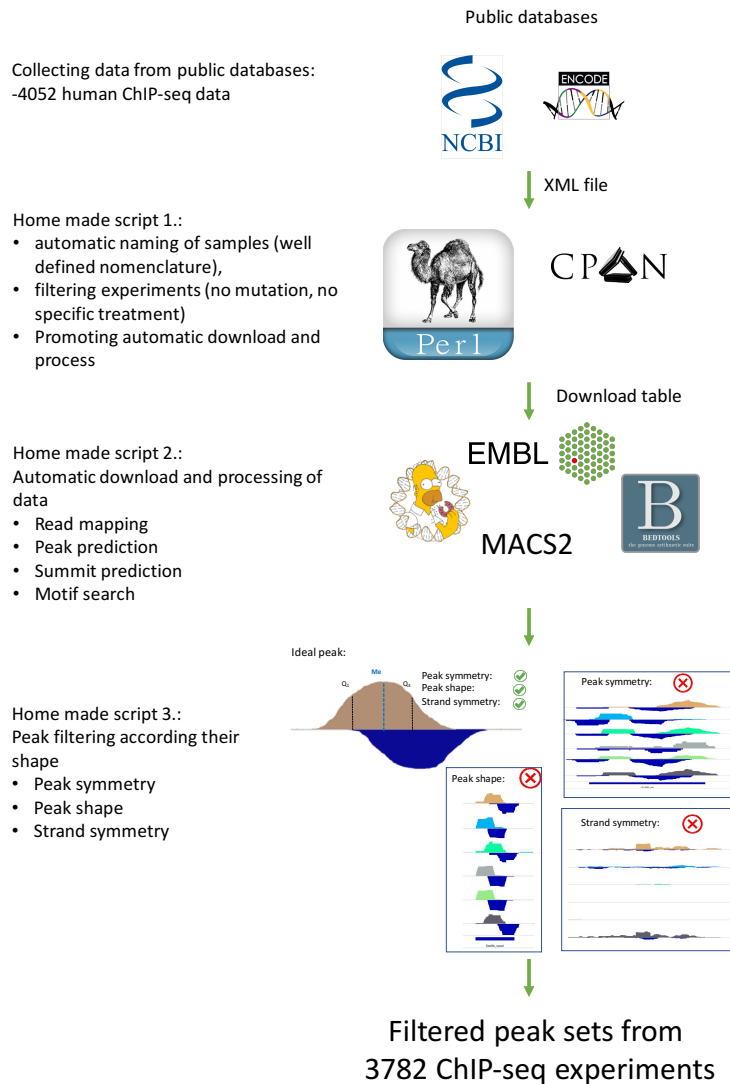
The peak's summit (maxima) shows the highest coverage of the region and is known to more-or-less coincide with the center of corresponding DNA elements in the case of transcription factors (Zhang et al., 2008).

These summits correlate with the accurate contact positions of the proteins on the DNA and can be used to determine the topological arrangements of the binding proteins relative to the strand-specific transcription factor binding sites (transcription factor binding motifs (TFBMs)). Earlier we showed (Nagy et al., 2016) that the exact positions of DNA binding proteins on the DNA can be extracted from the ChIP-seq data by identifying the peak summit positions.

Our goal was to create a global database which was based on combining the location of identified Transcriptional regulatory elements (TREs) with the positional information of the co-bound regulatory proteins (using publicly available ChIP-seq data, targeting as many proteins as we could). By investigating a global picture of different transcription factors and cofactors we can identify previously unknown transcriptional regulatory networks. Using the database, we can browse co-bound proteins on TREs and acquire information about their positioning relative to each other and the bound transcription factor motif.

Processing the data

Data from 4068 ChIP-seq experiments, covering a wide range of transcription factors (vagy DNA-binding proteins) and cell types, were collected from the NCBI SRA and ENCODE databases (Leinonen, Sugawara, Shumway, & Collaboration, 2011 "An integrated encyclopedia of DNA elements in the human genome," 2012) Processing of the downloaded raw data was carried out using an in-house developed ChIP-seq analysis pipeline, which involves the following steps, among others, mapping (Li & Durbin, 2010), peak calling (Zhang et al., 2008), tagdirectory creation and data visualization (Heinz et al., 2010). Following this analysis, the semi-processed data were further processed by various steps.



