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. Protein-protein interactions, co-binding, and DNA looping can explain this phenomenon. We uniformly processed 3727 human ChIP-seq data sets and determined the cistrome for 292 TFs together with the distances between the TF binding site (TFBS) centers and the ChIP-seq peak summits. In addition to providing a comprehensive human TFBS collection, the ChIPSummitDB database

and web interface allows to examine the topological arrangements of TF complexes on the DNA.

Keywords

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

BACKGROUND

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 different ChIP-seq experiments, and visualization, for example generating profiles, heat-maps and Venn-diagrams) [7]. The most critical step in such a pipeline is the peak calling. Different peak calling algorithms provides different results and the number of the determined peaks also depends on the number of the sequenced reads [8].

Today, raw data from more than 85000 human and mouse ChIP-seq experiments are available [9], which gives the opportunity to perform further analyses and/or to set up secondary databases using those data. Previously, such databases have been built based on different parameters of ChIP-seq analyses. For some databases (CODEX, BloodChIP, hmChIP), the experimental metadata were collected and the experiments were classified by the cell type [10–12]. For other databases, for example Cistrome Data Browser, GTDR, ChIP-Atlas, Factorbook, raw data from as many as possible ChIP-seq experiments were collected and processed ([13–16]. Most of these databases are not only a simple collection of ChIP-seq data and a display of ChIP-seq peaks. Factorbook, for example, has an interactive tool to examine the nucleosome and histone modification profiles around the ENCODE TF ChIP-seq peaks [16,17]. The gene transcription regulation database (GTRD) project, among other things, focuses on improving the peak calling procedure [14]. They use several peak calling algorithms and make clusters of overlapping results. ChIP-Atlas provides a tool for extensive co-localization and enrichment analyses [15]. TFBSbank focuses on annotating genomic localizations, finding co-binding proteins, and searching for *de novo* and known motifs within the peaks [18]. The Cistrome Data Browser combines ChIP-seq data with chromatin accessibility data and provides a convenient web interface to browse and download these data [13]. Most of the above-mentioned databases contain not only human but mouse data too (Cistrome Data Browser, Factorbook, CODEX, hmChIP), and, in some cases, ChIP-seq data for other species (GTRD, ChIP-Atlas, TFBSbank).

Enhancers are distant regulatory elements relative to transcription start sites [19]. They can be characterized by TF binding (GTRD, TFBSbank), certain histone marks (SEdb), and enhancer transcription (HACER) [20,21]. Because both TF binding and histone modifications can be detected by ChIP-seq, secondary ChIP-seq database tools can

predict enhancer sequences as well if they determine the cognate binding sites under the peaks of TF ChIP-seq experiments. The ChIP-seq peaks and the protein-binding DNA-motifs under them can be further analyzed, and the resulted data can be integrated into a higher level database, such as the TRRUST [22].

ChIP-seq databases can provide tools to search, analyze, visualize, and download existing ChIP-seq data. Our previous results, however, demonstrated that ChIP-seq peak summits can also help to understand the topological arrangements, the spatial position(s) of different proteins bound to the DNA double helix. Therefore, our aim was to extend our CCCTC-binding factor (CTCF) peak summit-based analysis to every available ChIP-seq transcription factor, which was examined in human ChIP-seq experiments [23]. For the study, we have manually selected 3727 ChIP-seq experiments, which representing a very large number of human TFs and co-factors [9]. Since determining the correct positions of peak summits is critical for the analysis, we have developed a robust peak filtering pipeline, by which the positions of peak summits and the mapped TFBS motifs could be compared. Therefore, based on the ChIP-seq peak regions for each TF, we defined the corresponding consensus motif binding site sets for each of them. In addition to the consensus motif binding site sets, the ChIPSummitDB contains the distances between each pair of mapped consensus motifs and ChIP-seq peak summit.

The web interface of the ChIPSummitDB can display data in different views. Using the GenomeView module, users can browse the genome for ChIP-seq peaks and consensus motif binding site sets. The MotifView option can display the average distances between the centers of consensus motif binding sites and the ChIP-seq peak summit positions for each ChIP-seq experiment with overlapping peaks. Three different ChIP-seq experiments can be compared in the PairShiftView module. Using the dbSNPView

97 option, users can evaluate whether a dbSNP entry overlaps with one or more consensus
98 motif binding sites from our database [24].
99 In addition to browsing all processed data, the analysis of the scatterplots provided in
100 the MotifView led us to hypothesize that the extent of the standard deviation of the
101 motif center vs peak summit distances may be proportional to the closeness of the given
102 protein to the DNA double helix.

103 CONSTRUCTION AND CONTENT

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

114 1. ChIP-seq data collection from public databases

115 We searched for human ChIP-seq experiments in the NCBI SRA database according to
116 its status as per November 1, 2017. We used the NCBI's run selector feature to
117 download all available metadata associated with the selected experiments [9]. A custom
118 PERL script was used to process the downloaded data in XML table format and to give
119 a unique (descriptive) name to every experiment. The names include the species, the
120 tissue, the cell line (if available), the pathology (eg. normal or cancer), the ChIP target

121 protein, and the experiment's SRA database ID. Our aim was to restrict the analysis to
122 transcription factors and to other non-histone proteins, thus experiments with other type
123 of proteins were filtered out by using a script. For simplicity and to avoid redundancy,
124 we processed only the normal (without any specific treatment, which influence the TF-
125 DNA interaction) ChIP-seq experiments. The final list contained 4052 experiments and
126 was converted into a table with the proper format (BED, BEDGRAPH) for further
127 processing (Additional file 1: Figure S1).

