An approach to identify potential transcription factors co-localized with CTCF in breast cancer cells

LCG UNAM 2022

Daianna Gonzalez Padilla

2022-02-18

Contents

Keywords

CTCF, ChIP-seq, cis-regulatory modules, breast cancer

Introduction

The ubiquitously expressed and highly conserved transcription factor CTCF is perhaps mostly known for its role as a genome organizer that interacts with cohesin to establish and/or maintain the chromatin loops and TAD structure. However, CTCF also has a crucial role in transcriptional regulation and it is well known that CTCF can act both as transcriptional repressor and activator [1-3] Additionally, it can also help to tether distal enhancers to their cognate promoters [4]. This transcription factor (TF) binds to DNA through its 11 zinc-finger domain (ZF) and maps of its binding sites through ChIP-seq have revealed that CTCF binds to tens of thousands of genomic sites including enhancers, gene promoters and regions inside gene bodies [5,6] indicating its wide-range regulatory function in the genome. Some of these sites are tissue-specific and others ultra-conserved.

CTCF can also attract other transcription factors and repressors and its exact functions appears to be determined by these associated transcription factors, by the location of the binding site relative to TSS of

genes and by the site's engagement in chromatin loops with other CTCF-binding sites, enhancers or gene promoters [7]. Since CTCF performs multiple roles, this protein shares chromatin binding sites with many other factors [8-10] to regulate gene expression. This shared DNA binding sites make up **cis-regulatory modules** (CRM), stretches of DNA, usually 100–1000 DNA base pairs in length, where a number of transcription factors bind to regulate expression of nearby genes. Unsurprisingly, cohesin binds to many of the chromosomal binding sites of CTCF [11-15] to mediate chromatin loop formation.

Importantly, there is evidence pointing to CTCF as a major tumor suppressor gene. It has been seen that CTCF is highly mutated in breast cancer [16]. In fact, more than half of the known missense CTCF mutations in cancer are located in its ZFs disrupting its DNA binding capability at specific loci leading to aberrant expression of cancer-related genes such as Myc, ARF and Igf2 [17] whose deregulation may contribute to malignant tumor phenotypes. Looping formation disruption by CTCF and cohesin loss-of-function could establish communication between an enhancer and a nearby promoter of an oncogene, thereby activating that gene. Tumor suppressor function of CTCF could also account for its role in DNA double-strand break repair as it is recruited in those sites [18]. Particularly, CTCF's role in breast cancer has been of recent interest since it shares binding sites with the Estrogen Receptor a (ERa) an enhancer activating transcription factor [19] and a key driver of breast cancer. ERa is currently the main target for breast cancer therapy, however, in such a complex disease as it is cancer, it is not expected to attribute it to a sole cause but to multiple factors that account for it. Thereby, identifying CTCF-interacting TFs within CRM in breast cell lines, can help to understand the underlying regulatory processes that promote proliferation in this type of cancer so that its causes could be broader understood and therapies ultimately more efficient.

This work pretends to gain insights into CTCF interactions with other TFs in breast cancer cells through an approach based on identifying TFs that co-occupy CTCF binding sites in the MCF-7 breast cell line based not directly on ChIP-seq identified sites but on motifs of those sequences that present similarity and high correlation with other known TFBS motifs. This approach presents the advantage of comparing the CTCF motifs discovered with multiple TFBS motifs in a database. In contrast, studies such as the one that revealed shared sites between CTCF and ERa [19] needed to analyze ChIP-seq data for each TF under study in order to identify overlapping binding sites between them.

Methods

ChIP-seq data of CTCF binding sites in MCF-7 cell line was obtained from **ReMap** database (See data sources and details in *Supplementary*). Next, **JASPAR** database is used to download in *TRANSFAC* format the position frequency matrix (PFM) of CTCF known binding sites as a reference motif for a positive control (See details in *Supplementary*).

Regulatory Sequence Analysis Tools **RSAT** is then used to make peak analysis and motif discovery through its Teaching tools: **fetch-sequences** is used to extract the peak sequences of the *.bed* file with the genomic coordinates using *hg38* human genome version, used in **ReMap** too. Then these sequences are sent as input to **peak-motifs** to get a composition analysis of the peak sequences and discover *de novo* motifs with them through ***oligo-analysis*** that compares the observed to expected frequencies of *k-mers* of lengths 6 and 7 and returns the top 5 of each one with higher significance scores under a Binomial distribution and a background model that takes *k-mers* frequencies from the test sequences. It also compute ***position-analysis*** that retains the top 5 oligos of the same lengths with heterogeneous distribution along the input sequences, comparing its observed occurrences per window to the expected ones under a homogeneous distribution background model. The third analysis picked is ***dyad-analysis*** which fundamentally accomplishes the same analysis as ***oligo-analysis*** but allows gaps within oligos to find noncontinuous motifs.

These oligos and dyads do not reflect the motifs themselves but compose them. So **peak-motifs** aligns and groups these sequences based on their similarity to create matrices and get the consensus motifs for their comparison with known motifs in *core non-redundant vertebrates* collection of **JASPAR** database in order to get the TFs those motifs belong to. The PFM of the reference CTCF motif is added to compare it with the discovered motifs and the rest of the options are taken as default. (See server commands and bioinformatic resources details in *Supplementary*).

