ChIPSummitDB: A ChIP-seq- based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them

Erik Czipa[1†], Mátyás Schiller[2†], Tibor Nagy[1], Levente Kontra[1,2], László Steiner[4], Júlia Koller[1,3], Orsolya Pálné-Szén[1] and Endre Barta[1,2*]

[†]These authors contributed equally

[1]Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen, Hungary

[2]Agricultural Genomics and Bioinformatics Group, Agricultural Biotechnology Institute, NARIC, Gödöllő, Hungary

[3]Semmelweis University, Institute of Genomic Medicines and Rare Disorders, Budapest, Hungary

[4]UD-GenoMed Medical Genomic Technologies Research & Development Services Ltd., Nagyerdei krt. 98., Debrecen, H-4032 Hungary

* To whom correspondence should be addressed. Email: barta.endre@unideb.hu.

ABSTRACT

ChIP-Seq reveals genomic regions where proteins, e.g. transcription factors (TFs) interact with DNA. A substantial fraction of these regions, however, do not contain the cognate binding site for the TF of interest. This phenomenon might be explained by protein-protein interactions and co-precipitation of interacting gene regulatory elements. We uniformly processed 3,727 human ChIP-Seq data sets and determined the cistrome of 292 TFs, as well as the distances between the TF binding motif centers and the ChIP-Seq peak summits.

ChIPSummitDB enables the analysis of ChIP-Seq data using multiple approaches. The 292 cistromes and corresponding ChIP-Seq peak sets can be browsed in GenomeView.

Overlapping SNPs can be inspected in dbSNPView. Most importantly, the MotifView and PairShiftView pages show the average distance between motif centers and overlapping ChIP-Seq peak summits and distance distributions thereof, respectively.

In addition to providing a comprehensive human TF binding site collection, the ChIPSummitDB database and web interface allows for the examination of the topological arrangement of TF complexes genome-wide. ChIPSummitDB is freely accessible at http://summit.med.unideb.hu/summitdb/. The database will be regularly updated and extended with the newly available human and mouse ChIP-Seq data sets.

Keywords

ChIP-seq, peak summit, TF topology, transcription factor binding site (TFBS)

INTRODUCTION

ChIP-seq (Chromatin ImmunoPrecipitation followed by high-throughput sequencing) is a powerful technique, which reveals the genome-wide positions of those DNA sequences that co-precipitate with a given protein, which was used to generate the antibody for the IP, (1,2). The interaction between the protein and the DNA can be direct or indirect. Direct interactions can be specific, i.e. when a protein (transcription factor) recognizes and binds to a DNA sequence motif; or it can be nonspecific, as in the case of histones or cohesins (3–5). Indirect interactions between DNA and proteins occur through transcriptional regulatory complexes and/or DNA looping. In such cases, the cognate binding site for the given TF is not present under the ChIP-seq peaks (Additional file 1) (6).

In a typical primary ChIP-seq analysis pipeline the sequence reads are mapped to a reference genome, areas with the highest coverage (peaks) are determined, and the enriched *de novo* or known motifs at the peaks are identified. These steps are followed by downstream analyses, which typically involve peak annotation, comparison of

53 different ChIP-seq experiments, and visualization, for example generating profiles,
54 heat-maps and Venn-diagrams) (7). The most critical step in such a pipeline is the peak
55 calling. Different peak calling algorithms provides different results and the number of
56 the determined peaks also depends on the number of the sequenced reads (8).
57 Today, raw data from more than 85000 human and mouse ChIP-seq experiments are
58 available (9), which gives the opportunity to perform further analyses and/or to set up
59 secondary databases using those data. Previously, such databases have been built based
60 on different parameters of ChIP-seq analyses. Some databases (CODEX, BloodChIP,
61 hmChIP) put more focus on the experimental metadata collection and the classification
62 of the experiments by the cell type (10–12). In addition, CODEX provides a
63 visualization tool for examining peaks (10). Other databases, for example Cistrome
64 Data Browser, GTDR, ChIP-Atlas, Factorbook, carry out different downstream
65 analyses to show further details (13–16). Most of these databases are not only a simple
66 collection of ChIP-seq data and a display of ChIP-seq peaks. Factorbook, for example,
67 has an interactive tool to examine the nucleosome and histone modification profiles
68 around the ENCODE TF ChIP-seq peaks (16,17). The gene transcription regulation
69 database (GTRD) project, among other things, focuses on improving the peak calling
70 procedure (14). They use several peak calling algorithms and make clusters of
71 overlapping results. ChIP-Atlas provides a tool for extensive co-localization and
72 enrichment analyses (15). TFBSbank focuses on annotating genomic localizations,
73 finding co-binding proteins, and searching for *de novo* and known motifs within the
74 peaks (18). The Cistrome Data Browser combines ChIP-seq data with chromatin
75 accessibility data and provides a convenient web interface to browse and download
76 these data (13). Most of the above-mentioned databases contain not only human but

77 mouse data too (Cistrome Data Browser, Factorbook, CODEX, hmChIP), and, in some
78 cases, ChIP-seq data for other species (GTRD, ChIP-Atlas, TFBSbank).
79 Enhancers are distant regulatory elements relative to transcription start sites (19). They
80 can be characterized by TF binding (GTRD, TFBSbank), certain histone marks (SEdb),
81 and enhancer transcription (HACER) (20,21). Because both TF binding and histone
82 modifications can be detected by ChIP-seq, secondary ChIP-seq database tools can
83 predict enhancer sequences as well if they determine the cognate binding sites under
84 the peaks of TF ChIP-seq experiments. The ChIP-seq peaks and the protein-binding
85 DNA-motifs under them can be further analyzed, and the resulted data can be integrated
86 into a higher level database, such as the TRRUST (22).
87 ChIP-seq databases can provide tools to search, analyze, visualize, and download
88 existing ChIP-seq data. Our previous results, however, demonstrated that ChIP-seq
89 peak summits can also help to understand the topological arrangements, the spatial
90 position(s) of different proteins bound to the DNA double helix. Therefore, our aim
91 was to extend our CCCTC-binding factor (CTCF) peak summit-based analysis to every
92 available ChIP-seq transcription factor, which was examined in human ChIP-seq
93 experiments (23). For the study, we have manually selected 3727 ChIP-seq
94 experiments, which representing a very large number of human TFs and co-factors (9).
95 Since determining the correct positions of peak summits is critical for the analysis, we
96 have developed a robust peak filtering pipeline, by which the positions of peak summits
97 and the mapped TFBS motifs could be compared. Therefore, based on the ChIP-seq
98 peak regions for each TF, we defined the corresponding consensus motif binding site
99 sets for each of them. In addition to the consensus motif binding site sets, the
100 ChIPSummitDB contains the distances between each pair of mapped consensus motifs
101 and ChIP-seq peak summit.

102 The web interface of the ChIPSummitDB can display data in different views. Using the
103 GenomeView module, users can browse the genome for ChIP-seq peaks and consensus
104 motif binding site sets. The MotifView option can display the average distances
105 between the centers of consensus motif binding sites and the ChIP-seq peak summit
106 positions for each ChIP-seq experiment with overlapping peaks. Three different ChIP-
107 seq experiments can be compared in the PairShiftView module. Using the dbSNPView
108 option, users can evaluate whether a dbSNP entry overlaps with one or more consensus
109 motif binding sites from our database (24).
110 In addition to browsing all processed data, the analysis of the scatterplots provided in
111 the MotifView led us to hypothesize that the extent of the standard deviation of the
112 motif center vs peak summit distances may be proportional to the closeness of the given
113 protein to the DNA double helix.