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

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

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

141 The basic analysis provided the peak regions for each ChIP-seq experiment in two
142 forms, a BED file containing the borders of the HOMER predicted peak regions and a
143 BEDGRAPH file with the coverage values of the extended reads within these peak
144 regions [27]. It is possible to have more than one binding site within a peak region,

145 which can result in more than one summit. We employed the PeakSplitter program to
146 determine such summit positions using the BEDGRAPH files as input [28]. The result
147 of this step is a BED file for each experiment containing the positions of the identified
148 summit(s) for each peak (Additional file 1: Figure S2).

149 **4. Peak filtering**

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

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

165 We assume that if the antibody used during the immunoprecipitation is against a TF,
166 then the cognate binding site for that TF will be enriched in the peak set of the given
167 ChIP-seq experiment. During the basic analysis, we determined the enriched motifs for
168 each peak set. The problem was that even if the immunoprecipitated TF was the same,

169 the resulting *de novo* motif could be slightly different. Also, the *de novo* motif finding
170 algorithms usually give more than one enriched motif and it is unclear whether the best
171 one is the cognate binding site. To precisely determine the peak summit motif center
172 distances, however, we need to use the same binding site for every experiment with the
173 same TFBS. Therefore, we chose a reverse approach. Based on the antibody used
174 during the immunoprecipitation, we assigned a JASPAR core consensus motif to each
175 experiment by hand. This resulted in a table where there is a corresponding JASPAR
176 consensus motif for each TF experiment (Additional file 1: Figure S4, Additional file
177 2) [29].

178 **6. Motif optimization**

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

189 **7. Remapping motifs**

190 Once we had all the optimized consensus motif matrices, we needed to map them back
191 to the genome. For each optimized JASPAR consensus motif matrix, we had a list of
192 corresponding ChIP-seq experiments (Additional file 1: Figure S4) [29]. We took the

193 peak regions for those experiments and merged them to create the merged peak regions
194 for each matrix. This is an important and, so far, a unique step among the similar ChIP-
195 seq databases. Using this method, we specifically determined the possible transcription
196 factor binding sites. For mapping, we used three different programs and we kept only
197 the positions where at least two of them gave a hit for the final consensus motif binding
198 site sets (Additional file 1: Figure S6) [27,30,31].

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

200 After the peak filtering in step 4, we got the list of peak summits for each ChIP-seq
201 experiment. In step 7, we got the ChIP-seq verified positions of TFBSs for each
202 consensus motif. The majority of the peak regions do not contain the cognate binding
203 site, even for TF ChIP-seq experiments. We hypothesize that in these cases the given
204 TF can be part of a complex, which is bound to the DNA through another TF, which in
205 turn is bound to its cognate binding site, which is also present in our consensus binding
206 site set. Therefore, we investigated such cases as well. To do this, we calculated every
207 distance between consensus motif centers and the nearby ChIP-seq peak summit
208 positions. The calculation is motif based, which means that the output is a list for each
209 consensus motif set (i.e. a given transcription factor binding site), which contains the
210 distances of every peak summit inside the 50 bp range of the given consensus motif
211 binding site (Additional file 1: Figure S7) [32]. Practically, if we see a consensus motif
212 (e.g. CTCF), we will have the ChIP-seq verified instances of that motif in the genome
213 (88906) and a list of distances for each of the 3727 ChIP-seq experiments. These lists
214 were processed further in the database. The number of distances in these lists were used
215 as a cutoff value in the MotifView. The standard deviation of the distances in these lists
216 was also calculated and indicated in the Y-axis of the MotifView.

217 This processing of raw ChIP-seq read and metadata resulted in the following tables and
218 files:

219 1. The HOMER *homerpeaks.txt files for each experiment. These files contain the
220 peak regions and other parameters used in peak finding.
221 2. A peak region and a summit bed format file for each experiment.
222 3. A table with all the metadata related to the ChIP-seq experiments.
223 4. A table with the average distances between each consensus binding site set and
224 experiment.
225 5. A table with all the distances between individual consensus binding sites and
226 each overlapping peak summit.
227 6. The HOMER *de novo* motif finding results for every experiment.
228 7. The optimized consensus motif matrices for every consensus binding site set.

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

235 **UTILITY AND DISCUSSION**

236 The ChIPSummitDB address two main aims. First, the database is a comprehensive
237 collection of ChIP-seq verified binding sites for 292 different TFs. Second, the database
238 provides a new tool for analyzing the topological relationship between bound and co-
239 bound proteins on the TFBSs.

240 There are six different views for exploring the database. From the main page, the user
241 can either choose one of them under the search view tab or click any of the example
242 pictures in order to go directly to a view page.

243 **User interfaces**

244 MotifView is the summary page for each of the examined TFs. During the construction
245 of the database, we have determined the ChIP-seq verified binding site sets for 292
246 different TFs (Additional file 2). For each ChIP-seq experiment, we also counted how
247 many times the filtered peak summits of that experiment overlapped with a binding site
248 of the chosen TF. Furthermore, we calculated the exact distances between each
249 occurrence of a peak summit and binding site pair (Additional file 1: Figure S7). In
250 MotifView, one can see the averages of these distances for each experiment, where the
251 number of calculated distances is above the given threshold. Thus, the scatterplot
252 presented in MotifView shows the average distances of the overlapping peak summits
253 relative to the center of the adjusted TFBS. In the Y-axis, one can choose between
254 showing the number of overlapping peak summits and TFBS pairs or the standard
255 deviation of those distances.
256 Each point on the plot represents one ChIP-seq experiment. Circles represent
257 experiments where the antibody was against a TF with a defined TFBS, while triangles
258 show co-factors and other proteins attached to the primarily bound TF. On the right
259 side of the plot, there is a color-coded list of antibodies or cell types used in these
260 experiments. The list can be sorted alphabetically or by the number of occurrences. By
261 clicking on these squares (TFs) or triangles (co-factors and other proteins), one can
262 change the displayed experiments. For example, the user can choose and examine only
263 the experiments belonged to one or more antibodies or cell types. This allows users to