Peak splitting and filtering

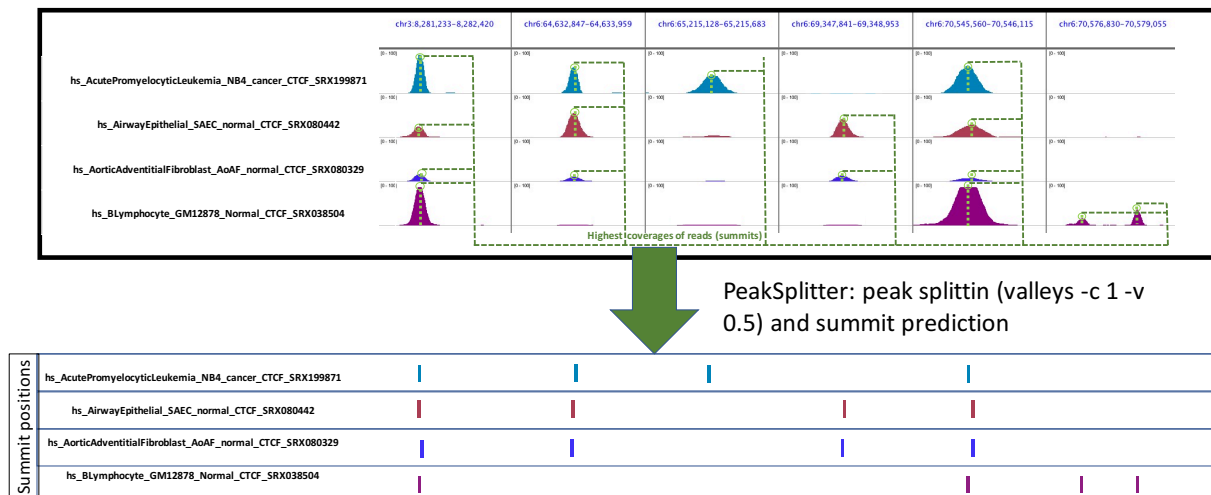
Identifying peaks with a well-defined maxima was crucial at the early stage of data processing because false positive peaks could result in false prediction of the protein's position. The peak summits (maxima) show the highest coverage for the peak region and coincide reasonably with the center of the corresponding DNA elements bound by transcription factors. Therefore, the identification of regions suitable for the clear determination of the summit position(s) were required. Current software packages use different strategies, such as the evaluation of peak prediction reproducibility or using false discovery rates (FDR), for peak prediction ((Taslim, Huang, Huang, & Lin, 2012), which dramatically decrease the false positive rates.

Unfortunately, by using these methods, the filtering algorithms needed to be configured differently for each experiment, which makes the automatization of the processing of large datasets more difficult. For better filtering, we have developed a pipeline, which reduces the false positive discovery rate even further.

In this pipeline, first we used PeakSplitter, which was developed to split sub-peaks when overlapping peaks are present, for summit predictions, and thus more accurate

local maxima could be obtained. The peaks for transcription factors bound sequences are usually concentrated to a narrow area, considerably showing a Gaussian distribution due to the random fragmentation and their narrow binding surface. This was especially observable after extension of reads to the expected fragment length (Zhang et al., 2008) Remarkably high signal and weak enrichment can refer to insufficient discarding of read duplicates or to library preparation artefacts (Star et al., 2014; Steven R. Head et al. 2014).

Summit prediction and peak splitting



To avoid false positive results, we filtered out duplicated reads by using a step in ChIP-seq analysis pipeline, and developed a Perl script which classified and filtered the sub-peaks based on their size and shape. In the script two parameters are responsible for the detection of the previously mentioned large signal intensity increase.