114 **MATERIAL AND METHODS**

115 To construct the ChIPSummitDB, primary raw read data of ChIP-seq experiments and
116 the accompanying metadata were obtained from the NCBI SRA database (9).
117 Processing of the downloaded data into their final appearance in ChIPSummitDB is
118 summarized in Fig. 1. To determine peak regions, peak summits, consensus motif
119 positions, and the distances between peak summits and motif centers, we carried out
120 eight different processing steps (Additional file 1: Figure S1-S7). The scripts used
121 during the process were deposited to the GitHub repository
122 (https://github.com/summitdb). Below we provide some description of each step while
123 more details and explanation of them are on the ChIPSummitDB web site
124 (http://summit.med.unideb.hu/summitdb/Documentation.php) and in Additional file 1.

125 **1. ChIP-seq data collection from public databases**

126 We searched for human ChIP-seq experiments in the NCBI SRA database according to
127 its status as per November 1, 2017. We used the NCBI's run selector feature to
128 download all available metadata associated with the selected experiments (9). A custom
129 PERL script was used to process the downloaded data in XML table format and to give
130 a unique (descriptive) name to every experiment. The names include the species, the
131 tissue, the cell line (if available), the pathology (eg. normal or cancer), the ChIP target
132 protein, and the experiment's SRA database ID. Our aim was to restrict the analysis to
133 transcription factors and to other non-histone proteins, thus experiments with other type
134 of proteins were filtered out by using a script. For simplicity and to avoid redundancy,
135 we processed only the normal (without any specific treatment, which influence the TF-
136 DNA interaction) ChIP-seq experiments. The final list contained 4052 experiments and
137 was converted into a table with the proper format (BED, BEDGRAPH) for further
138 processing (Additional file 1: Figure S1).

139 **2. Basic ChIP-seq analysis**

140 For the basic processing, we used a modified version of our ChIP-seq_anal BASH script
141 (Additional file 1: Figure S1) (7). Briefly, the script needs two input files. The first is
142 the above-mentioned table with the experiments, while the second contained the SRR
143 IDs for each SRX ID. After downloading the SRR files and converting them into fastq
144 format, we mapped the reads to the hg19 (GRCh37) human reference genome using the
145 Burrows-Wheeler Aligner (BWA) program (25,26). Peak calling, generation of the
146 bedgraph coverage files, and *de novo* identification of protein-binding DNA motifs
147 were performed by using the HOMER package. The output of the analysis contained
148 the BAM files with the read sequences, the HOMER tag directories, the peaks in BED
149 file format, the BEDGRAPH files, and both the *de novo* identified and previously
150 known motifs, in a single html report (27).

151 **3. Peak splitting and summit prediction**

152 The basic analysis provided the peak regions for each ChIP-seq experiment in two
153 forms, a BED file containing the borders of the HOMER predicted peak regions and a
154 BEDGRAPH file with the coverage values of the extended reads within these peak
155 regions (27). It is possible to have more than one binding site within a peak region,
156 which can result in more than one summit. We employed the PeakSplitter program to
157 determine such summit positions using the BEDGRAPH files as input (28). The result
158 of this step is a BED file for each experiment containing the positions of the identified
159 summit(s) for each peak (Additional file 1: Figure S2).

160 **4. Peak filtering**

161 The usefulness of the ChIPSummitDB largely relies on the correct determination of
162 peak regions. Different peak -finding algorithms can give surprisingly diverse peak sets
163 using the same ChIP-seq experiments. A number of parameters in the experiment, such
164 as the number of the cells in the biological sample, the conditions of the sonication, the
165 quality of the antibody, the library preparation, and especially the depth of the
166 sequencing affect the number of detected peaks. There are different approaches to
167 determine the biologically most relevant peak regions. One approach was to apply
168 different peak calling programs and find the consensus peak sets from the results.
169 Furthermore, we also developed a different approach, in which we used only the
170 HOMER peak calling program, but applied a rigorous filtering, which was based on the
171 shape of the peaks, to reduce the number of false positives/artifact peaks in the peak
172 sets obtained in the previous step for downstream summit based topology analysis
173 (Additional file 1: Figure S3). Using this approach, we filtered out about 35 percent of
174 all peaks (27) (Additional file 3).

175 **5. Assigning consensus motifs to TF ChIP-seq experiments**

176 We assume that if the antibody used during the immunoprecipitation is against a TF,
177 then the cognate binding site for that TF will be enriched in the peak set of the given
178 ChIP-seq experiment. During the basic analysis, we determined the enriched motifs for
179 each peak set. The problem was that even if the immunoprecipitated TF was the same,
180 the resulting *de novo* motif could be slightly different. Also, the *de novo* motif finding
181 algorithms usually give more than one enriched motif and it is unclear whether the best
182 one is the cognate binding site. To precisely determine the peak summit motif center
183 distances, however, we need to use the same binding site for every experiment with the
184 same TFBS. Therefore, we chose a reverse approach. Based on the antibody used
185 during the immunoprecipitation, we assigned a JASPAR core consensus motif to each
186 experiment by hand. This resulted in a table where there is a corresponding JASPAR
187 consensus motif for each TF experiment (Additional file 1: Figure S4, Additional file
188 2) (29).

189 **6. Motif optimization**

190 Many of the consensus motifs in the JASPAR database are based on *de novo* motif
191 finding in ChIP-seq peak regions (29). Since we now have a good collection of
192 representative ChIP-seq experiments for the targeted transcription factors, we decided
193 to further optimize the consensus motif matrices. For this, we first merged the peaks of
194 the experiments belonging to the same consensus motif, and then we used the HOMER
195 package to optimize the matrices on these merge peak regions (Additional file 1: Figure
196 S5) (27). The resulting optimized matrices were further inspected and adjusted
197 manually. The motif optimization resulted in more than one similar motif in a few cases.
198 The decision between them couldn't be automatized. In these cases, the most analogous
199 motif was chosen.

200 **7. Remapping motifs**

201 Once we had all the optimized consensus motif matrices, we needed to map them back
202 to the genome. For each optimized JASPAR consensus motif matrix, we had a list of
203 corresponding ChIP-seq experiments (Additional file 1: Figure S4) (29). We took the
204 peak regions for those experiments and merged them to create the merged peak regions
205 for each matrix. This is an important and, so far, a unique step among the similar ChIP-
206 seq databases. Using this method, we specifically determined the possible transcription
207 factor binding sites. For mapping, we used three different programs and we kept only
208 the positions where at least two of them gave a hit for the final consensus motif binding
209 site sets (Additional file 1: Figure S6) (27,30,31).

210 **8. Motif center and summit distance calculation**

211 After the peak filtering in step 4, we got the list of peak summits for each ChIP-seq
212 experiment. In step 7, we got the ChIP-seq verified positions of TFBSs for each
213 consensus motif. The majority of the peak regions do not contain the cognate binding
214 site, even for TF ChIP-seq experiments. We hypothesize that in these cases the given
215 TF can be part of a complex, which is bound to the DNA through another TF, which in
216 turn is bound to its cognate binding site, which is also present in our consensus binding
217 site set. Therefore, we investigated such cases as well. To do this, we calculated every
218 distance between consensus motif centers and the nearby ChIP-seq peak summit
219 positions. The calculation is motif based, which means that the output is a list for each
220 consensus motif set (i.e. a given transcription factor binding site), which contains the
221 distances of every peak summit inside the 50 bp range of the given consensus motif
222 binding site (Additional file 1: Figure S7) (32). Practically, if we see a consensus motif
223 (e.g. CTCF), we will have the ChIP-seq verified instances of that motif in the genome
224 (88906) and a list of distances for each of the 3727 ChIP-seq experiments. These lists

225 were processed further in the database. The number of distances in these lists were used
226 as a cutoff value in the MotifView. The standard deviation of the distances in these lists
227 was also calculated and indicated in the Y-axis of the MotifView.
228 This processing of raw ChIP-seq read and metadata resulted in the following tables and
229 files:

230 1. The HOMER *homerpeaks.txt files for each experiment. These files contain the
231 peak regions and other parameters used in peak finding.
232 2. A peak region and a summit bed format file for each experiment.
233 3. A table with all the metadata related to the ChIP-seq experiments.
234 4. A table with the average distances between each consensus binding site set and
235 experiment.
236 5. A table with all the distances between individual consensus binding sites and
237 each overlapping peak summit.
238 6. The HOMER *de novo* motif finding results for every experiment.
239 7. The optimized consensus motif matrices for every consensus binding site set.