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

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

314 In the recent release, the ChIPSummitDB contains data from the analysis of 3727
315 human ChIP-seq experiments. In the ExperimentView, the user can see an overview of
316 the main attributes for each experiment. Most importantly, there are links to the NCBI
317 SRA database [26]. The number of sequencing reads and the number of filtered peaks
318 called by the HOMER findPeaks program is also listed [27]. During the analysis
319 pipeline, we have determined the most enriched *de novo* and known motifs for each
320 experiment. The link for the results of this search is also located on this page.
321 The GenomeView allows users to see the database content (consensus binding sites,
322 peaks, etc.) in a genomic context through a web browser interface. This view is
323 implemented in the JBrowse framework [34]. The GenomeView can be used as a
324 standalone web page where the users can select tracks to load and display. Users can
325 select from the 292 consensus motif sets, the 3727 experiments or from miscellaneous
326 tracks like genomic features or know SNPs. Therefore, users can compare any
327 combinations of experiments versus consensus motif binding sites. Users can also get
328 to this JBrowse interface from the MotifView, the PairShiftView, and the VennView
329 after selecting up to three experiments. In this case, the consensus motif set of the
330 chosen motif and the overlapping peaks and summits of the three chosen experiments
331 will be displayed initially. The user can display any other previously mentioned tracks.
332 Our database provides a comprehensive catalogue of experimentally verified
333 transcription factor binding sites in the human genome. As the cost of whole genome
334 sequencing is drastically decreasing, the number of variations associated with a certain
335 phenotype is rapidly increasing. Most of these variations are in intronic or intergenic
336 regions. Therefore, there is a great interest in determining the overlap of transcription
337 factor binding sites. The dbSNPView allows users to check these cases. Users can enter
338 either a genomic region or a dbSNP ID [24]. In the first case, the webpage will then

339 display the given region with the variations on it and also the overlapping consensus
340 motif binding sites from our database. The overlapping SNPs are highlighted in red.
341 Either clicking on them or entering the dbSNP accession number directly into the search
342 field leads to the enlarged dbSNPView. Here the reference genome sequence together
343 with the logo of the consensus motif and the overlapping SNPs can be seen. This view
344 is useful for examining how severely the altered base can affect the transcription factor
345 binding. There is also a button to check which experiments give an overlapping peak
346 with the given variation.

347 **The Novelty of the ChIPSummitDB**

348 The ChIPSummitDB is the first ChIP-seq related database that analyzes and shows the
349 peak summit and consensus motif binding site center distances. The web interface of
350 the database provides several tools for displaying this kind of topological data. In the
351 MotifView, every experiment (antibodies used for the immunoprecipitation) that is
352 above the threshold number of overlapping peaks for a chosen consensus motif is
353 shown. Either direct binding (in the cases where the immunoprecipitated protein
354 recognizes and binds to the given binding site) or indirect binding is indicated. We
355 hypothesize that the standard deviation of the distances of the peak summit and binding
356 site centers shows the directness of the binding. For example, based on the standard
357 deviations, there are at least three groups of interacting proteins at the YY1 consensus
358 motif binding site
359 (http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=10000&minid=1&m
360 nelem=1000&mxelem=120000&motive=YY1). In the first group, the YY1 ChIP-seq
361 experiments have a standard deviation value between 16 and 22 [35]. In the next layer,
362 the CTCF and the cohesin proteins and some other TF and co-factor have standard
363 deviation values between 23 and 27 [6]. Interestingly, there are many other ChIP-seq

364 experiments, which have a standard deviation above 27 but still have more than 1000
365 peaks (the threshold value set here), which are overlapping with the YY1 consensus
366 motif binding sites throughout the genome. If we further analyze the scatterplot, we
367 may notice that the YY1 experiments are grouped around 2 at the X-axis, while the
368 CTCF experiments are grouped around -2, as most of the other TFs and cofactors from
369 the third group. Based on these observations, we can hypothesize that the YY1 binds to
370 its cognate binding site. The CTCF-cohesin complex binds to the already bound YY1
371 protein in the same horizontal arrangement as can be seen at the CTCF binding sites
372 (http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=10000&minid=1&m
373 nelem=100&mxelem=120000&motive=CTCF). The proteins in the third group most
374 likely bind to the CTCF-cohesin complex. There are, of course, not only CTCF-cohesin
375 proteins in the second group (e.g. SRF or auts2) [36]. Proteins in the second group are
376 most probably directly bound to the YY1 protein.
377 Besides the already analyzed CTCF-cohesin topology, the ChIPSummitDB provides
378 further evidence that the peak summit versus the binding site based analysis indicates
379 topological arrangements for TF-DNA complexes. One good example is the
380 GATA1::TAL1 composite element (http://jaspar.genereg.net/matrix/MA0140.2/) [37].
381 Here, the ChIPSummitDB clearly confirms the experimental results. The topological
382 arrangement of the two transcription factors is clearly visible at both the MotifView
383 (Fig. 3a)
384 (http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=10000&minid=1&m
385 nelem=100&mxelem=120000&motive=GATA1%3A%3ATAL1) and at the
386 PairShiftView (Fig. 3a)
387 (http://summit.med.unideb.hu/summitdb/paired_shift_view.php?exp1=220&exp2=21