In these analyses, the peaks are considered coverage histograms and the positions of the median, first and third quartile values were used. The „ideal” transcription factor peak has three attributes; i) the read distribution of both strands have symmetrically curved shoulder, if the median value is the symmetry axis); ii) the peak’ shape displays a belllike curve; iii) the maxima of the ChIP-seq signal is approximately equal between the Watson and the Crick strands (Figure 2A). The first two steps are required for filtering out peak positions which have large gaps in their ChIP-seq signal intensity even after the read extension by peak caller software. The first formula calculates how symmetrical the two sides of the peak are (Figure 2B). For this, the maxima was used as the axis of symmetry. The second formula quantifies the shape of the peak based on the distances between the minimum, maximum, 2nd and 3rd quartiles values of ChIP-seq signal intensities within the peaks. This resulted value can change on a scale of 0 to 1. If we connect the four above-mentioned values with a straight line (where the x axis represents the position of the signal and y axis represents the signal intensity), the peaks which have a “0” shape value would be shaped like a triangle. In contrast, if the value converges to 0.5, the shape of the peak would resemble a square (Figure 2C).

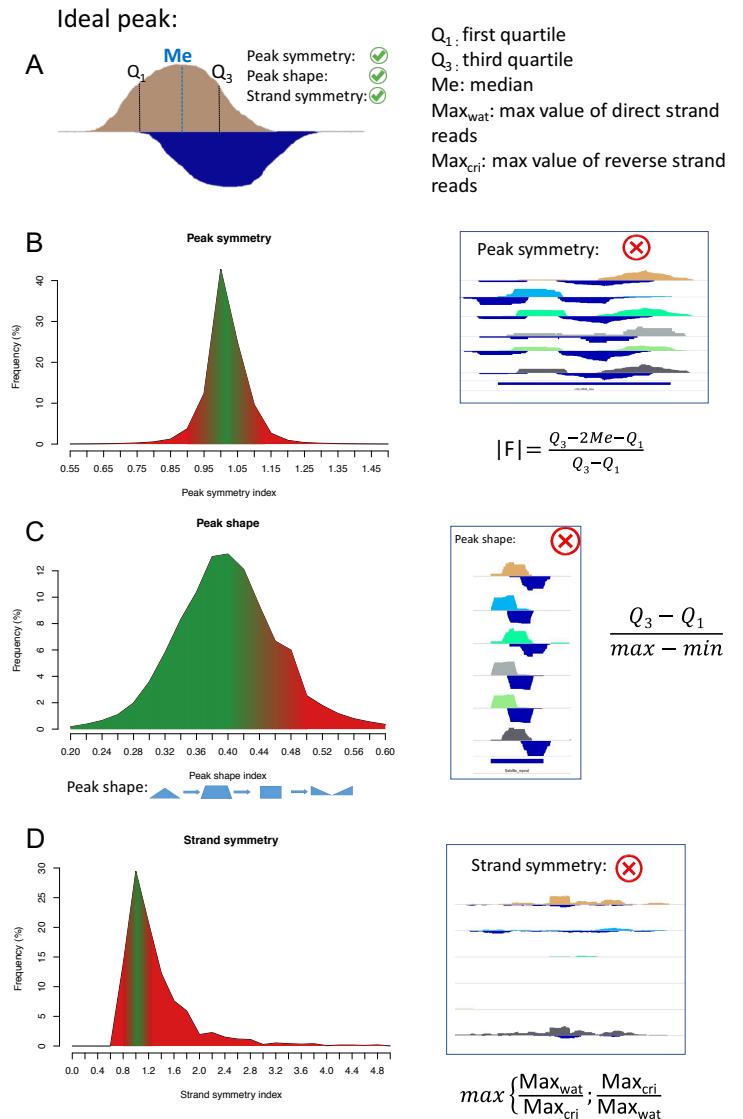
Optimally, the forward and the reverse tag counts (in a peak) have, approximately, the same size due to the ChIP-seq method. The third formula calculates the symmetry between the reverse and the direct strand tag counts (Figure 2D).

Due to the ChIP-Seq technology, at each protein-DNA binding site the tag counts from the forward strands are located on the left-hand side of the binding site, and the tag counts observed from the reverse strand are located on the right-hand side. This is an aspect which is considered and used by several peak-calling (e.g.: macs2) softwares to extend reads by an average value during peak identification. We used this parameter to filter data. We calculated forward-reverse maxima distances and values which could be found in the 90 percentile passed this filtering step.