240 Based on these files and tables, the physical and logical scheme of the database was
241 developed using Oracle SQL Developer Database Modeler (version 4.2). For data
242 storage, MariaDB (version 5.5.60) was chosen because it is free and open source.
243 The website is hosted in CentOS Linux with Apache HTTP server and pages were
244 developed using PHP. Graphs and other visual elements representing statistical results
245 were made using the D3.js (version 3) JavaScript library.

246 **RESULTS AND DISCUSSION**

247 The ChIPSummitDB addresses two main aims. First, the database is a comprehensive
248 collection of ChIP-seq verified binding sites for 292 different TFs. Second, the database

249 provides a new tool for analyzing the topological relationship between bound and co-
250 bound proteins on the TFBSs.
251 There are six different views for exploring the database. From the main page, the user
252 can either choose one of them under the search view tab or click any of the example
253 pictures in order to go directly to a view page.

254 **User interfaces**

255 **MotifView** is the summary page for each of the examined TFs. The scatterplot
256 presented in MotifView shows the average distances of the overlapping peak summits
257 relative to the center of the adjusted TFBS. In the Y-axis, one can choose between
258 showing the number of overlapping peak summits and TFBS pairs or the standard
259 deviation of those distances.
260 Each point on the plot represents one ChIP-seq experiment. Circles represent
261 experiments where the antibody was against a TF with a defined TFBS, while triangles
262 show co-factors and other proteins attached to the primarily bound TF. On the top of
263 the scatterplot box there are three buttons. On default, all the experiments are displayed
264 which have the element number above the threshold set at the left side bellow the box.
265 In many cases when the "Minimum overlap number between motifs and peaks of
266 experiment" threshold is set to too low, the box is overloaded with scatters. In this case
267 there are two options. The "Only direct binding" button selects only those ChIP-seq
268 experiments, which have been done with the antibody corresponding to the given
269 consensus motif. The second option is to hide all the scatters. It allows then to
270 individually change the experiments to be shown because on the right side of the plot,
271 there is a color-coded list of antibodies or cell types used in these experiments. The list
272 can be sorted alphabetically or by the number of occurrences. By clicking on these
273 squares (TFs) or triangles (co-factors and other proteins), one can change the displayed

274 experiments. For example, the user can choose and examine only the experiments
275 belonged to one or more antibodies or cell types. This allows users to flexibly change
276 the displayed experiments. On the scatterplot, one can select up to three experiments
277 for further analysis in the other views.
278 The **PairShiftView** allows users to examine the distribution of individual peaks vs.
279 summit consensus motif distances for three experiments. To smooth the graph, a 5 bp
280 rolling bin is used in the histogram. Each curve represents one ChIP-seq experiment.
281 During the creation of the database, both the summit locations for all peaks for a given
282 experiment and the consensus motif binding sites of the chosen TF have been
283 determined. In the latter case, only those binding sites that have been counted are
284 shown, where the binding site in the genome overlaps at least one peak of a ChIP-seq
285 experiment immunoprecipitated with the antibody against the corresponding TF. In the
286 histogram, the highest points of the curves show the most likely topological position
287 for a given protein on the DNA double helix relative to the consensus motif. The height
288 of the curve depends on the number of overlapping consensus motif summit pairs
289 (element number). In most cases, the curve has one peak, but sometimes it also has a
290 shoulder. This indicates a more complex topological arrangement or a consequence of
291 other, still unknown, reasons. It is remarkable for example that in many cases the
292 distance between the main summit and the shoulder (or between two summits from
293 different experiments with the same antibody) is approximately 11 or 22 bp (Fig. 2)
294 (http://summit.med.unideb.hu/summitdb/paired_shift_view.php?exp1=419&exp2=19
295 60&exp3=3681&motive=FOXA1&motifid=77&limit=25&low_limit=-
296 25&formminid=1&formmaxid=10000&mnelem=100&formmaxelem=120000),
297 which represents one or two turn in the double helix and can be interpreted as a protein
298 on the same side of the DNA (33).

299 Although the highest point of the curve in the PairShiftView and the position of the
300 corresponding scatter in the X-axis indicate the same topological positioning of the
301 given protein on the DNA double helix, these values can be different. In most cases,
302 this is because in the histogram the distribution curve either has a shoulder or is simply
303 asymmetric. Therefore, we assume that while the scatterplot in the MotifView gives a
304 good overview of the differences in the topological arrangements of the given
305 consensus motif experiments pairs, the PairShiftView is a more accurate tool to
306 examine the relative topological arrangements of the three chosen experiments.
307 Although in most cases one gets to the PairShiftView from the MotifView by selecting
308 three experiments, the user can change both the consensus motif and the selected
309 experiments from inside the PairShiftView. This allows a quick comparison of any
310 three experiments.
311 Besides seeing the topological arrangement of three chosen immunoprecipitated
312 proteins relative to the given motif center, the extent to which the peaks of the three
313 ChIP-seq experiments share the same binding sites is also interesting. The **VennView**
314 is designed to allow this. During the processing of the ChIP-seq experiments, we
315 determined not only the filtered peaks for each ChIP-seq experiment but also the
316 consensus motif binding sets for each transcription factor. Thus, for each binding site
317 in the genome, the user can see which ChIP-seq experiment has an overlapping peak
318 (technically this means a peak summit position within 50 bp in either direction). In this
319 way, having the three chosen experiments and the consensus motif, we can count how
320 many sites have overlapping peak summits for each of the seven possible combinations.
321 In the VennView, the user can see these values in a Venn diagram. This can be useful
322 for comparing three experiments with not only the same antibody, but from different
323 tissues / developmental stages / treatments and also in examining the extent of

324 overlapping binding of a given transcription factor and its co-factors or co-bound
325 proteins.
326 In the recent release, the ChIPSummitDB contains data from the analysis of 3727
327 human ChIP-seq experiments. In the **ExperimentView**, the user can see an overview
328 of the main attributes for each experiment. Most importantly, there are links to the
329 NCBI SRA database (26). The number of sequencing reads and the number of filtered
330 peaks called by the HOMER findPeaks program is also listed (27). During the analysis
331 pipeline, we have determined the most enriched *de novo* and known motifs for each
332 experiment. The link for the results of this search is also located on this page.
333 The **GenomeView** allows users to see the database content (consensus binding sites,
334 peaks, etc.) in a genomic context through a web browser interface. This view is
335 implemented in the JBrowse framework (34). The GenomeView can be used as a
336 standalone web page where the users can select tracks to load and display. Users can
337 select from the 292 consensus motif sets, the 3727 experiments or from miscellaneous
338 tracks like genomic features or know SNPs. Therefore, users can compare any
339 combinations of experiments versus consensus motif binding sites. Users can also get
340 to this JBrowse interface from the MotifView, the PairShiftView, and the VennView
341 after selecting up to three experiments. In this case, the consensus motif set of the
342 chosen motif and the overlapping peaks and summits of the three chosen experiments
343 will be displayed initially. The user can display any other previously mentioned tracks.
344 Our database provides a comprehensive catalogue of experimentally verified
345 transcription factor binding sites in the human genome. As the cost of whole genome
346 sequencing is drastically decreasing, the number of variations associated with a certain
347 phenotype is rapidly increasing. Most of these variations are in intronic or intergenic
348 regions. Therefore, there is a great interest in determining the overlap of transcription

349 factor binding sites. The **dbSNPView** allows users to check these cases. Users can enter
350 either a genomic region or a dbSNP ID (24). In the first case, the webpage will then
351 display the given region with the variations on it and also the overlapping consensus
352 motif binding sites from our database. The overlapping SNPs are highlighted in red.
353 Either clicking on them or entering the dbSNP accession number directly into the search
354 field leads to the enlarged dbSNPView. Here the reference genome sequence together
355 with the logo of the consensus motif and the overlapping SNPs can be seen. This view
356 is useful for examining how severely the altered base can affect the transcription factor
357 binding. There is also a button to check which experiments give an overlapping peak
358 with the given variation.

