Comprehensive Identification and Annotation of Cell Type-Specific and Ubiquitous CTCF-Binding Sites in the Human Genome

Hebing Chen, Yao Tian, Wenjie Shu, Xiachen Bo, Shengqi Wang

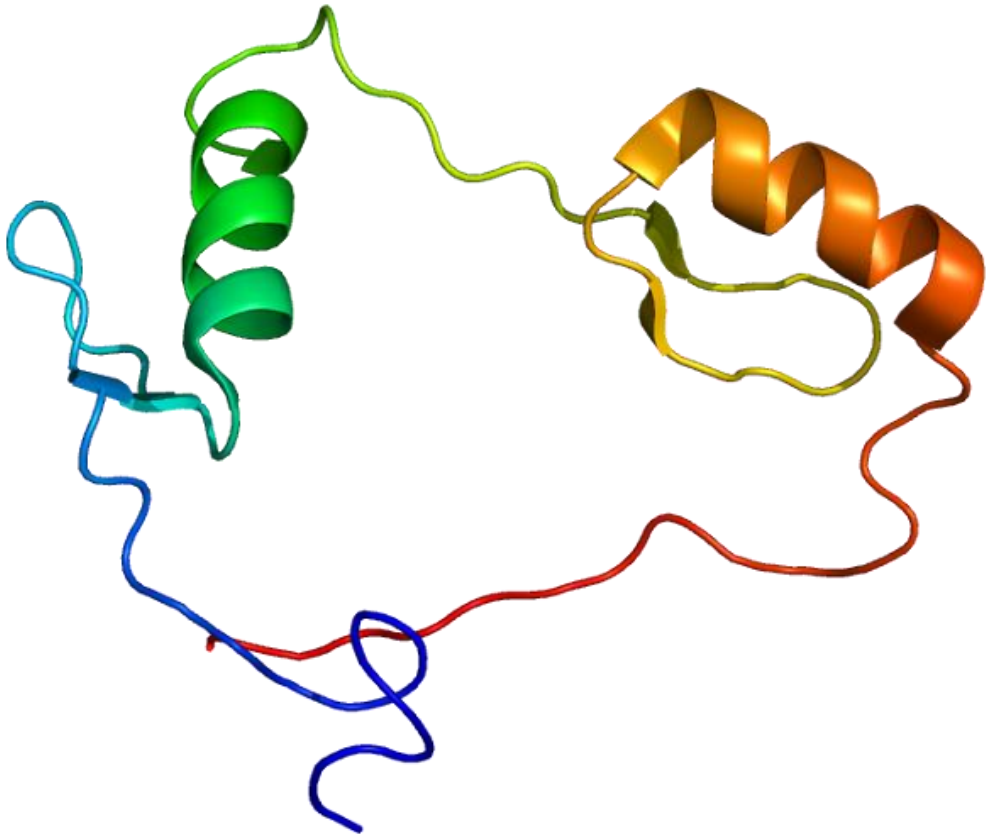
Nithya Sreenivasan

OBJECTIVE

- The paper aims to identify and characterize cell type-specific and ubiquitous CTCF-binding sites in the human genome across 38 cell types designated by the Encyclopedia of DNA Elements (ENCODE) consortium.

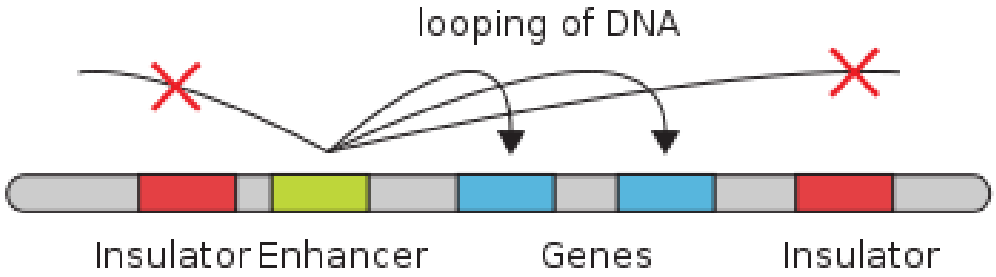
CTCF

- CTCF stands for **CCCTC-binding factor**
- It is a **11-zinc finger DNA binding protein** and is encoded by the **CTCF gene**
- Functions:
  - **Regulates 3D structure of chromatin**
  - Anchors DNA to cellular structures such as nuclear lamina
  - Plays a major roll in transcriptional regulation
- CTCF **helps in chromatin insulator activity** due to the regulation of 3D structure.