Motif optimization and determining of their locations

Identification of the exact positions of TF binding sites is the basis of our ChIPSummitDB. These motif positions were not only a collection of regulatory regions but also the motif centres were also used as reference points for summit position analysis. Our primary goal was to create consensus binding site sets for as many transcription factors as possible. To do this, we used the JASPAR CORE database, which is a “curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes” (Khan et al., 2018) and incorporates 579 non-redundant motifs. We attempted to collect all motifs with ChIP-seq experiments from our collection. Several motifs are lacking NGS data for historical reasons, thus the JASPAR CORE was built to create families of binding profiles for as many structural transcription factor classes as possible. Despite of this, we could allocate only 338 motifs to at least one ChIP-seq experiment because in human, a number of cases no sequence and NGS data were available.

To optimise these allocated motifs, the peak regions of the corresponding ChIP-seq experiments were scanned for similar motif enrichments (homer link). The optimized motifs were manually curated and the most identical ones were paired with the given antibodies. This step maximized the number of specific motif instances, which were identified in the next step.



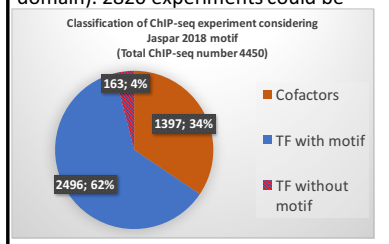
List of all collected ChIP-seq experiments peak sets 4068

hs_AcutePromyelocyticLeukemia_NB4_cancer_CTCF_SRX199871
hs_AirwayEpithelial_SAE normal_CTCF_SRX080442
hs_AorticAdventitialFibroblast_AoAF_normal_CTCF_SRX080329
hs_BLymphocyte_GM12878_Normal_CTCF_SRX038504
hs_BreastAdenoCarcinoma_MCF7_cancer_P300_SRX176885
hs_BreastCancer_T47D_cancer_CTCF_SRX100393
hs_BreastCancer_T47D_cancer_P300_SRX1012606
hs_CD59_U937_undef_PU1_ERX626807-homerpeaks.bed
hs_lymphoblastoid_GM12878_normal_PU1_SRX100576
hs_MonocyteDerived_macrophage_normal_PU1_SRX093189
hs_primaryadulTCD34HSP_primaryadulTCD34HSP_undef_PU1_SRX1089833
hs_primaryfetaliverCD3_primaryfetaliverCD3_undef_PU1_SRX1089832
hs_PulmonaryArteryFibroblasts_HPAF_normal_CTCF_SRX080344
hs_SkinFibroblast_BJ_normal_CTCF_SRX080340

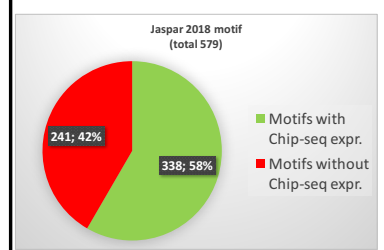
JASPAR CORE motifs (579 non-redundant motifs)

MA0018.3	CREB1	Homo sapiens	Basic leucine zipper factors (bZIP)	CREB-related factors	
MA0139.1	CTCF	Homo sapiens	C2H2 zinc finger factors	More than 3 adjacent zinc finger factors	
MA0467.1	Crx	Mus musculus	Homeo domain factors	Paired-related HD factors	
MA0608.1	Creb3l2	Mus musculus	Basic leucine zipper factors (bZIP)	CREB-related factors	
MA0080.2	SPI1	Homo sapiens	Tryptophan cluster factors	Ets-related factors	

2921 (from 3782 successfully processed) ChIP-seq targets were classified as transcription factor (with DNA binding domain). 2820 experiments could be