388 8&exp3=undefined&motive=GATA1::TAL1&motifid=89&limit=25&low_limit=-
389 25&formminid=1&formmaxid=10000&mnelem=100&formmaxelem=120000).
390 Theoretically, we expect that the peak summit will be in the middle of the transcription
391 factor binding site. We have already shown, in the case of the CTCF-cohesin complex,
392 that this is not necessarily true [38]. The ChIPSummitDB provides even more extreme
393 cases. For example, in the case of the NFYB motif observed in the MotifView (Fig. 3b)
394 (http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=10000&minid=1&m
395 nelem=2000&mxelem=120000&motive=NFYB), the average distance values that the
396 NFYB binds are upstream, while the USF1 binds downstream of the motif. If we further
397 scrutinize the topology on the PairShiftView, which shows the real distribution of the
398 distances, we can see even more extremities (Fig. 3b)
399 (http://summit.med.unideb.hu/summitdb/paired_shift_view.php?exp1=2301&exp2=7
400 61&exp3=1597&motive=NFYB&motifid=175&limit=40&low_limit=-
401 40&mnelem=2000). In the case of NFYB, most of the distances between the
402 overlapping consensus motifs and peak summits are around the +15 - +17 positions. In
403 the case of USF1, however, the majority of the distance values are clustered around the
404 -18 position. It is also remarkable that in the distribution curves, there are also other
405 smaller shoulders.

406 **Conclusions**

407 ChIPSummitDB is the first ChIP-seq database based on a transcription factor binding
408 site centered analysis of peak summits. The database convincingly confirms our
409 previous hypothesis that if the different proteins are sitting on the DNA not exactly
410 above each other, then the average peak summit position will display a shifted value
411 relative to the motif center. There can be numerous reasons for this phenomenon. One
412 obvious example is when two different transcription factor binds nearby on a composite

413 element [37]. Surprisingly, there are cases, when a transcription factor is bound to its
414 cognate site in the DNA and somehow a different protein without a DNA binding
415 domain that is bound to that TF gets so close to the DNA double helix that it will
416 crosslink to it during the experiment. This will result in a shifted peak summit versus
417 motif center value in our analysis (as can be seen in MotifView and PairShiftView in
418 our database). We showed this shift for cohesion proteins (RAD21, SA1/2, and
419 SMC1/3), but this shift can also be recognized in other cases (e.g. at the ETS TFs like
420 the ELF1,
421 http://summit.med.unideb.hu/summitdb/motif_view.php?maxid=2000&minid=1&mn
422 elem=3000&mxelem=120000&motive=ELF1) [23].
423 The detailed analysis of the ChIP-seq summit and motif center positions led us to a new
424 hypothesis: Taking a consensus binding site set (ChIP-seq verified binding sites for a
425 given transcription factor), the closer a given protein is to the DNA, the lower the
426 standard deviation of the distances between overlapping peak summits versus motif
427 center pairs. In other words, if a protein is very close to the DNA double helix, which
428 means in most cases that the protein is bound to the DNA, the resulting ChIP-seq peak
429 summits will more likely be centered in the middle of the DNA region covered by the
430 protein. This recognition can help us better understand how the protein complexes are
431 built on the DNA starting from binding of a transcription factor to its cognate binding
432 site.
433 In the MotifView of the different consensus binding site sets, ChIP-seq experiments
434 made by different protein antibodies give overlapping peaks on the same sites. An
435 interesting biological question is, if we take three different experiments (e.g. CTCF,
436 RAD21, and Znf143) from the same type of cells (GM12878) and analyze their peaks,
437 is there overlapping at the CTCF consensus binding sites and are they giving peaks on

438 the same sites? The VennView in the ChIPSummitDB gives an answer to this question
439 (http://summit.med.unideb.hu/summitdb/venn_diagramm.php?exp1=2672&exp2=273
440 4&exp3=2714&motive=CTCF&motifid=33&mnelem=10000) and, thus, may help to
441 elucidate the role of the different TF protein complexes on the DNA.
442 Besides these completely new features, the ChIPSummitDB provides a comprehensive,
443 experimental based collection of transcription factor binding sites. The site can be
444 browsed in the GenomeView and we have also developed a dbSNP view that allows
445 users to check whether a given SNP is overlapping a TFBS [24,34]. This feature can be
446 useful in determining the consequences of non-coding mutations.

447 **List of abbreviations**

448 bp: base pair
449 BWA: Burrows-Wheeler Aligner
450 ChIP-seq: Chromatin Immunoprecipitation Sequencing
451 CTCF: CCCTC-binding factor
452 DB: Database
453 dbSNP: Single Nucleotide Polymorphism Database
454 ELF1: E74-like factor 1
455 ETS1: E26 Oncogene Homolog 1
456 GATA1: GATA-binding factor 1
457 GM12878: Lymphoblastoid cell line
458 HOMER: Hypergeometric Optimization of Motif EnRichment
459 HTML: HyperText Markup Language
460 HTTP: HyperText Transfer Protocol
461 ID: identification number/identifier
462 NCBI: National Center for Biotechnology Information

463 NFYB: Nuclear transcription factor Y subunit beta

464 PHP: Hypertext Preprocessor

465 PWM: Position weight matrix

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

467 SA1: Stromal Antigen 1

468 SMC1/3: Structural maintenance of chromosomes proteins

469 SNP: Single Nucleotide Polymorphism

470 SQL: Structured Query Language

471 SRA: Sequence Read Archive

472 TAL1: T-cell acute lymphocytic leukemia protein 1

473 TF: transcription factor

474 TFBS: transcription factor binding site

475 USF1: Upstream stimulatory factor 1

476 XML: Extensible Markup Language

477 YY1: Yin Yang 1

478 ZNF143: Zinc Finger Protein 143

479 **Declarations**

480 **Ethics approval and consent to participate**

481 Since this study carries out exclusively in silico analysis of already published data, it

482 does not need ethics committee approval, as it did not constitute biomedical research.