359 **The Novelty of the ChIPSummitDB**

360 The ChIPSummitDB is the first ChIP-seq related database that analyzes and shows the
361 peak summit and consensus motif binding site center distances. The web interface of
362 the database provides several tools for displaying this kind of topological data. In the
363 MotifView, every experiment (antibodies used for the immunoprecipitation) that is
364 above the threshold number of overlapping peaks for a chosen consensus motif is
365 shown. Either direct binding (in the cases where the immunoprecipitated protein
366 recognizes and binds to the given binding site) or indirect binding is indicated. We
367 hypothesize that the standard deviation of the distances of the peak summit and binding
368 site centers shows the directness of the binding. For example, based on the standard
369 deviations, there are at least three groups of interacting proteins at the YY1 consensus
370 motif binding site (Fig. 3a, modified from:
371 http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=10000&minid=1&m
372 nelem=1000&mxelem=120000&motive=YY1). In the first group, the YY1 ChIP-seq
373 experiments have a standard deviation value between 16 and 22 (35). In the next layer,

374 the CTCF and the cohesin proteins and some other TF and co-factor have standard
375 deviation values between 23 and 27 (6). Interestingly, there are many other ChIP-seq
376 experiments, which have a standard deviation above 27 but still have more than 1000
377 peaks (the threshold value set here), which are overlapping with the YY1 consensus
378 motif binding sites throughout the genome. If we further analyze the scatterplot, we
379 may notice that the YY1 experiments are grouped around 2 at the X-axis, while the
380 CTCF experiments are grouped around -2, as most of the other TFs and cofactors from
381 the third group. Based on these observations, we can hypothesize that the YY1 binds to
382 its cognate binding site. The CTCF-cohesin complex binds to the already bound YY1
383 protein in the same horizontal arrangement as can be seen at the CTCF binding sites
384 (Fig. 3b, modified from
385 http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=10000&minid=1&m
386 nelem=100&mxelem=120000&motive=CTCF). The proteins in the third group most
387 likely bind to the CTCF-cohesin complex. There are, of course, not only CTCF-cohesin
388 proteins in the second group (e.g. SRF or auts2) (36). Proteins in the second group are
389 most probably directly bound to the YY1 protein.
390 Besides the already analyzed CTCF-cohesin topology, the ChIPSummitDB provides
391 further evidence that the peak summit versus the binding site based analysis indicates
392 topological arrangements for TF-DNA complexes. One good example is the
393 GATA1::TAL1 composite element (http://jaspar.genereg.net/matrix/MA0140.2/) (37).
394 Here, the ChIPSummitDB clearly confirms the experimental results. The topological
395 arrangement of the two transcription factors is clearly visible at both the MotifView
396 (Fig. 4a, modified from:
397 http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=2000&minid=1&mn
398 elem=1000&mxelem=120000&motive=GATA1%3A%3ATAL1, with the "only direct

399 binding" button) and at the PairShiftView (Fig. 4a, modified from:
400 http://summit.med.unideb.hu/summitdb/paired_shift_view.php?exp1=220&exp2=218
401 &exp3=undefined&motive=GATA1::TAL1&motifid=89&limit=25&low_limit=-
402 25&formminid=1&formmaxid=10000&mnelem=100&formmaxelem=120000).
403 Theoretically, we expect that the peak summit will be in the middle of the transcription
404 factor binding site. We have already shown, in the case of the CTCF-cohesin complex,
405 that this is not necessarily true (38). The ChIPSummitDB provides even more extreme
406 cases. For example, in the case of the NFYB motif observed in the MotifView (Fig. 4b,
407 modified from:
408 http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=10000&minid=1&m
409 nelem=2000&mxelem=120000&motive=NFYB), the average distance values that the
410 NFYB binds are upstream, while the USF1 binds downstream of the motif. If we further
411 scrutinize the topology on the PairShiftView, which shows the real distribution of the
412 distances, we can see even more extremities (Fig. 4b, modified from:
413 http://summit.med.unideb.hu/summitdb/paired_shift_view.php?exp1=2301&exp2=76
414 1&exp3=1597&motive=NFYB&motifid=175&limit=40&low_limit=-
415 40&mnelem=2000). In the case of NFYB, most of the distances between the
416 overlapping consensus motifs and peak summits are around the +15 - +17 positions. In
417 the case of USF1, however, the majority of the distance values are clustered around the
418 -18 position. It is also remarkable that in the distribution curves, there are also other
419 smaller shoulders.

420 **Conclusions**

421 ChIPSummitDB is the first ChIP-seq database based on a transcription factor binding
422 site centered analysis of peak summits. The database convincingly confirms our
423 previous hypothesis that if the different proteins are sitting on the DNA not exactly

424 above each other, then the average peak summit position will display a shifted value
425 relative to the motif center. There can be numerous reasons for this phenomenon. One
426 obvious example is when two different transcription factor binds nearby on a composite
427 element (37). Surprisingly, there are cases, when a transcription factor is bound to its
428 cognate site in the DNA and somehow a different protein without a DNA binding
429 domain that is bound to that TF gets so close to the DNA double helix that it will
430 crosslink to it during the experiment. This will result in a shifted peak summit versus
431 motif center value in our analysis (as can be seen in MotifView and PairShiftView in
432 our database). We showed this shift for cohesion proteins (23), but this shift can also
433 be recognized in other cases.
434 The detailed analysis of the ChIP-seq summit and motif center positions led us to a new
435 hypothesis: Taking a consensus binding site set (ChIP-seq verified binding sites for a
436 given transcription factor), the closer a given protein is to the DNA, the lower the
437 standard deviation of the distances between overlapping peak summits versus motif
438 center pairs. In other words, if a protein is very close to the DNA double helix, which
439 means in most cases that the protein is bound to the DNA, the resulting ChIP-seq peak
440 summits will more likely be centered in the middle of the DNA region covered by the
441 protein. This recognition can help us better understand how the protein complexes are
442 built on the DNA starting from binding of a transcription factor to its cognate binding
443 site.
444 Besides these completely new features, the ChIPSummitDB provides a comprehensive,
445 experimental based collection of transcription factor binding sites. The site can be
446 browsed in the GenomeView and we have also developed a dbSNP view that allows
447 users to check whether a given SNP is overlapping a TFBS (24,34). This feature can be
448 useful in determining the consequences of non-coding mutations.