338 motif could be paired to any ChIP-seq experiments

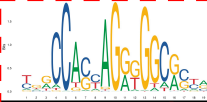



ChIP-seq – JASPAR motif table

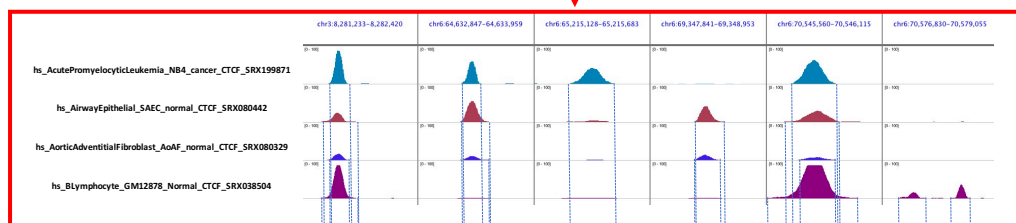
MOTIF	ChIP target name	Name of belonging experiments	Number of experiments	Position Weight Matrix
CTCF	CTCF	hs_AcutePromyelocyticLeukemia_NB4_cancer_CTCF_SRX199871	328	
		hs_AirwayEpithelial_SAE normal_CTCF_SRX080442		
		hs_AorticAdventitialFibroblast_AoAF_normal_CTCF_SRX080329		
		hs_BLymphocyte_GM12878_Normal_CTCF_SRX038504		
		...		
SPI1	PU1, Pu1, Pu.1, SPI1	hs_CD59_U937_undef_PU1_ERX626807-homerpeaks.bed	16	
		hs_lymphoblastoid_GM12878_normal_PU1_SRX100576		
		hs_MonocyteDerived_macrophage_normal_PU1_SRX093189		
		hs_primaryadulTCD34HSP_primaryadulTCD34HSP_undef_PU1_SRX1089833		
		hs_primaryfetaliverCD3_primaryfetaliverCD3_undef_PU1_SRX1089832		

Numerous tools can be used to find the occurrences of individual motifs. Instead of choosing one single tool, we combined 3 popular methods: HOMER, FIMO and MAST (Finak et al., 2015; Grant, Bailey, & Noble, 2011; Lee et al., 2013).

The positions which were identified to a given motif by at least two programs were selected in the first step of filtering. Using the default motif scores obtained by the above-mentioned three programmes and the distance of the closest summit, from the list of paired motif-ChIP-seq experiments, a weighted motif value was calculated. All identified ChIP-seq peaks were attempted to pair with the closest motif possessing the highest weighted motif value. The distance cutoff was +/- 50 base pair. Following this step, sets of non-redundant motifs were created by filtering out the motifs with identical position and direction. Even in the case of palindromic sequences, identifying motif directions was possible due to the flanking regions and the positional preferences of the peak summits.

MOTIF	ChIP target name	Name of belonging experiments	Number of experiments	Position Weight Matrix
CTCF	CTCF	hs_AcutePromyelocyticLeukemia_NB4_cancer_CTCF_SRX199871	328	
		hs_AirwayEpithelial_SAEC_normal_CTCF_SRX080442		
		hs_AorticAdventitialFibroblast_AoAF_normal_CTCF_SRX080329		
		hs_BLymphocyte_GM12878_Normal_CTCF_SRX038504		
SPI1	PU1, Pu1, Pu.1, SPI1	hs_CD59_U937_undef_PU1_ERX626807-homerpeaks.bed	16	
		hs_lymphoblastoid_GM12878_normal_PU1_SRX100576		
		hs_MonocyteDerived_macrophage_normal_PU1_SRX093189		
		hs_primaryadultCD34HSP_primaryadultCD34HSP_undef_PU1_SRX1089833		
...	...	hs_primaryfetalliverCD3_primaryfetalliverCD3_undef_PU1_SRX1089832
		...		

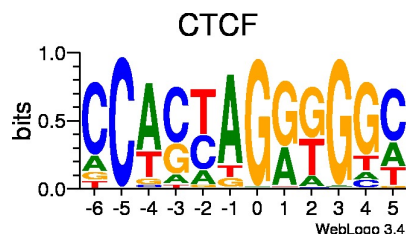
Take all CTCF ChIP-seq data
(328)