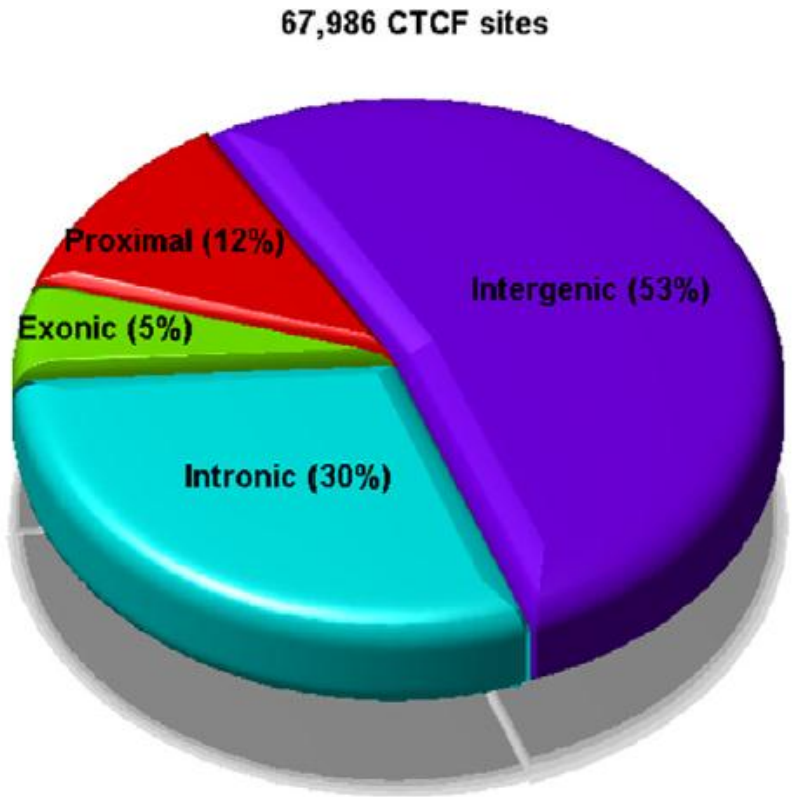
CHROMATIN INSULATOR

- It is a **DNA element** that **regulates the level of gene expression**
- Insulators **block the interaction between enhancers and promoters**
- This is essential when **adjacent genes have different transcription patterns**
- In vertebrates, the **establishment of insulators is achieved only with the help of CTCF**

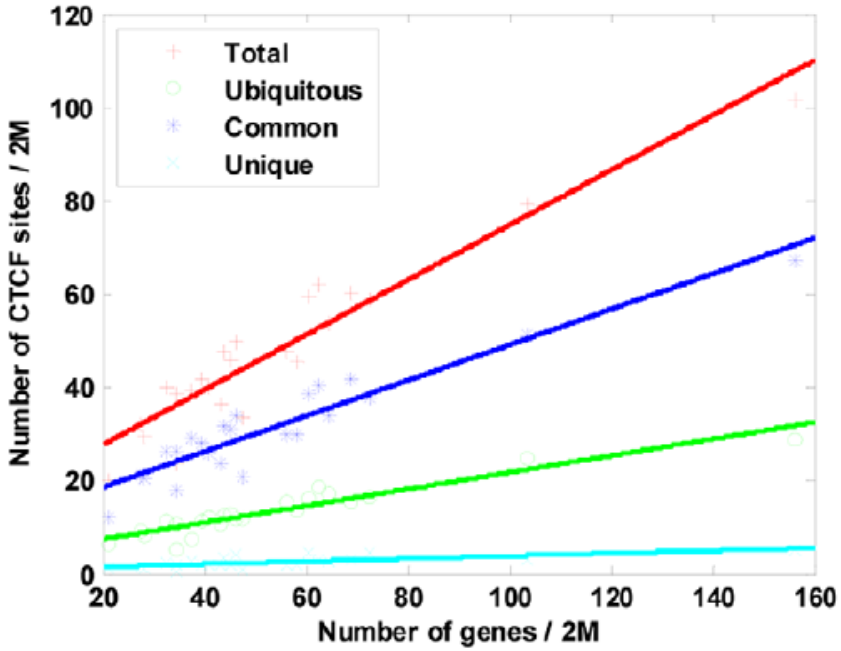


CTCF-BINDING SITES ARE UNIQUELY DISTRIBUTED IN THE HUMAN GENOME

- Ubiquitous and cell-type specific CTCF-binding sites were **universally present** throughout the 38 cell lines
  - >**50%** occurred remotely from TSSs
  - **30 %** of cell-type specific sites were predominantly found in introns and **5%** in exons
  - Ubiquitous binding sites were found in intergenic regions (role of an insulator)