483 However, reporting is consistent with all ethical requirements.

484 **Consent for publication**

485 No consent for publication is required.

Funding

The financial support is provided 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 the statistical analyses and data filtering. EB and EC evaluated the results and wrote the manuscript.

Availability of data and materials

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

Competing interests

The authors declare no conflict of interest.

Acknowledgments:

The authors would like to thank Ferenc Marincs and Karen L. Uray for their critical discussion and technical assistance. Data analyses were performed on the High Performance Computing cluster at the Genomic Medicine and Bioinformatic Core

508 Facility of Department of Biochemistry and Molecular Biology, University of
509 Debrecen.

510 **References**

511 1. Ren B, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, et al. Genome-
512 Wide Location and Function of DNA Binding Proteins. 2000;2306–10.
513 2. Vishwanath R. Iyer et al. Genomic binding sites of the yeast cell-cycle
514 transcription factors SBF and MBF. 2001;
515 3. Gordan R, Hartemink AJ, Bulyk ML. Distinguishing direct versus indirect
516 transcription factor-DNA interactions. Genome Res. United States; 2009;19:2090–
517 100.
518 4. Xiao T, Wallace J, Felsenfeld G. Specific sites in the C terminus of CTCF interact
519 with the SA2 subunit of the cohesin complex and are required for cohesin-dependent
520 insulation activity. Mol Cell Biol. United States; 2011;31:2174–83.
521 5. Bartke T, Vermeulen M, Xhemalce B, Robson SC, Mann M, Kouzarides T.
522 Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell
523 [Internet]. 2010;143:470–84. Available from:
524 https://www.ncbi.nlm.nih.gov/pubmed/21029866
525 6. Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, et al. YY1 and CTCF
526 orchestrate a 3D chromatin looping switch during early neural lineage commitment.
527 2017;1–14.
528 7. Barta E. Command line analysis of ChIP-seq results. EMBnet.journal; Vol 17, No 1
529 Next Gener Seq Data Anal - 1014806/ej171209 [Internet]. 2011; Available from:
530 http://journal.embnet.org/index.php/embnetjournal/article/view/209/480
531 8. Koohy H, Down TA, Spivakov M, Hubbard T. A Comparison of Peak Callers Used
532 for DNase-Seq Data. 2014;

533 9. Leinonen R, Sugawara H, Shumway M, Collaboration on behalf of the INSD. The
534 Sequence Read Archive. Nucleic Acids Res. Oxford University Press; 2011;39:D19–
535 21.
536 10. Sánchez-Castillo M, Ruau D, Wilkinson AC, Ng FSL, Hannah R, Diamanti E, et
537 al. CODEX: a next-generation sequencing experiment database for the
538 haematopoietic and embryonic stem cell communities. Nucleic Acids Res [Internet].
539 2014/09/30. Oxford University Press; 2015;43:D1117–23. Available from:
540 https://www.ncbi.nlm.nih.gov/pubmed/25270877
541 11. Chacon D, Beck D, Perera D, Wong JWH, Pimanda JE. BloodChIP: a database of
542 comparative genome-wide transcription factor binding profiles in human blood cells.
543 Nucleic Acids Res. England; 2014;42:D172-7.
544 12. Chen L, Wu G, Ji H. hmChIP: a database and web server for exploring publicly
545 available human and mouse ChIP-seq and ChIP-chip data. Bioinformatics [Internet].
546 2011/03/30. Oxford University Press; 2011;27:1447–8. Available from:
547 https://www.ncbi.nlm.nih.gov/pubmed/21450710
548 13. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a
549 data portal for ChIP-Seq and chromatin accessibility data in human and mouse.
550 Nucleic Acids Res. England; 2017;45:D658–62.
551 14. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F. GTRD: a database of
552 transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids
553 Res. England; 2017;45:D61–7.
554 15. Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O. ChIP-Atlas : a data-mining
555 suite powered by full integration of public ChIP-seq data. 2018;1–10.
556 16. Wang J, Zhuang J, Iyer S, Lin X-Y, Greven MC, Kim B-H, et al. Factorbook.org:
557 a Wiki-based database for transcription factor-binding data generated by the

558 ENCODE consortium. Nucleic Acids Res [Internet]. 2012/11/29. Oxford University
559 Press; 2013;41:D171–6. Available from:
560 https://www.ncbi.nlm.nih.gov/pubmed/23203885
561 17. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al.
562 ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.
563 Genome Res [Internet]. Cold Spring Harbor Laboratory Press; 2012;22:1813–31.
564 Available from: https://www.ncbi.nlm.nih.gov/pubmed/22955991
565 18. Chen D, Jiang S, Ma X, Li F. TFBSbank: a platform to dissect the big data of
566 protein-DNA interaction in human and model species. Nucleic Acids Res [Internet].
567 2016/11/24. Oxford University Press; 2017;45:D151–7. Available from:
568 https://www.ncbi.nlm.nih.gov/pubmed/27899608
569 19. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer
570 action. Science. United States; 1998;281:60–3.
571 20. Jiang Y, Qian F, Bai X, Liu Y, Wang Q, Ai B, et al. SEdb: a comprehensive
572 human super-enhancer database. Nucleic Acids Res. England; 2019;47:D235–43.
573 21. Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y. HACER: an atlas of human
574 active enhancers to interpret regulatory variants. Nucleic Acids Res. England;
575 2019;47:D106–12.
576 22. Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded
577 reference database of human and mouse transcriptional regulatory interactions.
578 Nucleic Acids Res. England; 2018;46:D380–6.
579 23. Nagy G, Czipa E, Steiner L, Nagy T, Pongor S, Nagy L, et al. Motif oriented
580 high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and
581 cohesin proteins on DNA. BMC Genomics. 2016;17:637.
582 24. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al.