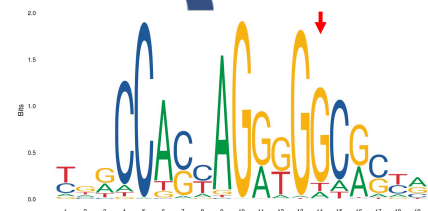
MergeBed

Merged CTCF region set

Motif
optimization



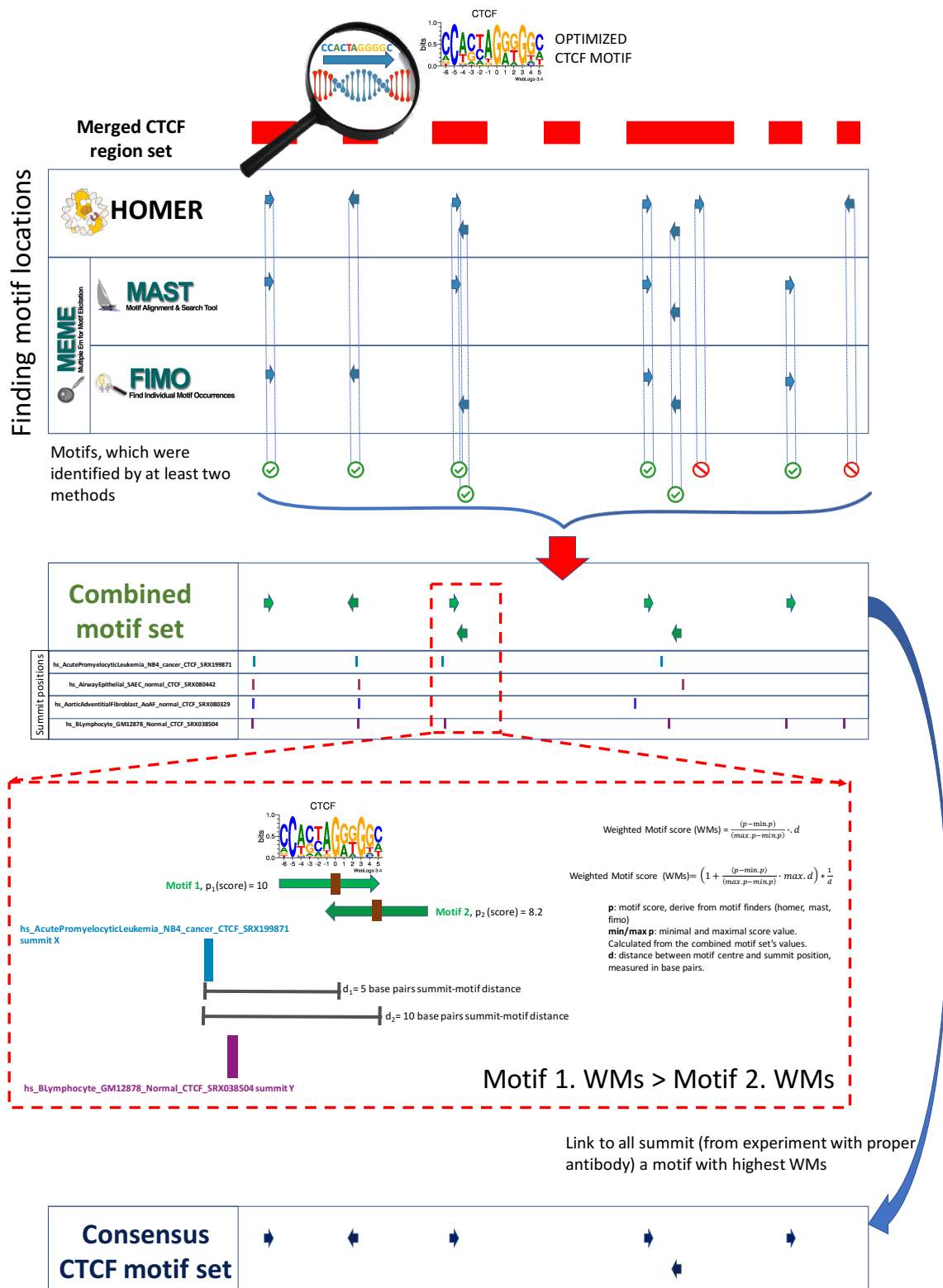
HOMER motif optimization
findMotifsGenome.pl -opt



In the previously mentioned step and in the following analysis, the closestBed, a tool of bedtools, was used to measure the distance between the centre of the motifs and the summits (Quinlan & Hall, 2010). If the length of an n bp long motif was even, then the $n/2+1$ bp from the 5' end of the sequence was considered as the centre of the motif. We created individual summit position pools for all motifs from their respective ChIP-seq experiments. Afterwards, the identified motifs and summits were combined using the closestBed program. This step resulted in a table where all of the summits positions from the proper set are shown with the nearest (one or more) motif instances. Distances between the centers of the motifs and the summits can be calculated this way. This distance and the score of the motif were both taken into account during the pairing of the most probable motifs to each of the summits. We combined these scores into a formula and the motif with the resulting highest score was picked for each summit position (one summit could have more than one motif in its vicinity, but only the strongest motif was selected for the following steps). The