Then, the RSAT tool **matrix-clustering** is used to merge the motif sequences by their similarity and to visualize the resultant clusters. It takes the *TRANSFAC* matrix of the discovered motifs and the matrix of the reference motif. The root motifs are retrieved and compared with **JASPAR** *nonredundant vertebrate* database with **compare matrices** tool. Then **matrix-scan** is used to measure the rate of coverage of the peaks for each non redundant discovered motif and the reference motif from the fasta file of the peak sequences.

To generate a negative control, **random genome fragments** uses the peaks as a template to take the same number of random sequences from the human genome and with the same lengths as peak sequences. Same process with **peak-motifs** is done with these random sequences.

Results and discussion

ChIP-seq data consisted of 3781 peaks whose length sum was 622,361 bp, the mean length was 164.602 bp and the median was 154 bp. (See commands in *Supplementary*).

After running **peak-motifs**, the sequence composition analysis of peaks showed that the minimum peak length was 134 bp and the maximum 733 bp (See *Supplementary F1A*). The nucleotide composition profile showed there were not huge differences in transition frequencies of each nucleotide (See *Supplementary F2A*). However, the nucleotide position profile revealed peak sequences have higher frequencies of G and C at central positions (See *Supplementary F2B*) which initially describes the general composition of CTCF binding sites which are located in the peak centers due to the experimental ChIP-seq methodology in which after crosslinking genetic material with bound TFs, DNA is randomly fragmented by sonication, generating DNA fragments whose ends are TFBS- flanking regions since TFBS were protected by the bound TFs. Only sequences with CTCF bound are retained and sequenced by *short read sequencing*.

Notably, dinucleotide composition profiles showed that transition frequency from C to G is the lowest one (See **F1A**) and CG dinucleotide is the one with less occurrences at central positions of peaks (See **F1B**). CG underrepresentation accounts for CpG avoidance in vertebrate genomes as a consequence of the high mutation rate of methylated CpG sites: spontaneous deamination of methylated cytosine results in thymine and the mismatch is resolved to A:T, causing a transition mutation that could reduce TF affinity for its binding sites [20,21]. Additionally, in both cases nucleotide and dinucleotide occurrences drop at peak ends because there are fewer peak sequences with such lengths (See **F1B** and *Supplementary F2B*). It is also possible to see that dinucleotides containing A and T such as AA, AT, TA and TT are less frequent at the center, that is consistent with lower frequencies of A and T nucleotides in these positions (See *Supplementary F2B*).

pr\suf	a	c	g	t	P_prefix
a	0.274	0.209	0.333	0.184	0.233
c	0.284	0.309	0.114	0.293	0.266
g	0.216	0.289	0.311	0.184	0.266
t	0.154	0.245	0.323	0.278	0.235

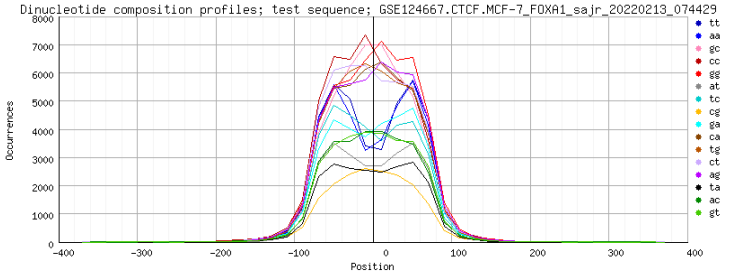

Figure 1: Dinucleotide composition profile. **A)** The table shows the frequencies of passing from one nucleotide (prefix) to another (suffix). **B)** Plot of dinucleotide frequency by position of all peaks.

Subsequently, there were 25 top discovered motifs from input peak sequences: 10 from oligo analysis, 10 from position analysis and 5 from dyad analysis. Of these, the ones characterized by over-representation with ***oligo-analysis*** had significance scores from 65.17 to 183.96 and E-values very close to 0. However, motifs from ***dyad-analysis*** had the highest scores, all with 350.00 and E-values of 0.00. Motifs characterized

by positional bias with ***position-analysis*** yielded the lowest significance scores from 34.85 to 60.85 and the highest E-values.

After the comparison between these motifs and *JASPAR* motifs, it resulted that all of them matched with at least one known motif. Importantly, all motifs discovered by the 3 previously described approaches matched the reference motif: two motifs from 7 nt oligo and position analysis had the highest normalized correlation scores (considering the percentage aligned of the match and the Pearson correlation), 0.905 and 0.906, respectively. In contrast, the 4 motifs with the lowest normalized correlation scores were all from position analysis and did not reach 0.6. These results served as a positive control that allowed to verify if discovered motifs from input ChIP-seq data certainly describe CTCF binding sites. This becomes more important in this study since a major issue to take into account when working with a breast cell line such as MCF-7 is the high mutagenesis of CTCF in its ZFs domain in this cancer [16] that may disrupt its DNA binding capability and therefore, the ChIP-seq results. All the same, the positive control confirmed CTCF bound to its sites.

When these 25 discovered motifs and the reference motif were clustered, they grouped in only 1 cluster (See **F2**), probably due to the small size of the data set used. Since there was only 1 cluster, there was only 1 root motif, the reference motif. When it was compared with *JASPAR* database it matched with its corresponding TF: CTCF (See *Supplementary F4*). The rate of coverage of the peaks with this motif was 89.2%, which means that 3373 of 3781 peaks had at least one match with it. The total number of matches in all peaks was 3861 (See commands in *Supplementary*). These results are in agreement with the fact of having found a match of this reference motif in all the discovered motifs from peak sequences.