583 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res [Internet]. Oxford
584 University Press; 2001;29:308–11. Available from:
585 https://www.ncbi.nlm.nih.gov/pubmed/11125122
586 25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
587 transform. Bioinformatics. England; 2009;25:1754–60.
588 26. Bethesda (MD): National Center for Biotechnology Information (US). Sequence
589 Read Archive Submissions Staff. Using the SRA Toolkit to convert .sra files into
590 other formats. 2011.
591 27. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple
592 combinations of lineage-determining transcription factors prime cis-regulatory
593 elements required for macrophage and B cell identities. Mol Cell. United States;
594 2010;38:576–89.
595 28. Salmon-Divon M, Dvinge H, Tammoja K, Bertone P. PeakAnalyzer: Genome-
596 wide annotation of chromatin binding and modification loci. BMC Bioinformatics
597 [Internet]. 2010;11:415. Available from: https://doi.org/10.1186/1471-2105-11-415
598 29. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee
599 R, et al. JASPAR 2018: update of the open-access database of transcription factor
600 binding profiles and its web framework. Nucleic Acids Res. 2018;46:D260–6.
601 30. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a
602 flexible statistical framework for assessing transcriptional changes and characterizing
603 heterogeneity in single-cell RNA sequencing data. Genome Biol. London: BioMed
604 Central; 2015;16:278.
605 31. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given
606 motif. Bioinformatics. Oxford University Press; 2011;27:1017–8.
607 32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing

608 genomic features. Bioinformatics. 2010;26.
609 33. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for
610 Deoxyribose Nucleic Acid. Nature [Internet]. Nature Publishing Group;
611 1953;171:737. Available from: https://doi.org/10.1038/171737a0
612 34. Skinner ME, Uzilov A V, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-
613 generation genome browser. Genome Res [Internet]. Cold Spring Harbor Laboratory
614 Press; 2009;19:1630–8. Available from:
615 https://www.ncbi.nlm.nih.gov/pubmed/19570905
616 35. Yant SR, Zhu W, Millinoff D, Slightom JL, Goodman M, Gumucio DL. High
617 affinity YY1 binding motifs: identification of two core types (ACAT and CCAT) and
618 distribution of potential binding sites within the human beta globin cluster. Nucleic
619 Acids Res [Internet]. 1995;23:4353–62. Available from:
620 https://www.ncbi.nlm.nih.gov/pubmed/7501456
621 36. Basu A, Wilkinson FH, Colavita K, Fennelly C, Atchison ML. YY1 DNA binding
622 and interaction with YAF2 is essential for Polycomb recruitment. Nucleic Acids Res
623 [Internet]. 2013/11/26. Oxford University Press; 2014;42:2208–23. Available from:
624 https://www.ncbi.nlm.nih.gov/pubmed/24285299
625 37. Han GC, Vinayachandran V, Bataille AR, Park B, Chan-Salis KY, Keller CA, et
626 al. Genome-Wide Organization of GATA1 and TAL1 Determined at High
627 Resolution. Mol Cell Biol [Internet]. American Society for Microbiology;
628 2015;36:157–72. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26503782
629 38. Nagy G, Czipa E, Steiner L, Nagy T, Pongor S, Nagy L, et al. Motif oriented
630 high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and
631 cohesin proteins on DNA. BMC Genomics. 2016;17.
632