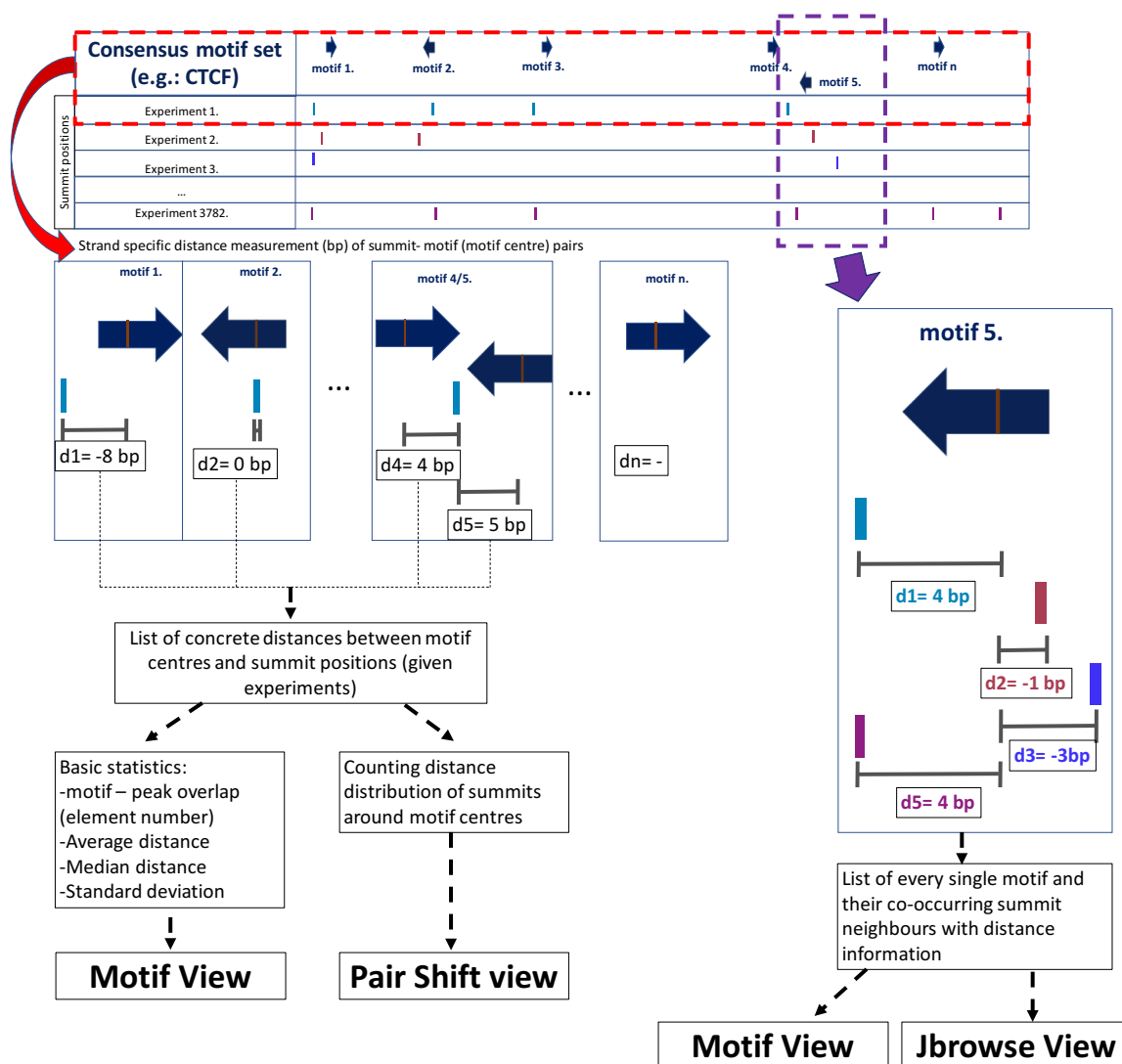
formula of our calculation can be found in the Figure 5 (WMs). The same motif was frequently paired to submits from different experiments. To avoid redundancy we removed the duplicates.