449 **List of abbreviations**

450 bp: base pair

451 BWA: Burrows-Wheeler Aligner

452 ChIP-seq: Chromatin Immunoprecipitation Sequencing

453 CTCF: CCCTC-binding factor

454 DB: Database

455 dbSNP: Single Nucleotide Polymorphism Database

456 ELF1: E74-like factor 1

457 ETS1: E26 Oncogene Homolog 1

458 GATA1: GATA-binding factor 1

459 GM12878: Lymphoblastoid cell line

460 HOMER: Hypergeometric Optimization of Motif EnRichment

461 HTML: HyperText Markup Language

462 HTTP: HyperText Transfer Protocol

463 ID: identification number/identifier

464 NCBI: National Center for Biotechnology Information

465 NFYB: Nuclear transcription factor Y subunit beta

466 PHP: Hypertext Preprocessor

467 PWM: Position weight matrix

468 RAD21: Double-strand-break repair protein (Scc1, Mcd1)

469 SA1: Stromal Antigen 1

470 SMC1/3: Structural maintenance of chromosomes proteins

471 SNP: Single Nucleotide Polymorphism

472 SQL: Structured Query Language

473 SRA: Sequence Read Archive

TAL1: T-cell acute lymphocytic leukemia protein 1

TF: transcription factor

TFBS: transcription factor binding site

USF1: Upstream stimulatory factor 1

XML: Extensible Markup Language

YY1: Yin Yang 1

ZNF143: Zinc Finger Protein 143

Declarations

Ethics approval and consent to participate

Since this study carries out exclusively in silico analysis of already published data, it does not need ethics committee approval, as it did not constitute biomedical research. However, reporting is consistent with all ethical requirements.

Consent for publication

No consent for publication is required.

Funding

This work was supported by the GINOP-2.3.2-15-2016-00044, the 2017-1.3.1-vke-2017-00026 and the FIKP_20428-3_2018_FELITSTRAT grants.

Authors' contributions

EB initiated the project. EB and EC conceived and designed the overall project. EC, JK and OPS carried out the data collection. EB carried out the basic data processing. EC designed and carried out the downstream analysis and the generation of data tables. The website was designed and created by MS and TN and LK. EC, TN and LS performed

496 the statistical analyses and data filtering. EB and EC evaluated the results and wrote the
497 manuscript.

498 **Availability of data and materials**

499 The datasets supporting the conclusions of this article are included within the article
500 and in Additional file 1-3. The full list of guidelines and a glossary of terms can be
501 found in ChIPSummitDB website (http://summit.med.unideb.hu/summitdb/) and in
502 Additional file 1. The scripts used during the analysis are available at the following
503 website: https://github.com/summitdb.

504 **Competing interests**

505 The authors declare no conflict of interest.

506 **Acknowledgments:**

507 The authors would like to thank Ferenc Marincs and Karen L. Uray for their critical
508 discussion and technical assistance. Data analyses were performed on the High
509 Performance Computing cluster at the Genomic Medicine and Bioinformatic Core
510 Facility of Department of Biochemistry and Molecular Biology, University of
511 Debrecen.

512 **References**

513 1. Ren, B., Wyrick, J. J., Aparicio, O., et al. (2000) Genome-Wide Location and
514 Function of DNA Binding Proteins, 2306–2310.
515 2. Vishwanath R. Iyer, et al (2001) Genomic binding sites of the yeast cell-cycle
516 transcription factors SBF and MBF.
517 3. Gordan, R., Hartemink, A. J. and Bulyk, M. L. (2009) Distinguishing direct

518 versus indirect transcription factor-DNA interactions., *Genome Res.*, **19**, 2090–
519 2100.
520 4. Xiao, T., Wallace, J. and Felsenfeld, G. (2011) Specific sites in the C terminus
521 of CTCF interact with the SA2 subunit of the cohesin complex and are required
522 for cohesin-dependent insulation activity., *Mol. Cell. Biol.*, **31**, 2174–2183.
523 5. Bartke, T., Vermeulen, M., Xhemalce, B., et al. (2010) Nucleosome-interacting
524 proteins regulated by DNA and histone methylation, *Cell*, **143**, 470–484.
525 6. Beagan, J. A., Duong, M. T., Titus, K. R., et al. (2017) YY1 and CTCF
526 orchestrate a 3D chromatin looping switch during early neural lineage
527 commitment, 1–14.
528 7. Barta, E. (2011) Command line analysis of ChIP-seq results, *EMBnet.journal;*
529 *Vol 17, No 1 Next Gener. Seq. Data Anal. - 10.14806/ej.17.1.209* .
530 8. Koohy, H., Down, T. A., Spivakov, M., et al. (2014) A Comparison of Peak
531 Callers Used for DNase-Seq Data.
532 9. Leinonen, R., Sugawara, H., Shumway, M., et al. (2011) The Sequence Read
533 Archive, *Nucleic Acids Res.*, **39**, D19–D21.
534 10. Sánchez-Castillo, M., Ruau, D., Wilkinson, A. C., et al. (2015) CODEX: a
535 next-generation sequencing experiment database for the haematopoietic and
536 embryonic stem cell communities, *Nucleic Acids Res.*, **43**, D1117–D1123.
537 11. Chacon, D., Beck, D., Perera, D., et al. (2014) BloodChIP: a database of
538 comparative genome-wide transcription factor binding profiles in human blood
539 cells., *Nucleic Acids Res.*, **42**, D172-7.
540 12. Chen, L., Wu, G. and Ji, H. (2011) hmChIP: a database and web server for
541 exploring publicly available human and mouse ChIP-seq and ChIP-chip data,
542 *Bioinformatics*, **27**, 1447–1448.

543 13. Mei, S., Qin, Q., Wu, Q., et al. (2017) Cistrome Data Browser: a data portal for
544 ChIP-Seq and chromatin accessibility data in human and mouse., *Nucleic Acids*
545 *Res.*, **45**, D658–D662.
546 14. Yevshin, I., Sharipov, R., Valeev, T., et al. (2017) GTRD: a database of
547 transcription factor binding sites identified by ChIP-seq experiments., *Nucleic*
548 *Acids Res.*, **45**, D61–D67.
549 15. Oki, S., Ohta, T., Shioi, G., et al. (2018) ChIP-Atlas : a data-mining suite
550 powered by full integration of public ChIP-seq data, 1–10.
551 16. Wang, J., Zhuang, J., Iyer, S., et al. (2013) Factorbook.org: a Wiki-based
552 database for transcription factor-binding data generated by the ENCODE
553 consortium, *Nucleic Acids Res.*, **41**, D171–D176.
554 17. Landt, S. G., Marinov, G. K., Kundaje, A., et al. (2012) ChIP-seq guidelines
555 and practices of the ENCODE and modENCODE consortia, *Genome Res.*, **22**,
556 1813–1831.
557 18. Chen, D., Jiang, S., Ma, X., et al. (2017) TFBSbank: a platform to dissect the
558 big data of protein-DNA interaction in human and model species, *Nucleic*
559 *Acids Res.*, **45**, D151–D157.
560 19. Blackwood, E. M. and Kadonaga, J. T. (1998) Going the distance: a current
561 view of enhancer action., *Science*, **281**, 60–63.
562 20. Jiang, Y., Qian, F., Bai, X., et al. (2019) SEdb: a comprehensive human super-
563 enhancer database., *Nucleic Acids Res.*, **47**, D235–D243.
564 21. Wang, J., Dai, X., Berry, L. D., et al. (2019) HACER: an atlas of human active
565 enhancers to interpret regulatory variants., *Nucleic Acids Res.*, **47**, D106–D112.
566 22. Han, H., Cho, J.-W., Lee, S., et al. (2018) TRRUST v2: an expanded reference
567 database of human and mouse transcriptional regulatory interactions., *Nucleic*

568 *Acids Res.*, **46**, D380–D386.

569 23. Nagy, G., Czipa, E., Steiner, L., et al. (2016) Motif oriented high-resolution
570 analysis of ChIP-seq data reveals the topological order of CTCF and cohesin
571 proteins on DNA, *BMC Genomics*, **17**, 637.

572 24. Sherry, S. T., Ward, M. H., Kholodov, M., et al. (2001) dbSNP: the NCBI
573 database of genetic variation, *Nucleic Acids Res.*, **29**, 308–311.

574 25. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with
575 Burrows-Wheeler transform., *Bioinformatics*, **25**, 1754–1760.