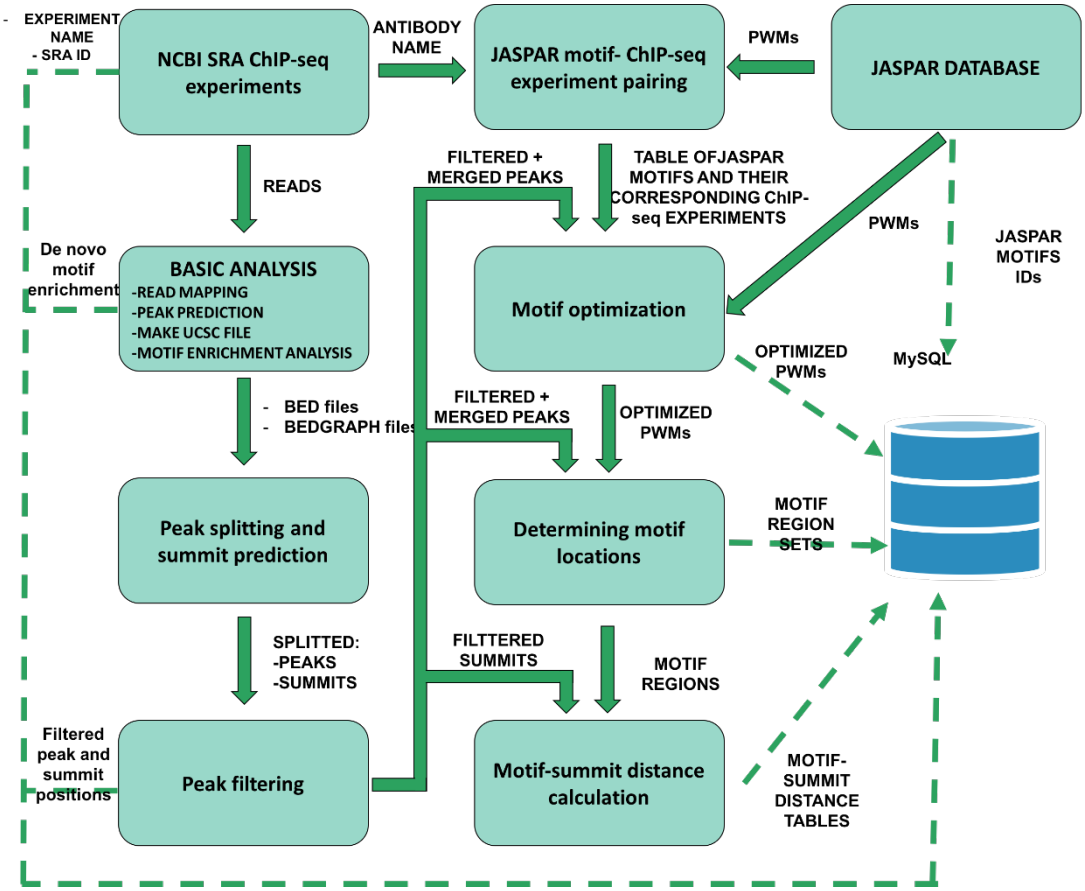

Fig. 1

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.

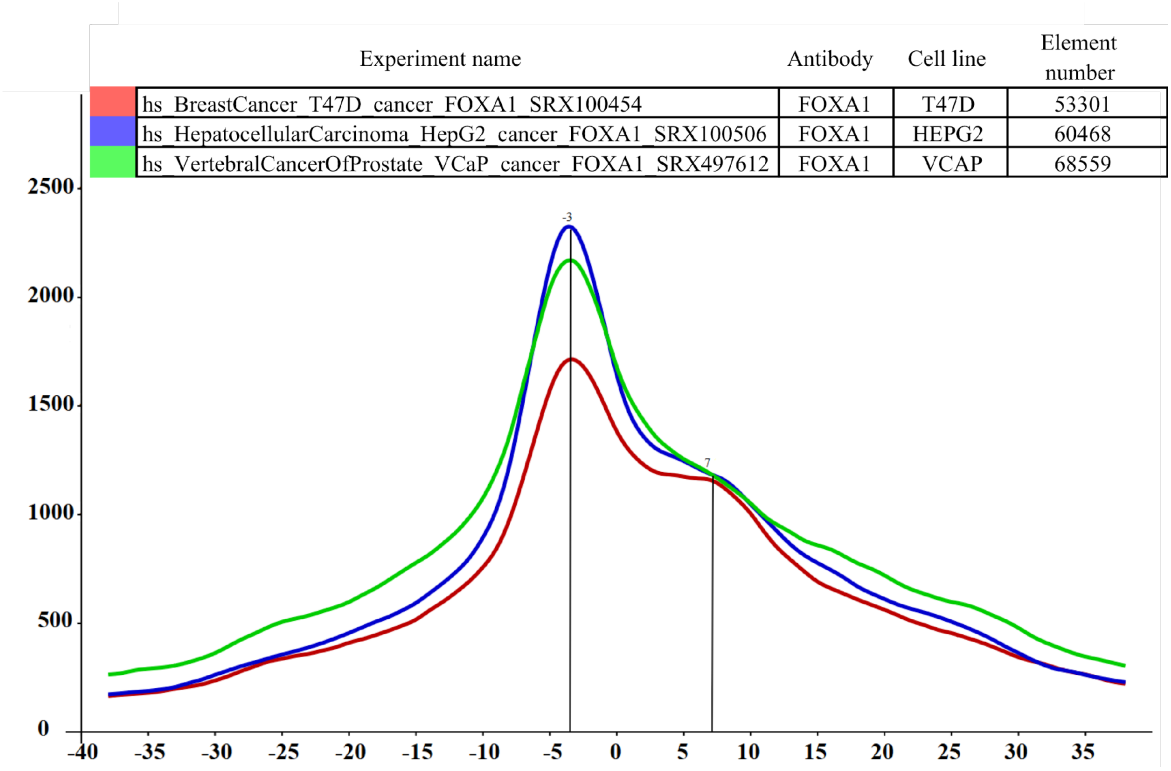

Fig. 2

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.