- Chromosomal distribution of CTCF-binding sites **correlated strongly** with genes.
  CTCF-binding sites correlated strongly with enhancers and active promoters – role of enhancer blocking insulator



- CTCF clusters with three or more overlapping members most likely represent real CTCF cluster events
  Clustering contributes to the insulator properties of CTCF-binding sites

MATERIALS & METHODS:

- All sequence and peak data of the 38 cell lines were downloaded from different tracks in the **University of Southern California, Santa Cruz (UCSC) Genome Browser**

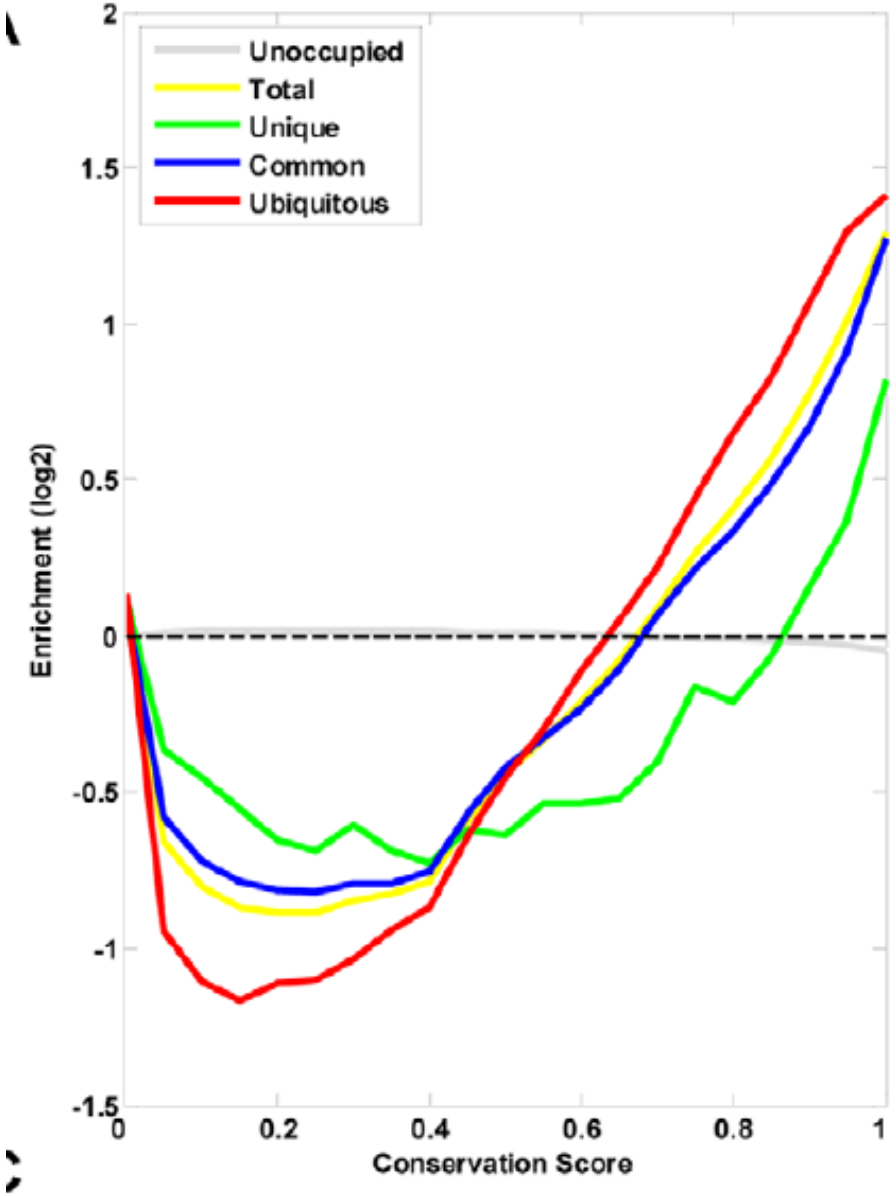
- Classification of CTCF-binding sites:
  All sites were **analyzed for overlaps** and then classified