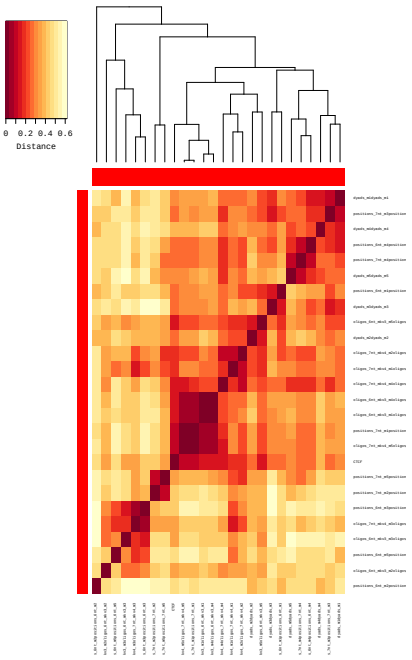

Figure 2: Cluster of the 25 discovered motifs and the reference motif. This heatmap shows how distant two motifs are based on their sequence similitude. The closer this distance is to 0, the more similar they are. Above is the dendrogram showing the relations between the motifs.

Apart from the reference motif, motifs also matched with others that belong to TFs other than CTCF: all motifs from oligo analysis, as well as 6 motifs from position analysis and 3 from dyad analysis matched with a motif of **CTCFL** with normalized correlations from 0.403 to 0.736. CTCFL is a paralogue TF of CTCF that harbors a nearly identical C2H2 ZF region to that of CTCF and consequently have similar binding specificity to DNA sequences, explaining why it was matched with CTCF discovered motifs [22]. Previous studies have reported that CTCFL gene is expressed in breast cell lines as well as in ~70% of the clinical specimens of breast cancer, but not in normal breast tissue [23], and that it is co-expressed with CTCF in

this type of cancer [23]. Additionally, CTCFL is more frequently amplified or transcriptionally activated, rather than mutated in cancers. And it has also been found that CTCFL contribute to the Warburg Effect in breast cancer cell lines through the alternative splicing of Pyruvate Kinase (PKM) to PKM2, an isoform that is known to contribute to aerobic glycolysis [25,26]. This supports that CTCFL was found as a possible CTCF-interacting TF in these MCF-7 cells.

Secondly, one of the motifs from oligo analysis also matched with a **THAP1** motif with a normalized correlation of 0.400. THAP1 harbors a C2H2 zinc-finger DNA-binding domain called THAP that recognizes an 11-nucleotide target sequence (agtacgGGCAa)[27]. This TF has shown to have significant association with overall survival of breast cancer patients [28] but no direct interaction with CTCF is reported. Besides, one motif of **INSM1** matched with one motif from oligo analysis and with another from position analysis with normalized correlations of 0.543 and 0.492, respectively. INSM1 is also a C2H2 five zinc-finger containing TF [29] whose relation with CTCF has not been described, nor its possible role in breast cancer. Three other motifs, two from oligo analysis and one from position analysis matched with a motif of **YY1** with normalized correlations of 0.499, 0.471 and 0.469, in that order. YY1 is a ubiquitous and multifunctional TF that binds to DNA through its C2H2 type zinc-fingers. YY1 contributes to enhancer-promoter structural interactions in a manner analogous to DNA interactions mediated by CTCF. YY1 binds to active enhancers and promoter-proximal elements and forms dimers that facilitate the interaction of these DNA elements [30]. The location of YY1-mediated interactions may be demarcated in development by a preexisting topological framework created by constitutive CTCF-mediated interactions [31]. Furthermore, several studies demonstrated YY1 overexpression in breast cancer cell lines and primary tumors leading to tumor promotion [32-35] but it has also been demonstrated that YY1 can act as a tumor suppressor [36,37]. Its similar functions to that of CTCF in looping formation suggest a feasible explanation of its similar DNA binding sites in enhancers and promoters and its found matched motif in this cell line could account for its role in breast cancer.

One motif of **ZEB1** matched with one discovered motif from position analysis with a normalized correlation of 0.435. This TF has two C2H2-type flanking zinc finger clusters which are responsible for interaction with paired CACCT(G) E-box-like promoter elements on DNA [38]. ZEB1 plays a vital role in embryonic development and cancer progression, including breast cancer. There is evidence for ZEB1's role in the progression of breast cancer mediated by inflammatory cytokines [39]; as well as for its involvement in the regulation of autophagy in several breast cancer cell models , augmenting the resistance of these cells to genotoxic drugs [40]. Much more research has been done around ZEB1 in breast cancer, confirming its relevant role in it. However, a functional association of this TF with CTCF has to be determined.

The last found TF was **SP9**, whose motif matched with one from oligo analysis with a normalized correlation of 0.473. This is another C2H2 TF with zinc finger DNA binding domain, that has already been reported by having multiple sites bound by CTCF at its locus through ChIP-seq analysis in E12.5 autopod cells [41] but further investigation of their common binding sites is required. Its role in breast cancer has not been studied.