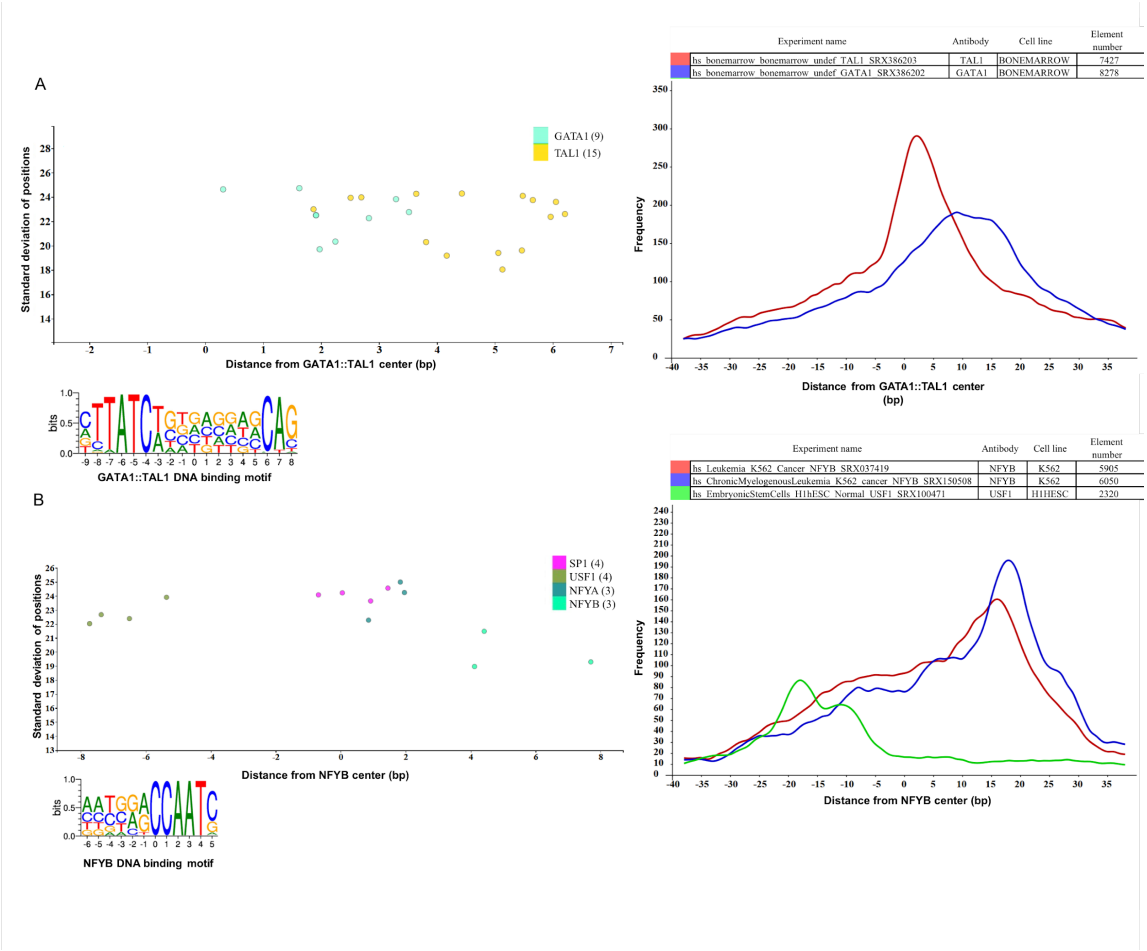

Fig. 3

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 scatter plots display the average summit positions (X axis) of factors relative to the center of the binding sites (motif logos represented below the scatterplots with reference points which are marked as “0” position). The distances are measured in base pairs (bp). The standard deviation of summit- motif center distances are shown on the Y axis. Each dot represents a single ChIP-seq experiment. 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

668 represented by a frequency curve (A) GATA1:TAL1: blue - SRX386203, red -
669 SRX386202; B) NFYB: red - SRX037419, blue -SRX150508, and green -
670 SRX100471), which are smoothed with a rolling mean with a 5 bp window. Element
671 numbers in the tables indicates the number of peak regions obtained in a ChIP-seq
672 experiment, which overlap with a particular consensus motif binding site set.

673 **Additional file 1. Supplementary notes and figures.**

674 **Figure S1: Schematic representation of the initial data processing.** Processing starts
675 with data collection and proper naming. After processing and filtering steps, we get the
676 transcription factor binding sites in bed and bedgraph formats.

677 **Figure S2: Summit prediction.** Identification of local maxima within peak regions.

678 **Figure S3: Peak filtering according to shape.** (A) Peak with well-defined summit.
679 We filtered peaks depending on the symmetry of their two side (summit positions
680 serves as a symmetry axis) (B), the positions of the 2nd and 3rd quartiles (C), and the
681 symmetry between the read coverage of the two strands (D).

682 **Figure S4: Pairing position weight matrices (PWMs) for processed ChIP-seq**
683 **experiments.** (A) 2758 experiments could be paired to a proper JASPAR motif from
684 the downloaded and processed 3727 ChIP-seq experiments (B). This paired to 338
685 JASPAR CORE motifs from the 579 (C). The result was a table where the PWMs are
686 paired to their corresponding ChIP-seq experiments (D).

687 **Figure S5: Motif optimization.** JASPAR CORE motifs were optimized with the
688 findMotifsGenome program, which used the original PWMs and the merged peak
689 region set of corresponding ChIP-seq experiments (determined in the Motif- ChIP-seq
690 experiment pairing step).

691 **Figure S6: Determining motif locations.** To identify the location of motif instances,
692 we combined three different motif finding methods (MAST, HOMER, FIMO). The

693 merged peak region set of corresponding ChIP-seq experiments was used in the
694 identification (Figure 6A). To filter the identified motifs, we used the presented
695 formula in Figure 6B. In the case of overlapping motifs, the motif with the highest
696 Weighted Motif score was selected.

697 **Figure S7: Measuring the distances between motif centers and the surrounding**
698 **summits.** We calculated the concrete distance between motifs and the neighboring
699 summits (measured in base pairs). We took into account all of the possible summits
700 from every experiment.

701 **Additional file 2. Summary table of the transcription factor motif, which has**
702 **identified instances in ChIPSummitDB.** The table describes the following basic
703 information for the motifs:

704 1. column: Name of the motif
705 2. column: JASPAR record/ID
706 3. column: motifURL
707 4. column: Motif length
708 5. column: SRA antibody name paired to a motif
709 6. column: used ChIP-seq experiments in consensus motif creation

710 **Additional file 3. Summary table of the ChIP-seq experiments.** The table describes
711 the following basic information for the experiments:

712 1. column: name of the experiments
713 2. column: ftp address of SRA files
714 3. column: layout of experient (paired/single)
715 4. column: HTTP address of SRA record of ChIP-seq experiment
716 5. column: Link to SRA record
717 6. column: all identified peaks

7. column: filtered peak number

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

9. column: ExperimentView HTTP address

10. column: Link to ExperimentView