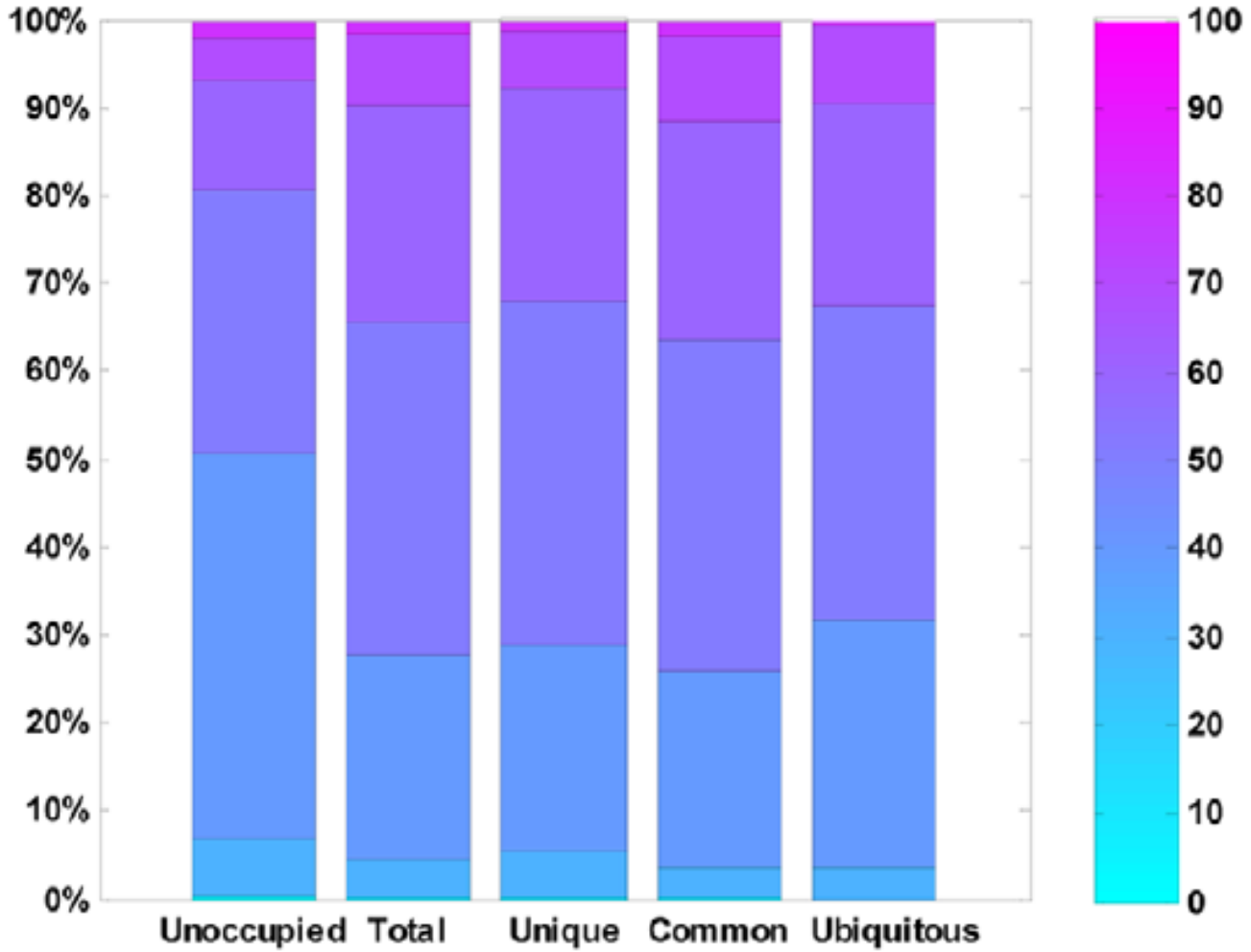
- Densities of gene and CTCF-binding sites:
  Linear regression