576 26. Bethesda (MD): National Center for Biotechnology Information (US) *Sequence*
577 *Read Archive Submissions Staff. Using the SRA Toolkit to convert .sra files into*
578 *other formats.*; 2011.

579 27. Heinz, S., Benner, C., Spann, N., et al. (2010) Simple combinations of lineage-
580 determining transcription factors prime cis-regulatory elements required for
581 macrophage and B cell identities., *Mol. Cell*, **38**, 576–589.

582 28. Salmon-Divon, M., Dvinge, H., Tammoja, K., et al. (2010) PeakAnalyzer:
583 Genome-wide annotation of chromatin binding and modification loci, *BMC*
584 *Bioinformatics*, **11**, 415.

585 29. Khan, A., Fornes, O., Stigliani, A., et al. (2018) JASPAR 2018: update of the
586 open-access database of transcription factor binding profiles and its web
587 framework, *Nucleic Acids Res.*, **46**, D260–D266.

588 30. Finak, G., McDavid, A., Yajima, M., et al. (2015) MAST: a flexible statistical
589 framework for assessing transcriptional changes and characterizing
590 heterogeneity in single-cell RNA sequencing data, *Genome Biol.*, **16**, 278.

591 31. Grant, C. E., Bailey, T. L. and Noble, W. S. (2011) FIMO: scanning for
592 occurrences of a given motif, *Bioinformatics*, **27**, 1017–1018.

593 32. Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for
594 comparing genomic features, *Bioinformatics*, **26**.
595 33. Watson, J. D. and Crick, F. H. C. (1953) Molecular Structure of Nucleic Acids:
596 A Structure for Deoxyribose Nucleic Acid, *Nature*, **171**, 737.
597 34. Skinner, M. E., Uzilov, A. V, Stein, L. D., et al. (2009) JBrowse: a next-
598 generation genome browser, *Genome Res.*, **19**, 1630–1638.
599 35. Yant, S. R., Zhu, W., Millinoff, D., et al. (1995) High affinity YY1 binding
600 motifs: identification of two core types (ACAT and CCAT) and distribution of
601 potential binding sites within the human beta globin cluster, *Nucleic Acids*
602 *Res.*, **23**, 4353–4362.
603 36. Basu, A., Wilkinson, F. H., Colavita, K., et al. (2014) YY1 DNA binding and
604 interaction with YAF2 is essential for Polycomb recruitment, *Nucleic Acids*
605 *Res.*, **42**, 2208–2223.
606 37. Han, G. C., Vinayachandran, V., Bataille, A. R., et al. (2015) Genome-Wide
607 Organization of GATA1 and TAL1 Determined at High Resolution, *Mol. Cell.*
608 *Biol.*, **36**, 157–172.
609 38. Nagy, G., Czipa, E., Steiner, L., et al. (2016) Motif oriented high-resolution
610 analysis of ChIP-seq data reveals the topological order of CTCF and cohesin
611 proteins on DNA, *BMC Genomics*, **17**.
612

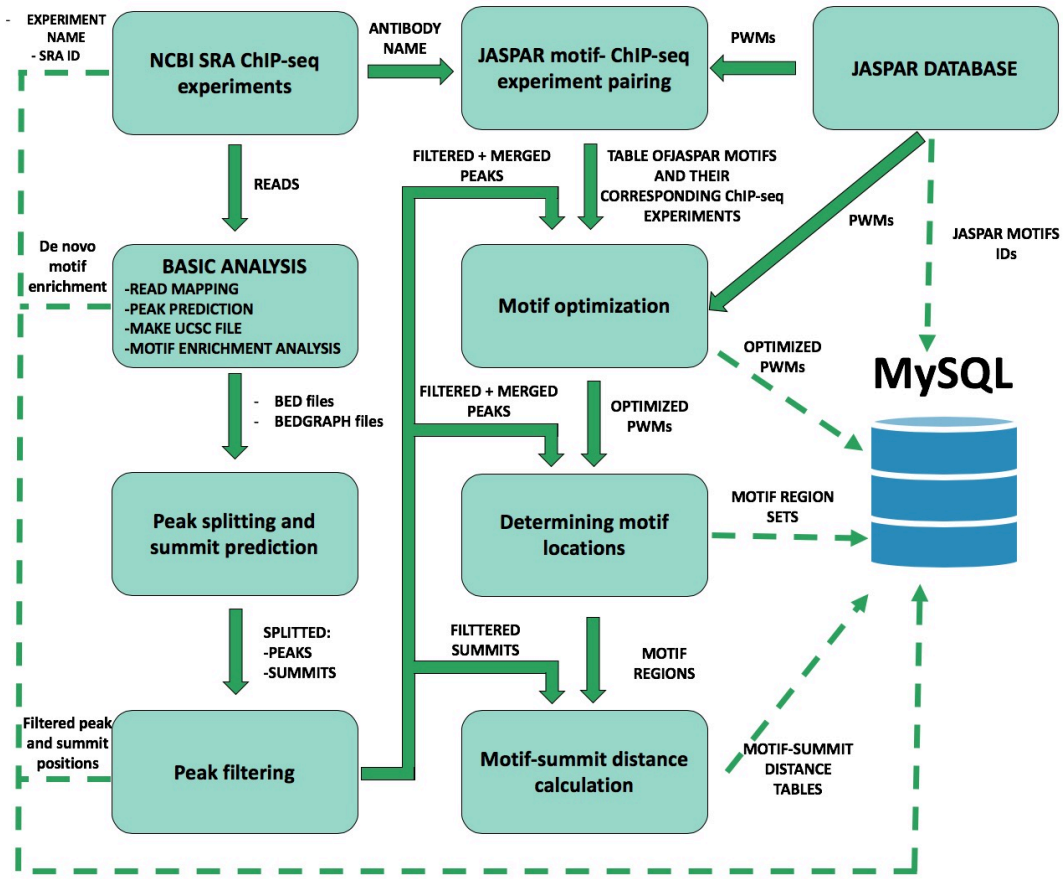


Fig. 1

Schematic overview of ChIP-seq data processing and imported content from MySQL. The analysis steps and data conversion are marked with thick arrows. The uploaded results/files are represented with dashed lines. A vast majority of processed data is available on ChIPSummitDB, including the predicted peak regions, optimized JASPAR CORE PWMs, identified TFBSs, and calculated protein position information.