As a negative control, these results were compared with the ones given by random sequences mostly lacking relevant meaning for TFs. The sequence lengths distribution is the same in both cases since there are the same number of sequences and with the same lengths (See *Supplementary F1*) as the peak sequences. Notably, nucleotide and dinucleotide composition profiles of these sequences show flatter curves at central positions, showing more homogenous frequencies than the ones from real TFBS which show variation according to the TFs affinity for their binding sites. It is also possible to notice that G and C nucleotides and G/C containing dinucleotides occurrences are lower than A and T nucleotides and A/T containing dinucleotides, the opposite from what real CTCF peaks showed (See **F2B**, **F3B** and *Supplementary F2B and F3B*). There were only 15 motifs found with significance scores from 1.67 to 76.46; even if these values are lower than those resulting from CTCF motifs, there still exist high significance scores that result from overrepresented sequences in the genome (not the peaks). E-values were from 0.021 to 3.5e-77, higher values than those of CTCF motifs. Again, dyad analysis gave the most significant results and position analysis did not yield any, which makes this last analysis the most reliable one to identify nonrandom motifs. As expected, none of these motifs matched with the reference CTCF motif and their normalized correlations with other motifs were from 0.465 to 0.753.

These results from random sequences prove that it is not the number and lengths of the peak sequences but

their content that actually yields significant results because they represent CTCF bounded sites.

pr\suf	a	c	g	t	P_prefix
a	0.336	0.168	0.232	0.264	0.299
c	0.357	0.249	0.046	0.349	0.201
g	0.297	0.204	0.248	0.251	0.200
t	0.226	0.199	0.240	0.335	0.300

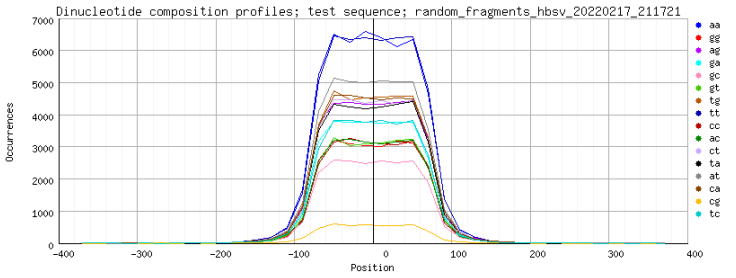

Figure 3: Dinucleotide composition profile of random sequences **A)** The table shows the frequencies of passing from one nucleotide (prefix) to another (suffix). **B)** Plot of dinucleotide frequencies by position of all random sequences.

In both peak motif analysis, position analysis yielded the less significant results in the discovered motifs, but as shown above, position analysis is very powerful to discard motifs from random sequences which may be overrepresented throughout the genome homogeneously. Its low values may be due to the small number of peaks of the data. On the other hand, the best results were obtained with dyad analysis. However, except for CTCFL, none of the found TFs resulted from this analysis. Such high values may be obtained for having higher probabilities to find dyads than their two composing monads together. For these reasons, analyzing the data with more than one program is convenient.

An important point to consider is that all TFs found from matched CTCF motifs, including CTCF itself, have C2H2-type zinc fingers domains which dictate high binding affinity to GC-boxes and lower binding affinities for CT and GT boxes [42], supporting the GC high frequencies observed in the discovered motifs (See **F1B** and *Supplementary F2B*). This GC affinity could only partly explain the similarities between their motifs because not all TFs with this DNA binding domain were identified, so other factors must explain it. Nevertheless, similar motifs are not contundent evidence of the presumed roles of their TFs in breast cancer: first, normalized correlations of these TFs motifs, except for CTCFL motif and the reference motif, are less than 0.6 so no strong inferences could be done around them; second, some of these elements could normally be associated with CTCF in such co-bound regions in other cell lines or types and not specifically in MCF-7; third, found TFs may be cell-line specific, as it has been seen that CTCF binding events are associated with genes highly expressed in each breast cell line [19], implicating that a TF do not have the same prevailing role in all breast cancer cell lines, making it hard to tell which TFs are effectively implicated in MCF-7 cells and to extrapolate their roles to other breast cancer cell lines.

Conclusions and perspectives

This study represents an initial approximation to the determination of TFs co-localized with CTCF in breast cancer cell lines for the future elucidation of their roles in regulatory processes involved in breast cancer, particularly, in MCF-7 cell line. Even if there are factors impeding to give concrete conclusions about the implications of finding these TFs in breast cancer cells, it is possible to say that this work showed CTCF have potential common binding sites with some TFs whose roles in breast cancer have already been studied and elucidated: CTCFL, THAP1, YY1 and ZEB1. Altogether these candidate TFs could form functional CRM with CTCF to regulate the transcription of their target genes in a coordinated manner.

As an initial study, further research must be done with other breast cancer cell lines and cell types, including normal healthy cells in order to compare TFs found in each case and determine which ones are effectively involved in breast cancer and to what extent in each cell line. It is also important to identify the nearby genes of these binding sites so that the regulation of target genes could be studied. The other two found

TFs, INSM1 and SP9, are cataloged for future research of their roles in breast cancer and their relationship with CTCF.

Supplementary material

Figures

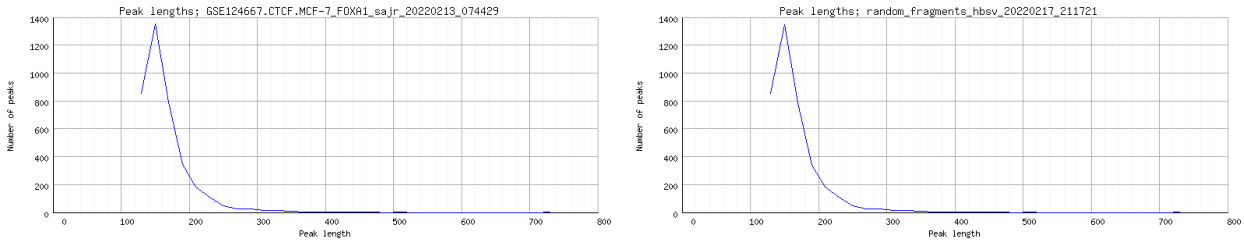

Figure 1: Peak and random sequence lengths distribution. **A)** Peak sequence lengths distribution from ChIP-seq data. **B)** Random sequence lengths distribution. Note that both graphs are the same since both data sets have the same dimensions.