- Identification and characterization of CTCF clusters:
  Monte Carlo simulation

CTCF IS A VERSATILE REGULATOR OF TRANSCRIPTION

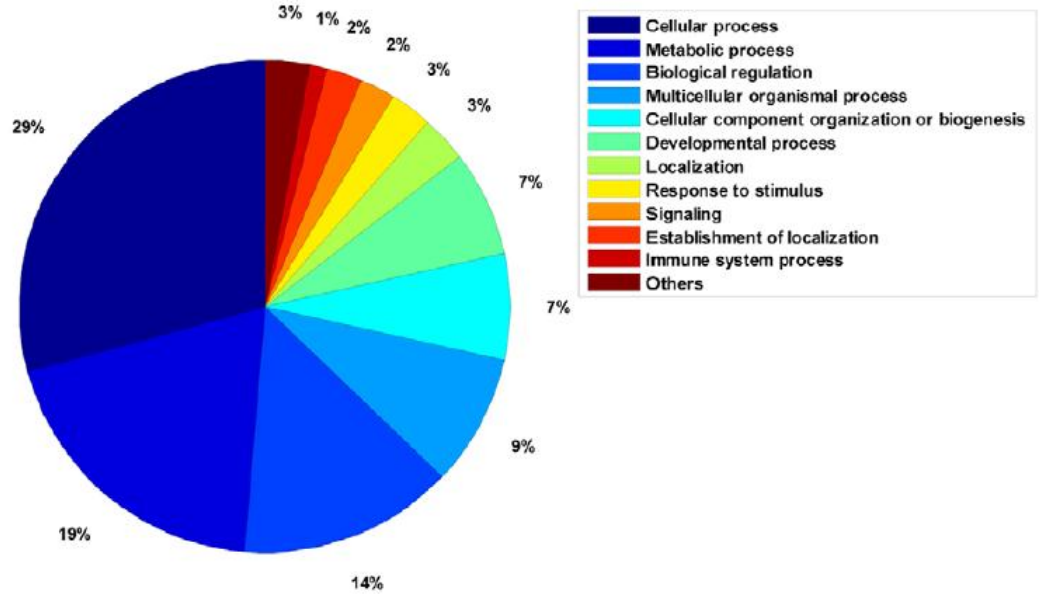
- **Conservation scores** and **GC content** were **significantly higher in ubiquitous CTCF-binding sites** than in cell type-specific sites
  - **High GC content is typically associated with gene-rich** areas and has functional relevance
  - CTCF is involved in **active gene regulation**

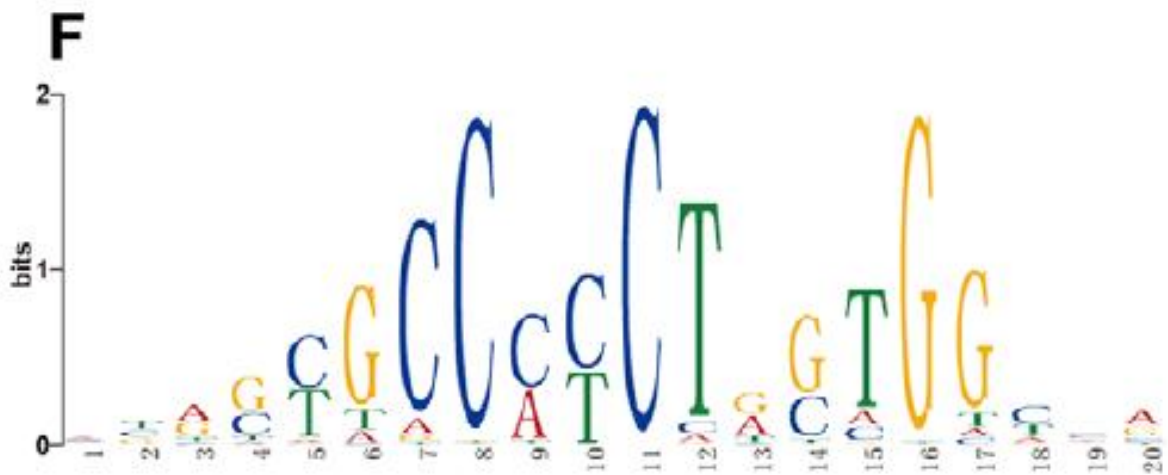

Unoccupied
Total
Unique
Common
Ubiquitous
Enrichment (log2)
Conservation Score
2
1.5
1
0.5
0
-0.5
-1
-1.5
0
0.2
0.4
0.6
0.8
1


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Unoccupied
Total
Unique
Common
Ubiquitous
100
90
80
70
60
50
40
30
20
10
0

- Biological processes such as **cellular processes, metabolic processes, and biologic regulation, may be regulated by CTCF**

  CTCF is a **multi-functional protein** involved in **gene regulation**





- **90% of ubiquitous CTCF-binding sites** were characterized by a specific **20-mer consensus motif** while only **27% cell-type specific CTCF-binding sites** shared this consensus motif

  Finding suggests that CTCF is an **evolutionary conserved**, yet versatile **transcriptional regulator**

MATERIALS & METHODS:

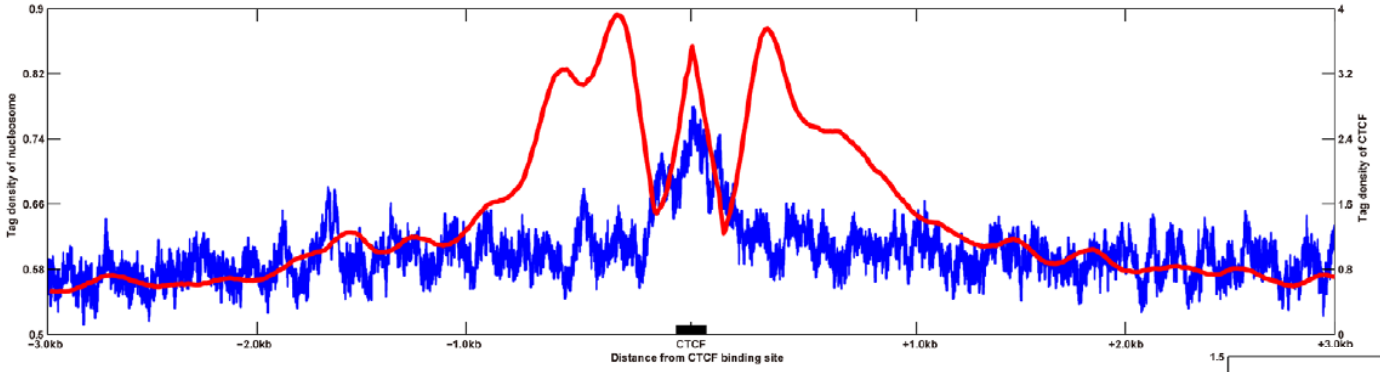
- Conservation analysis of CTCF-binding sites:
  - **phastCons** and **phyloP** scores were collected and **average conservation scores** were calculated for multiple alignments
  - **Wilcoxon rank sum test** to calculate conservation scores
- Analysis of CTCF-binding site GC content:

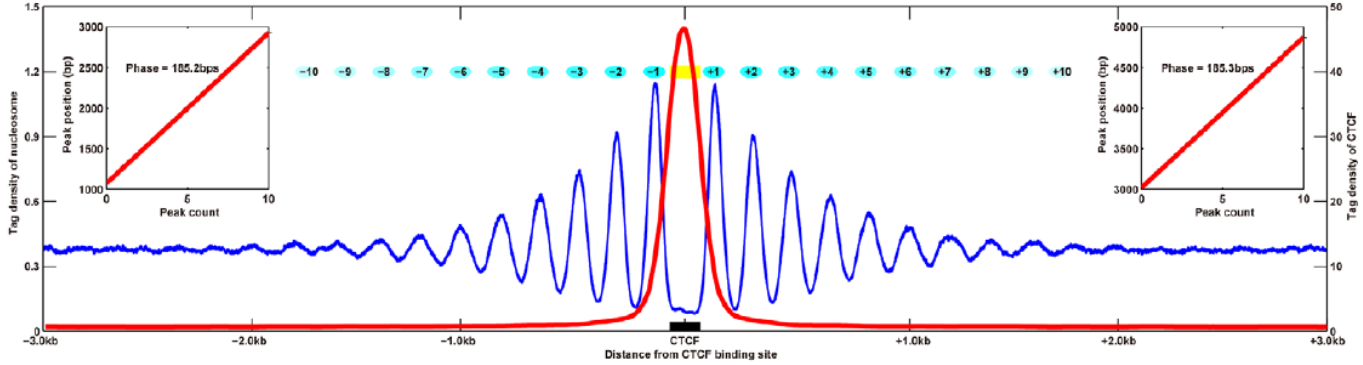
  Sequence data from UCSC Genome browser subjected to **two-sided Wilcoxon rank sum test** to calculate GC content scores
- Gene ontology analysis:
  - **EASE** was used to identify enriched GO categories.
  - Ultimately, the many functions of CTCF were identified
- Analysis of CTCF-binding site motifs:
  - **MEME** – discover consensus motifs
  - **GOMO** – to detect the association between TF-binding motifs and GO terms

CHROMATIN SIGNATURES DETERMINE CELL TYPE-SPECIFIC GENE EXPRESSION

- **Chromatin architecture at cell type-specific CTCF-binding sites is also cell type-specific.**

  20 well positioned nucleosomes flanked the ubiquitous CTCF-binding sites but no nucleosome flanked the cell type-specific sites





- A large number of **open chromatin** and **histone modifications** associated with active transcription, were observed within cell-type specific and ubiquitous binding sites

- Cell-type specific binding sights were **highly methylated** as well as compared to ubiquitous binding sites

- Dnase I HS sites **colocalize** with histone modified regions, p300 binding regions, and gene expression in a **cell-type specific manner**.