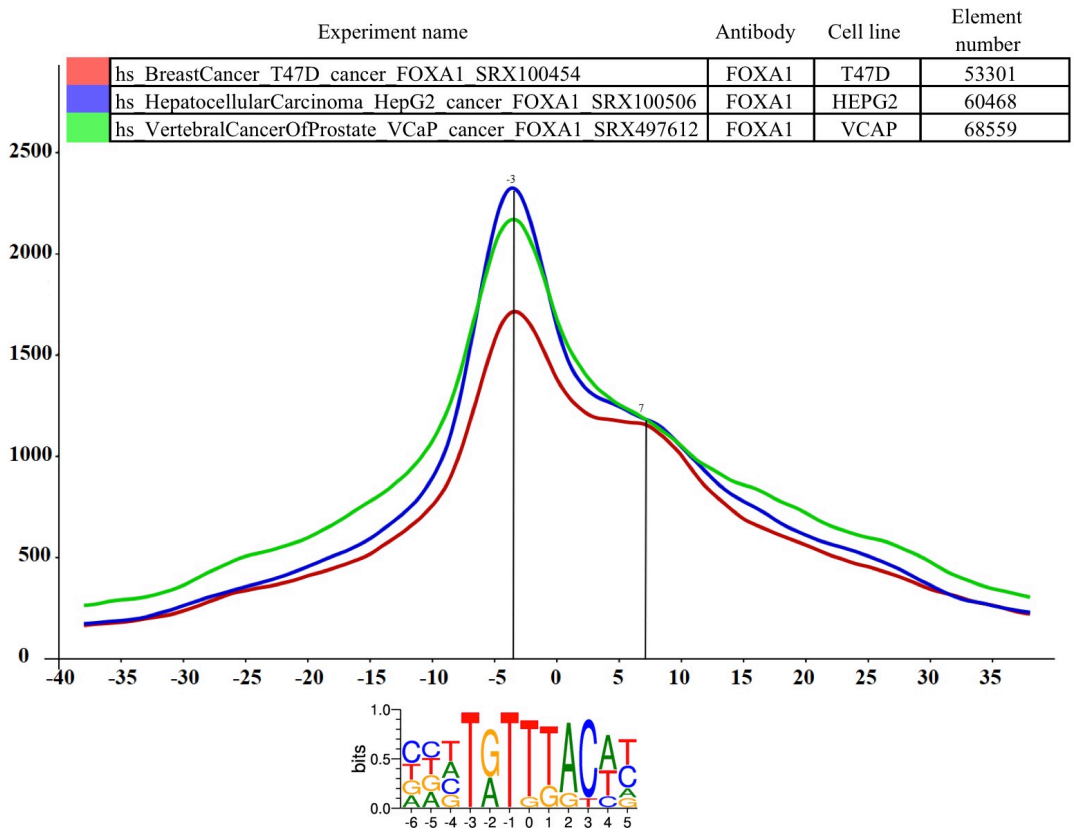
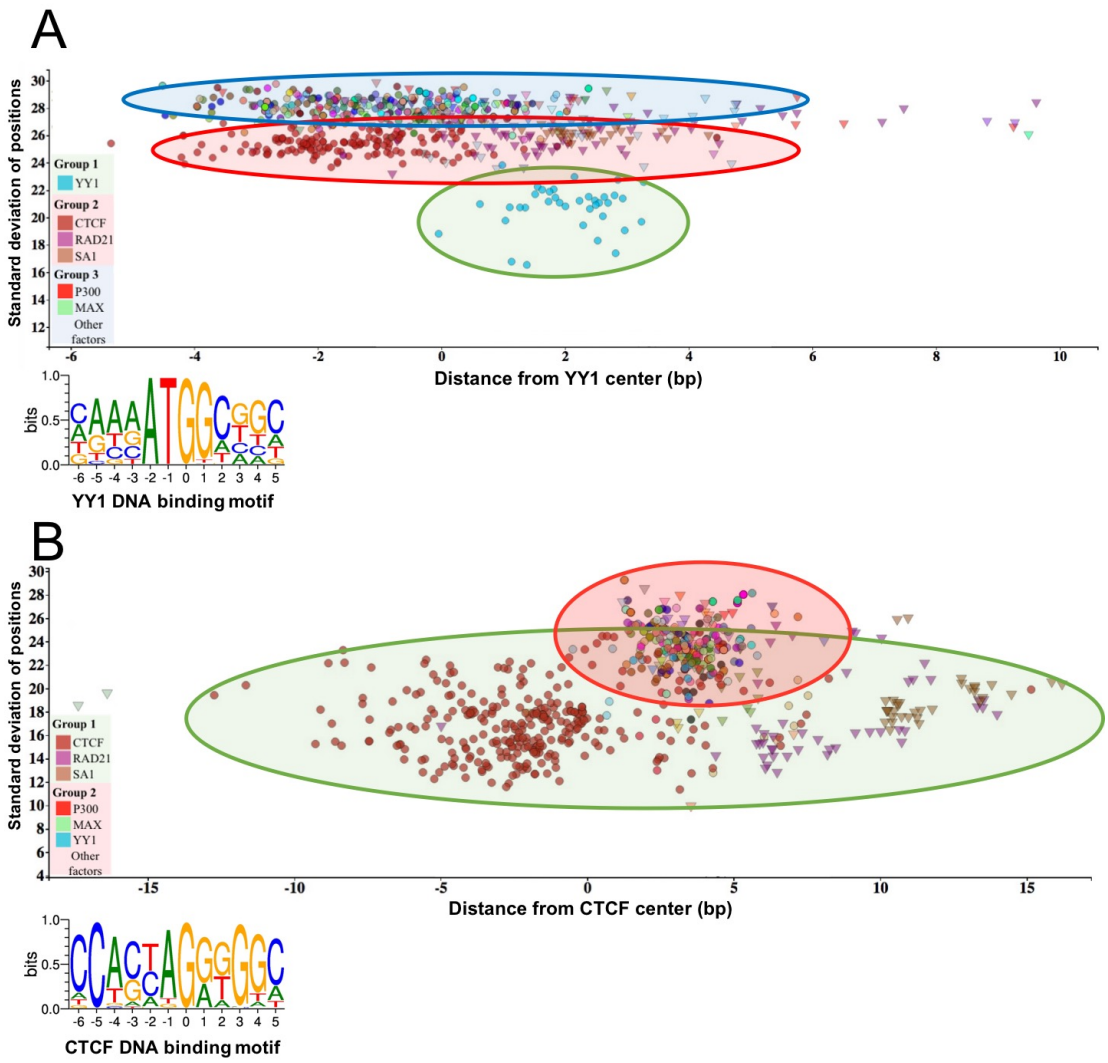


Fig. 2 The distance distribution of FOXA1 summits relative to the motif centers of FOXA1 binding sites. The horizontal axis represents the distance of summits in different cell lines (T47D (SRA ID: SRX100454, red curve), HepG2 (SRA ID: SRX100506 blue curve) and VCaP (SRA ID: SRX497612, green curve)) relative to the FOXA1 motif center. The vertical axis represents the distance frequencies. A rolling mean with a 5 bp window was applied to smooth the frequency curves. The distance between the maxima (main summit, maxima at -3 bp) and the shoulder (7 bp) is approximately 10 bp. Element numbers in the table indicates the number of peak regions obtained in a ChIP-seq experiment, which overlap with a particular consensus motif binding site set. Figure is adapted from ChIPSummitDB website: http://summit.med.unideb.hu/summitdb/paired_shift_view.php?exp1=419&exp2=1960&exp3=3681&motive=FOXA1&motifid=77&limit=25&low_limit=-

634 25&formminid=1&formmaxid=10000&mnelem=100&formmaxelem=120000

635



636

637 **Fig. 3**

638 The standard deviation of the distances of the peak summit and binding site centres
639 shows the DNA-protein proximity. Each scatter represents average summit position
640 from a single ChIP-seq experiment. The X axis represents the distance from the binding
641 site centre, which position is marked by „0" in the binding motif logo. The standard
642 deviation of the summit- motif centre distances are shown on the Y axis. A) The
643 proteins, which show interaction with YY1 binding sites, are arranged in three groups.
644 The lowest SD (between 16 and 22) belong to YY1 protein, which binds directly to the

645 YY1 DNA binding motifs. In the second layer, CTCF and cohesin subunit (RAD21,
646 SA1) ChIP-seq signals are the most common. The third group with high SD, above 27,
647 represents a diverse population, which consist of ChIP-seq experiments with different
648 protein targets and more than 1000 overlapping peaks. The P300 and MAX proteins
649 from group 3 are labelled by red and green colours, respectively. The figure was slightly
650 modified and adapted from ChIPSummitDB website:
651 http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=10000&minid=1&m
652 nelem=1000&mxelem=120000&motive=YY1 B) In the case of the CTCF binding
653 sites, only two layers can be distinguished. In the first group the directly interacting
654 CTCF, RAD21 and SA1 proteins can be found, while the YY1, P300, MAX and other
655 proteins are presented in the second group. Please note that the relative position of the
656 YY1 and CTCF proteins to each other is the same on both plots. The figure was slightly
657 modified and adapted from ChIPSummitDB website:
658 http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=10000&minid=1&m
659 nelem=5000&mxelem=120000&motive=CTCF
660