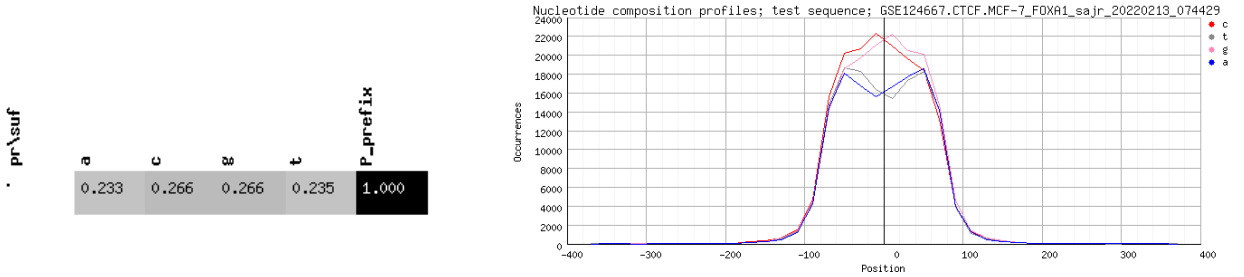

Figure 2: Nucleotide composition profile of CTCF peaks. **A)** The table shows the transition frequencies of each nucleotide in the peaks. **B)** Plot of nucleotide frequencies by position of all peaks.

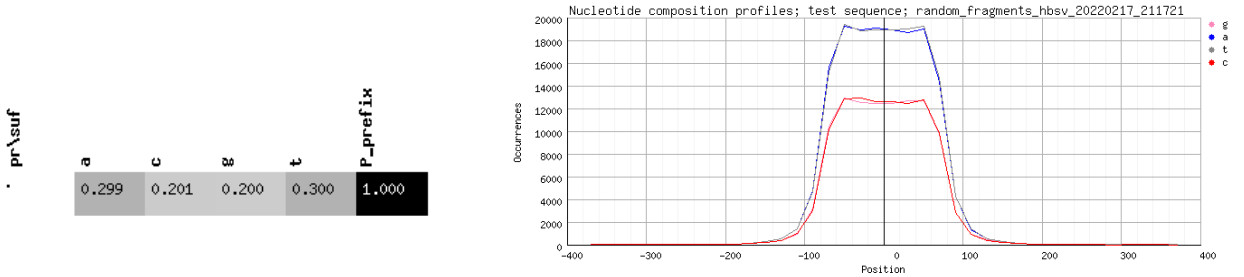

Figure 3: Nucleotide composition profile of random sequences. **A)** The table shows the transition frequencies of each nucleotide in the random sequences generated. **B)** Plot of nucleotide frequencies by position of all sequences.

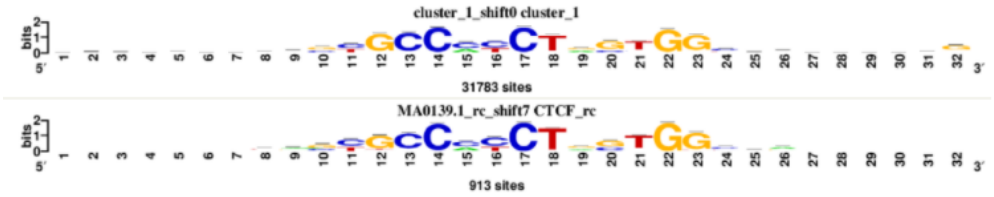

Figure 4: Match between the root motif of the cluster and the reference motif of *JASPAR*. They had 0.988 of Pearson correlation and a normalized correlation of 0.587.

Bioinformatics resources used for this work

The Table below indicates the bioinformatics resources (tools and databases) used for this analysis.

Acronym	Description	Version/release	URL
ReMap	Database of transcriptional regulators peaks derived from curated ChIP-seq, ChIP-exo, DAP-seq experiments in Human, Mouse, Fruit Fly and Arabidopsis Thaliana.	2022, 4rth human release	https://remap.univ-amu.fr/
JASPAR	Database of transcription factor binding profiles stored as position frequency matrices (PFMs) and TF flexible models (TFFMs) for TFs across multiple species in six taxonomic groups.	9th release (2022)	https://jaspar.genereg.net/
RSAT	Regulatory Sequence Analysis Tools	Feb 9 21:00:40 2022	http://rsat.eu/

Data sources

All the data sources used in this study are listed below.

The transcription factor CTCF (official gene name) is fully named as CCCTC-binding factor and its ChIP-seq data in MCF-7 cell line was obtained from **ReMap2022 human release**. The original experiment comes from **GEO dataset GSE124667**. The human genome version used to make these experiments was *hg38* and the biotype modification of the cell line is FOXA1.

The CTCF reference motif was extracted from **JASPAR 9th release** database. This motif belongs to *Homo sapiens* species, resulted from ChIP-seq data and its matrix ID is MA0139.1 (Base ID is M10139 and version 1). Data validation could be found with the PMID **17512414**.