Thus, we get non-redundant global consensus motif sets to 338 JASPAR CORE matrices.



Summit distance calculation

The identified consensus sequence locations serve not just as a global map to transcription binding sites but they can be also used as reference points in the landscaping of possible co-binding and in the measuring of motif-protein or protein-protein distances. All motif occurrences obtained from every set were screened to identify ChIP-seq experiments containing peak summits in the ± 50 bp vicinity to the motif centre). The concrete distances can be calculated between motif centres and summit positions. The resulting distance tables can be examined in a global or a local manner. The global analysis can highlight large-scale protein positioning information, such as co-location frequency, location preferences between proteins, possible members of complexes or patterns in the protein composition of different regulatory regions. In addition to the frequency and the median/average values, both calculated from the measured distances, the standard deviation can be also informative. It was observed that the preferred position of a given factor has a larger standard deviation (in relation to the positions of the motif centers) if it is physically far from the reference point.



ChIPSummitDB allows to take a closer look at a specific region of the genome. Investigation of the summit positions on a specific motif can provide detailed information about the composition of regulatory complexes, their topology and it's possible to make comparisons among different cell lines.

References

- Albert, I., Wachi, S., Jiang, C., & Pugh, B. F. (2008). GeneTrack--a genomic data processing and visualization framework. *Bioinformatics (Oxford, England)*, 24(10), 1305–1306. <https://doi.org/10.1093/bioinformatics/btn119>
- An integrated encyclopedia of DNA elements in the human genome. (2012). *Nature*, 489. <https://doi.org/10.1038/nature11247>
- Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., ... Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16, 278. <https://doi.org/10.1186/s13059-015-0844-5>
- Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7), 1017–1018. <https://doi.org/10.1093/bioinformatics/btr064>
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., ... Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38(4), 576–589. <https://doi.org/10.1016/j.molcel.2010.05.004>
- Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., ... Mathelier, A. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1), D260–D266. Retrieved from <http://dx.doi.org/10.1093/nar/gkx1126>
- Lee, M. T., Bonneau, A. R., Takacs, C. M., Bazzini, A. A., DiVito, K. R., Fleming, E. S., & Giraldez, A. J. (2013). Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature*, 503(7476), 360–364. <https://doi.org/10.1038/nature12632>
- Leinonen, R., Sugawara, H., Shumway, M., & Collaboration, on behalf of the I. N. S. D. (2011). The Sequence Read Archive. *Nucleic Acids Research*, 39(Database issue), D19–D21. <https://doi.org/10.1093/nar/gkq1019>
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows--Wheeler transform. *Bioinformatics*, 26. <https://doi.org/10.1093/bioinformatics/btp698>
- Nagy, G., Czipa, E., Steiner, L., Nagy, T., Pongor, S., Nagy, L., & Barta, E. (2016). Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA. *BMC Genomics*, 17(1), 637. <https://doi.org/10.1186/s12864-016-2940-7>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26. <https://doi.org/10.1093/bioinformatics/btq033>
- Star, B., Nederbragt, A. J., Hansen, M. H. S., Skage, M., Gilfillan, G. D., Bradbury, I. R., ... Jentoft, S. (2014). Palindromic sequence artifacts generated during next

- generation sequencing library preparation from historic and ancient DNA. *PLoS One*, 9(3), e89676. <https://doi.org/10.1371/journal.pone.0089676>
- Taslim, C., Huang, K., Huang, T., & Lin, S. (2012). Analyzing ChIP-seq data: preprocessing, normalization, differential identification, and binding pattern characterization. *Methods in Molecular Biology (Clifton, N.J.)*, 802, 275–291. https://doi.org/10.1007/978-1-61779-400-1_18
- Yamamoto, S., Wu, Z., Russnes, H. G., Takagi, S., Peluffo, G., Vaske, C., ... Polyak, K. (2014). JARID1B is a luminal lineage-driving oncogene in breast cancer. *Cancer Cell*, 25(6), 762–777. <https://doi.org/10.1016/j.ccr.2014.04.024>
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., ... Li, W. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol*, 9. <https://doi.org/10.1186/gb-2008-9-9-r137>