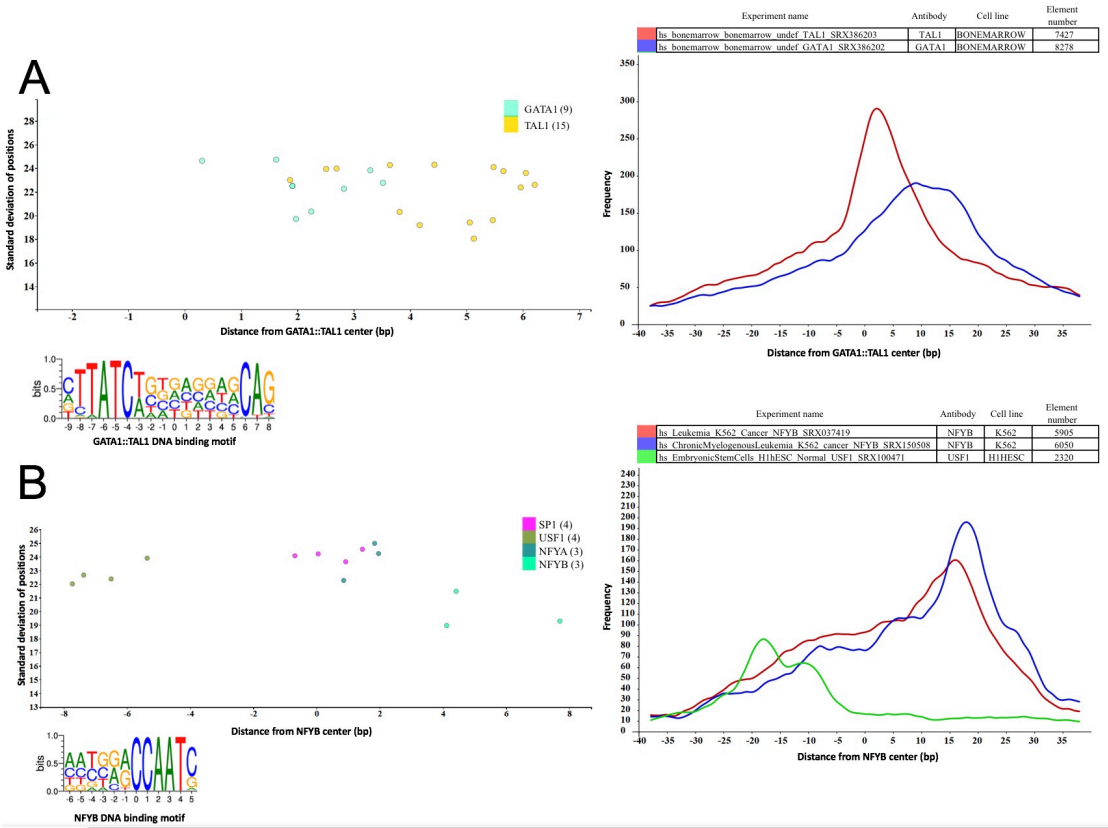


Fig. 4

Binding sites based analysis of topological arrangements of TF-DNA complexes as visualized in MotifView and PairShiftView. The plots show the preferred positions of different proteins on A) GATA1::TAL1 binding sites and on B) NFYB binding sites. The scatterplot follows the same logic as shown on Fig. 3. The figures derive from ChIPSummitDB, although the scatters were filtered to show only the presented factors. GATA1::TAL1: http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=2000&minid=1&mnelem=1000&mxelem=120000&motive=GATA1%3A%3ATAL1; NFYB: http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=2000&minid=1&mnelem=2000&mxelem=120000&motive=NFYB. The histograms (at right) show the distribution of the summits relative to the midpoint (motif centers). The horizontal axis shows the distance from motif center, measured in base pairs. The vertical axis displays the distance frequency of summits at the given positions. Each ChIP-seq experiment is

676 represented by a frequency curve A) GATA1:TAL1: blue - SRX386203, red -
677 SRX386202; B) NFYB: red - SRX037419, blue -SRX150508, and green -
678 SRX100471), which are smoothed with a rolling mean with a 5 bp window. Element
679 numbers in the tables indicates the number of peak regions obtained in a ChIP-seq
680 experiment, which overlap with a particular consensus motif binding site set. Figures
681 are adapted from ChIPSummitDB website: GATA1::TAL1:
682 http://summit.med.unideb.hu/summitdb/paired_shift_view.php?exp1=220&exp2=218
683 &exp3=5&motive=GATA1%3A%3ATAL1&motifid=89&limit=40&low_limit=-
684 40&mnelem=100; NFYB:
685 http://summit.med.unideb.hu/summitdb/paired_shift_view.php?exp1=2301&exp2=76
686 1&exp3=1597&motive=NFYB&motifid=175&limit=40&low_limit=-
687 40&mnelem=2000
688
689 **Additional file 1.** Supplementary notes and figures.
690 Figure S1: Schematic representation of the initial data processing. Processing starts
691 with data collection and proper naming. After processing and filtering steps, we get the
692 transcription factor binding sites in bed and bedgraph formats.
693 Figure S2: Summit prediction. Identification of local maxima within peak regions.
694 Figure S3: Peak filtering according to shape. (A) Peak with well-defined summit. We
695 filtered peaks depending on the symmetry of their two side (summit positions serves as
696 a symmetry axis) (B), the positions of the 2nd and 3rd quartiles (C), and the symmetry
697 between the read coverage of the two strands (D).
698 Figure S4: Pairing position weight matrices (PWMs) for processed ChIP-seq
699 experiments. (A) 2758 experiments could be paired to a proper JASPAR motif from
700 the downloaded and processed 3727 ChIP-seq experiments (B). This paired to 338

701 JASPAR CORE motifs from the 579 (C). The result was a table where the PWMs are
702 paired to their corresponding ChIP-seq experiments (D).
703 Figure S5: Motif optimization. JASPAR CORE motifs were optimized with the
704 findMotifsGenome program, which used the original PWMs and the merged peak
705 region set of corresponding ChIP-seq experiments (determined in the Motif- ChIP-seq
706 experiment pairing step).
707 Figure S6: Determining motif locations. To identify the location of motif instances, we
708 combined three different motif finding methods (MAST, HOMER, FIMO). The merged
709 peak region set of corresponding ChIP-seq experiments was used in the identification
710 (Figure 6A). To filter the identified motifs, we used the presented formula in Figure
711 6B. In the case of overlapping motifs, the motif with the highest Weighted Motif score
712 was selected.
713 Figure S7: Measuring the distances between motif centers and the surrounding
714 summits. We calculated the concrete distance between motifs and the neighboring
715 summits (measured in base pairs). We took into account all of the possible summits
716 from every experiment.
717 **Additional file 2.** Summary table of the transcription factor motif, which has identified
718 instances in ChIPSummitDB. The table describes the following basic information for
719 the motifs:
720 1. column: Name of the motif
721 2. column: JASPAR record/ID
722 3. column: motifURL
723 4. column: Motif length
724 5. column: SRA antibody name paired to a motif
725 6. column: used ChIP-seq experiments in consensus motif creation

726 **Additional file 3.** Summary table of the ChIP-seq experiments. The table describes the

727 following basic information for the experiments:

728 1. column: name of the experiments

729 2. column: ftp address of SRA files

730 3. column: layout of experient (paired/single)

731 4. column: HTTP address of SRA record of ChIP-seq experiment

732 5. column: Link to SRA record

733 6. column: all identified peaks

734 7. column: filtered peak number

735 8. column: Transcription factor /cofactor + DNA binding motif

736 9. column: ExperimentView HTTP address

737 10. column: Link to ExperimentView