Commands and parameters

The next commands were executed in Linux command line and RSAT server to get the results presented in this work.

```
## Decompress input ReMap data
gunzip GSE124667.CTCF.MCF-7_FOXA1.bed.gz

## Number of peaks
wc -l GSE124667.CTCF.MCF-7_FOXA1.bed
3781 GSE124667.CTCF.MCF-7_FOXA1.bed

## Sum of peak lengths
```

```
 awk '{ total += $3 -$2 } END { print total }' GSE124667.CTCF.MCF-7_FOXA1.bed
622361


# Mean peak lengths
 awk '{ total += $3 -$2; count++ } END { print total/count }' GSE124667.CTCF.MCF-7_
 FOXA1.bed
164.602

## Median peak lengths
awk '{print $3-$2}' GSE124667.CTCF.MCF-7_FOXA1.bed | sort | awk 'NR == 1891 {print $1}'
154

## Peak-motifs analysis to discover motifs and compare them
$RSAT/perl-scripts/peak-motifs  -v 1 -title 'GSE124667.CTCF.MCF-7_FOXA1_sajr_20220213/
_074429'  -i $RSAT/public_html/tmp/www-data/2022/02/13/peak-motifs.2022-02-13.184552_/
2022-02-13.184552_r240Dt/peak-motifspeak_seq  -max_seq_len 1000 -markov auto -disco/
oligos,dyads,positions -nmotifs 5  -minol 6 -maxol 7  -no_merge_lengths -2str  -origin/
center  -motif_db jaspar_core_nonredundant_vertebrates tf $RSAT/public_html/motif_data/
bases/JASPAR/Jaspar_2020/nonredundant/JASPAR2020_CORE_vertebrates_non-redundant_pfms.tf/
-motif_db personnal_collection tf $RSAT/public_html/tmp/www-data/2022/02/13/peak-motif/
s.2022-02-13.184552_2022-02-13.184552_r240Dt/peak-motifs_custom_motif_db.tf -scan_marko/
v 1 -task purge,seqlen,composition,disco,merge_motifs,split_motifs,motifs_vs_motifs,tim/
elog,archive,synthesis,small_summary,motifs_vs_db,motifs_vs_db,scan -prefix peak-motifs/
-noov -img_format png  -outdir $RSAT/public_html/tmp/www-data/2022/02/13/peak-motifs.20/
22-02-13.184552_2022-02-13.184552_r240Dt

## Cluster of the reference motif and discovered motifs by similitude
$RSAT/perl-scripts/matrix-clustering  -v 1 -max_matrices 300 -matrix peak-motifs_/
results $RSAT/public_html/tmp/www-data/2022/02/17/matrix-clustering_2022-02-17.06/
4113_EH7tUz/matrix-clustering_query_matrices.transfac transfac -matrix ctcf_ref_/
motif $RSAT/public_html/tmp/www-data/2022/02/17/matrix-clustering_2022-02-17.064/
113_EH7tUz/matrix-clustering_second_matrices. transfac -hclust_method average/
-calc sum -title 'ctcf_discovered_motifs_' -metric_build_tree 'Ncor' -lth w 5/
-lth cor 0.6 -lth Ncor 0.4 -quick-label_in_tree name -return json,heatmap  -o/
$RSAT/public_html/tmp/www-data/2022/02/17/matrix-clustering_2022-02-17.064113_/
EH7tUz/matrix-clustering 2> $RSAT/public_html/tmp/www-data/2022/02/17/matrix-/
clustering_2022-02-17.064113_EH7tUz/matrix-clustering_err.txt

## Compare root motifs with JASPAR nonredundant vertebrate database
compare-matrices -v 1 -format1 transfac -file1 $RSAT/public_html/tmp/www-data/
/2022/02/17/compare-matrices_2022-02-17.081823_UDCghk/compare-matrices_query_matrices/
.transfac -file2 $RSAT/public_html/motif_databases/JASPAR/Jaspar_2020/nonredundant/JA/
SPAR2020_CORE_vertebrates_non-redundant_pfms.tf -format2 tf -strand DR -lth cor 0.7/
-lth Ncor 0.4 -uth match_rank50 -return cor,Ncor,logoDP,NsEucl,NSW,match_rank,matrix_i/
d,matrix_name,width,strand,offset,consensus,alignments_1ton -o $RSAT/public_html/tmp/
/www-data/2022/02/17/compare-matrices_2022-02-17.081823_UDCghk/compare-matrices.tab

## Total nb. of matches in all peaks for the root motif
grep 'hg38' peaks_coverage_by_rootmotif.ft | sort | uniq | wc -l
3861
## Nb. of peaks with at least one match for the root motif
grep 'hg38' peaks_coverage_by_rootmotif.ft | cut -f1 |sort | uniq | wc -l
```

```
3373

## Generate random fragments for negative control
$RSAT/perl-scripts/random-genome-fragments  -template_format bed -i $RSAT/public_html/
/tmp/www-data/2022/02/17/random-genome-fragments_2022-02-17.211203_88YdTW_template.bed/
-org Homo_sapiens_GRCh38  -return coord -coord_format bed -v 1  -o $RSAT/public_html//
tmp/www-data/2022/02/17/random-genome-fragments_2022-02-17.211203_88YdT W_fragments./
bed 2> $RSAT/public_html/tmp/www-data/2022/02/17/random-genome-fragments_2022-02-17./
211203_88YdW_error_log.txt

## Fetch random sequences based on the peak sequences; peak analysis
$RSAT/perl-scripts/peak-motifs  -v 1 -title 'random_fragments_hbsv_20220217_211721'/
-i $RSAT/public_html/tmp/www-data/2022/02/17/peak-motifs.2022-02-17.212058_2022-02-/
17.212058_f5jaIE/peak-motifspeak_seq  -max_seq_len 1000 -markov auto -disco oligos,/
dyads,positions -nmotifs 5  -minol 6 -maxol 7  -no_merge_lengths -2str  -origin center/
-motif_db jaspar_core_nonredundant_vertebrates tf $RSAT/public_html/motif_databases//
JASPAR/Jaspar_2020/nonredundant/JASPAR2020_CORE_vertebrates_non-redundant_pfms.tf/
-motif_db personnal_collection tf $RSAT/public_html/tmp/www-data/2022/02/17/peak-motif/
s.2022-02-17.212058_2022-02-17.212058_f5jaIE/peak-motifs_custom_motif_db.tf -scan_mark/
ov 1 -task purge,seqlen,composition,disco,merge_motifs,split_motifs,motifs_vs_motifs,t/
imelog,archive,synthesis,small_summary,motifs_vs_db,motifs_vs_db,scan -prefix peak-mot/
ifs -noov -img_format png  -outdir $RSAT/public_html/tmp/www-data/2022/02/17/peak-moti/
fs.2022-02-17.212058_2022-02-17.212058_f5jaIE
```

References

1. Baniahmad A, Steiner C , Kohne AC, Rernkawitz R . Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. Cell 1990;61:505–14.
2. Burcin M,Arnold R, Lutz M, et al. Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. Mol Cell Biol 1997;17:1281–8.
3. Vostrov AA, Quitschke WW. The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. J Biol Chem 1997;272:33353–9.
4. Shen Y,Yue F, McCleary DF, et al. A map of the cis-regulatory sequences in the mouse genome. Nature 2012;488:116–20
5. Shen Y, et al. 2012. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–12010.1038/nature11243 (doi:10.1038/nature11243)
6. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128, 1231–124510.1016/j.cell.2006.12.048 (doi:10.1016/j.cell.2006.12.048)
7. Holwerda, S. J., & de Laat, W. (2013). CTCF: the protein, the binding partners, the binding sites and their chromatin loops. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 368(1620), 20120369. https://doi.org/10.1098/rstb.2012.0369
8. Wallace JA, Felsenfeld G. 2007. We gather together: insulators and genome organization. Curr. Opin. Genet. Dev. 17, 400–40710.1016/j.gde.2007.08.005 (doi:10.1016/j.gde.2007.08.005)
9. Zlatanova J, Caiafa P. 2009. CTCF and its protein partners: divide and rule? J. Cell Sci., 1275–128410.1242/jcs.039990 (doi:10.1242/jcs.039990)
10. Lee BK, Iyer VR. 2012. Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. J. Biol. Chem. 287, 30 906–30 91310.1074/jbc.R111.324962 (doi:10.1074/jbc.R111.324962)
11. Wendt KS, et al. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature 451, 796–80110.1038/nature06634 (doi:10.1038/nature06634)

12. Rubio ED, Reiss DJ, Welcsh PL, Disteche CM, Filippova GN, Baliga NS, Aebersold R, Ranish JA, Krumm A. 2008. CTCF physically links cohesin to chromatin. Proc. Natl Acad. Sci. USA 105, 8309–831410.1073/pnas.0801273105 (doi:10.1073/pnas.0801273105)
13. Parelho V, et al. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell 132, 422–43310.1016/j.cell.2008.01.011 (doi:10.1016/j.cell.2008.01.011)
14. Stedman W, Kang H, Lin S, Kissil JL, Bartolomei MS, Lieberman PM. 2008. Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. EMBO J. 27, 654–66610.1038/emboj.2008.1 (doi:10.1038/emboj.2008.1)
15. Xiao T, Wallace J, Felsenfeld G. 2011. Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. Mol. Cell. Biol. 31, 2174–218310.1128/MCB.05093-11 (doi:10.1128/MCB.05093-11)
16. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495–501
17. Filippova GN, Qi C-F, Ulmer JE, et al.Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter tts DNA-binding specificity. Cancer Res 2002;62:48–52
18. Cao, K., & Shilatifard, A. (2018). Enhancers in cancer: Genetic and epigenetic deregulation. In Encyclopedia of Cancer (pp. 559-568). Elsevier.
19. Ross-Innes CS, Brown GD, Carroll JS. 2011. A co-ordinated interaction between CTCF and ER in breast cancer cells. BMC Genomics 12, 593.10.1186/1471-2164-12-593 (doi:10.1186/1471-2164-12-593)
20. Walsh CP, Xu GL (2006). "Cytosine methylation and DNA repair". Curr Top Microbiol Immunol. Current Topics in Microbiology and Immunology. 301: 283–315. doi:10.1007/3-540-31390-7_11. ISBN 3-540-29114-8. PMID 16570853.
21. Arnheim N, Calabrese P (2009). "Understanding what determines the frequency and pattern of human germline mutations". Nat Rev Genet. 10 (7): 478–488. doi:10.1038/nrg2529. PMC 2744436. PMID 19488047.
22. Loukinov D. I., Pugacheva E., Vatolin S., et al. BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. Proceedings of the National Academy of Sciences of the United States of America . 2002;99(10):6806–6811. doi: 10.1073/pnas.092123699.
23. D'Arcy V, Pore N, Docquier F, et al: BORIS, a paralogue of the transcription factor, CTCF, is aberrantly expressed in breast tumours. Br J Cancer. 98:571–579. 2008.PubMed/NCBI
24. Loukinov DI, Pugacheva E, Vatolin S, et al: BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. Proc Natl Acad Sci USA. 99:6806–6811. 2002.
25. Christofk HR, Vander Heiden MG, Wu N, Asara JM, Cantley LC: Pyruvate kinase M2 is a phosphotyrosine-binding protein. Nature 2008, 452:181–186.
26. Singh S, Narayanan SP, Biswas K, Gupta A, Ahuja N, Yadav S, Panday RK, Samaiya A, Sharan SK, Shukla S: Intragenic DNA methylation and BORIS-mediated cancer-specific splicing contribute to the Warburg effect. Proc Natl Acad Sci U S A 2017, 114:11440–11445.
27. Clouaire, T., Roussigne, M., Ecochard, V., Mathe, C., Amalric, F., & Girard, J. P. (2005). The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. Proceedings of the National Academy of Sciences, 102(19), 6907-6912.
28. Dong, C., Liu, J., Chen, S.X. et al. Highly robust model of transcription regulator activity predicts breast cancer overall survival. BMC Med Genomics 13, 49 (2020). https://doi.org/10.1186/s12920-020-0688-z
29. Zhang T, Wang H, Saunee NA, Breslin MB, Lan MS. Insulinoma-associated antigen-1 zinc-finger transcription factor promotes pancreatic duct cell trans-differentiation. Endocrinology. 2010 May;151(5):2030-9. doi: 10.1210/en.2009-1224. Epub 2010 Mar 9. PMID: 20215568; PMCID: PMC2869251.
30. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, Guo YE, Hnisz D, Jaenisch R, Bradner JE, Gray NS, Young RA. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell. 2017 Dec 14;171(7):1573-1588.e28. doi: 10.1016/j.cell.2017.11.008. Epub 2017 Dec 7. PMID: 29224777; PMCID: PMC5785279.

31. Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, Lachanski CV, Gillis DR, Phillips-Cremins JE. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. 2017 Jul;27(7):1139-1152. doi: 10.1101/gr.215160.116. Epub 2017 May 23. PMID: 28536180; PMCID: PMC5495066.
32. Wan M, Huang W, Kute TE, Miller LD, Zhang Q, Hatcher H, et al. . Yin Yang 1 plays an essential role in breast cancer and negatively regulates p27. Am J Pathol. (2012) 180:2120–33. 10.1016/j.ajpath.2012.01.037
33. Allouche A, Nolens G, Tancredi A, Delacroix L, Mardaga J, Fridman V, et al. . The combined immunodetection of AP-2alpha and YY1 transcription factors is associated with ERBB2 gene overexpression in primary breast tumors. Breast Cancer Res. (2008) 10:R9. 10.1186/bcr1851
34. Begon DY, Delacroix L, Vernimmen D, Jackers P, Winkler R. Yin Yang 1 cooperates with activator protein 2 to stimulate ERBB2 gene expression in mammary cancer cells. J Biol Chem. (2005) 280:24428–34. 10.1074/jbc.M503790200
35. Thomassen M, Tan Q, Kruse TA. Gene expression meta-analysis identifies metastatic pathways and transcription factors in breast cancer. BMC Cancer. (2008) 8:394. 10.1186/1471-2407-8-394
36. Lee MH, Lahusen T, Wang RH, Xiao C, Xu X, Hwang YS, et al. . Yin Yang 1 positively regulates BRCA1 and inhibits mammary cancer formation. Oncogene. (2012) 31:116–27. 10.1038/onc.2011.217
37. Lieberthal JG, Kaminsky M, Parkhurst CN, Tanese N. The role of YY1 in reduced HP1alpha gene expression in invasive human breast cancer cells. Breast Cancer Res. (2009) 11:R42. 10.1186/bcr2329
38. Vandewalle, C., Van Roy, F., and Berx, G. (2009). The role of the ZEB family of transcription factors in development and disease. Cell Mol. Life Sci. 66, 773–787. doi: 10.1007/s00018-008-8465-8
39. Katsura, A., Tamura, Y., Hokari, S., Harada, M., Morikawa, M., Sakurai, T., Takahashi, K., Mizutani, A., Nishida, J., Yokoyama, Y., Morishita, Y., Murakami, T., Ehata, S., Miyazono, K., & Koinuma, D. (2017). ZEB1-regulated inflammatory phenotype in breast cancer cells. Molecular oncology, 11(9), 1241–1262. https://doi.org/10.1002/1878-0261.12098
40. Fedorova O, Daks A, Parfenyev S, Shuvalov O, Netsvetay S, Vasileva J, Gudovich A, Golotin V, Semenov O, Petukhov A, Baiduik E, Berdigaliyev N, Tulchinsky EM, Barlev NA. Zeb1-mediated autophagy enhances resistance of breast cancer cells to genotoxic drugs. Biochem Biophys Res Commun. 2022 Jan 22;589:29-34. doi: 10.1016/j.bbrc.2021.11.088. Epub 2021 Nov 26. PMID: 34883287.
41. Fabre, P.J., Leleu, M., Mormann, B.H. et al. Large scale genomic reorganization of topological domains at the HoxD locus. Genome Biol 18, 149 (2017). https://doi.org/10.1186/s13059-017-1278-z
42. Specificity Protein Transcription Factors and Cancer: Opportunities for Drug Development Stephen Safe1 , James Abbruzzese2 , Maen Abdelrahim3 , and Erik Hedrick1