MATERIALS & **METHODS**:

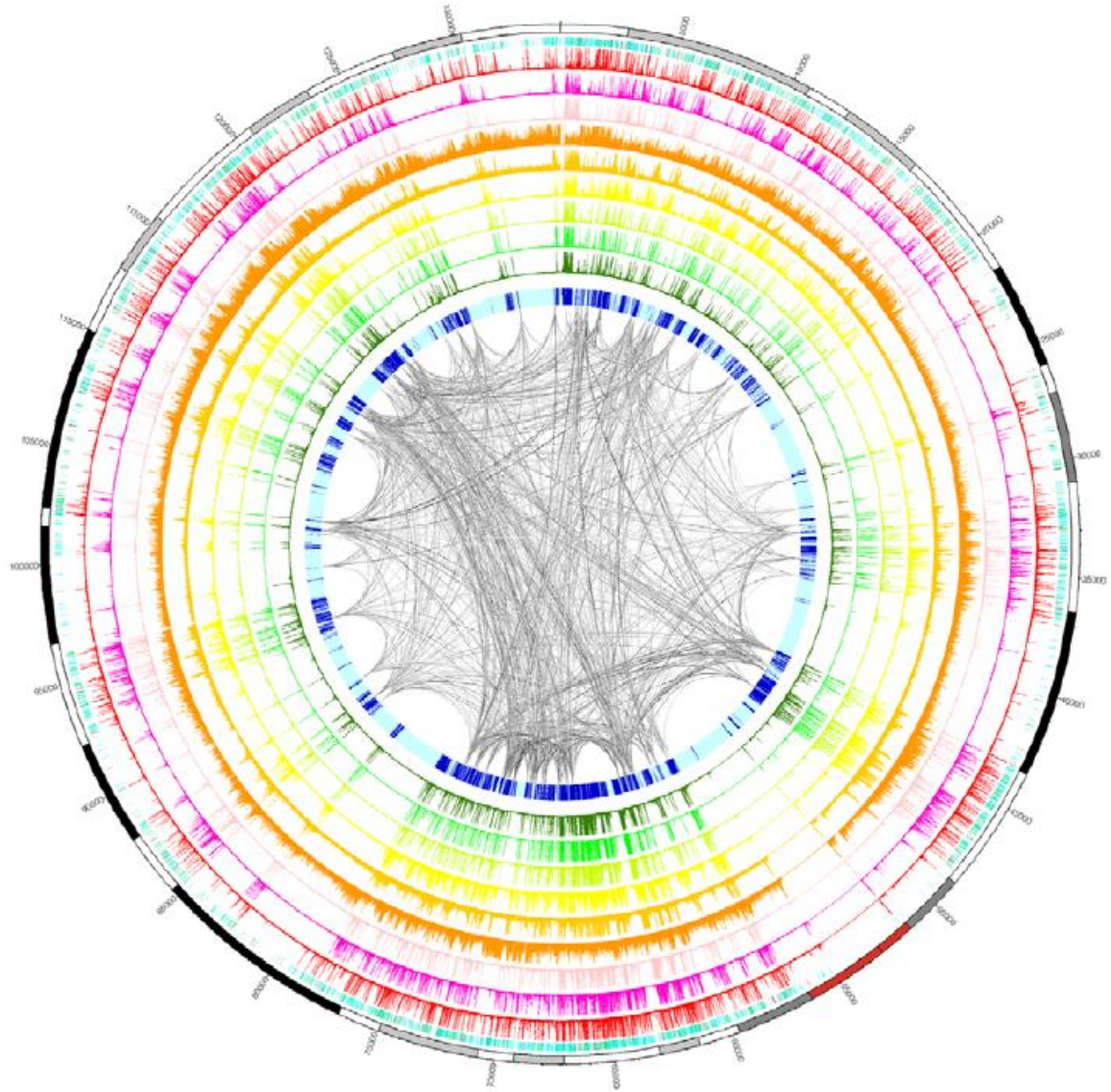
- Scatter correlations of CTCF-binding sites with genome annotations:
  - **Average RPKM value** of CTCF-binding and of genome annotations: nucleosome positioning, histone modifications, and open chromatin were calculated.
  - Evaluated by **Pearson's correlation**
- Quantification of tag densities:

  Done with **RPKM values for the calculation of tag density** for DNase I HS, histone modifications, and nucleosome positioning.
- DNA methylation of CTCF-binding sites:

  Reduced Representation Bisulfite Sequencing (RRBS)
- Colocalization analysis:
  - **Chi-square test** – ubiquitous and common sites were omitted due to many overlaps
  - Carried out on two types of binding sites or regions

CTCF ORGANIZES HIGHER-ORDER CHROMATIN STRUCTURE

- CTCF is known to **demarcate boundaries between euchromatin and heterochromatin**



- CTCF functions as a **domain barrier** that separates the heterochromatin and euchromatin domains
  - Coincides with the fact that **insulators delimit organizational domains of a genome**
  - Barrier function identified due to **DNase I HS** and **specific histone-modification binding sites** located between domains

- **Intrachromosomal interactions** between barrier CTCF-binding sites of chromatin are **much stronger** than all interactions across the human genome

  Suggests that CTCF-mediated chromatin interactions **may produce loops** that behave as functional and structural barriers.

MATERIALS & METHODS:

- Identification of chromatin domains:
  - Genome-wide data sets made with **ChIP-Seq** and **DNaseI-Seq**
  - **HMMSeg** - used to identify heterochromatin and euchromatin domains

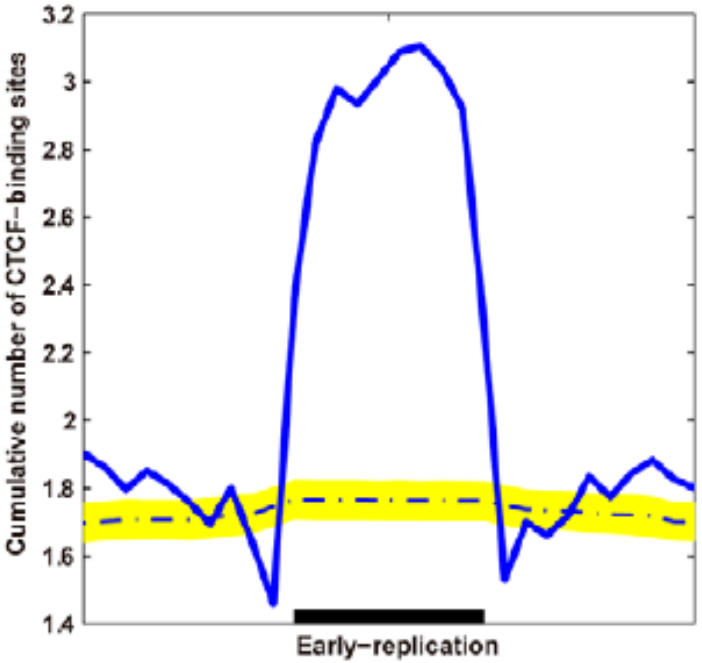
- Identification of barrier CTCF-binding site:

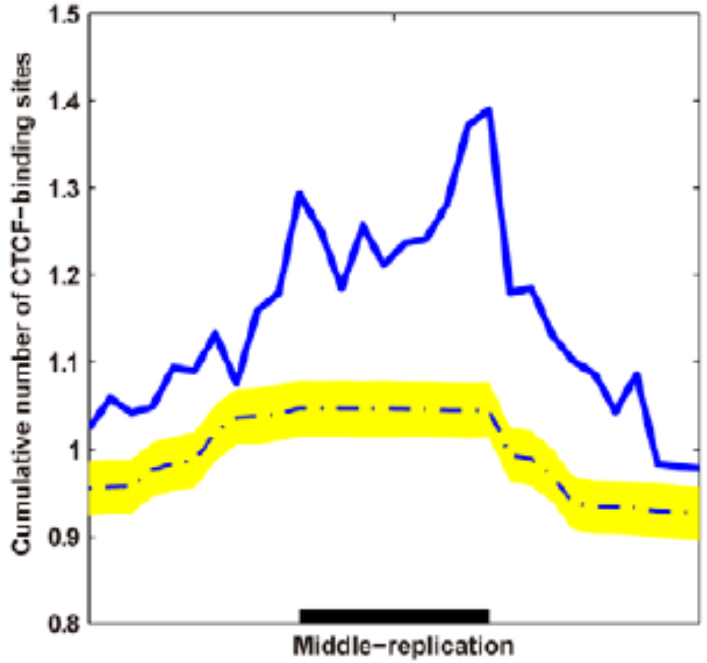
  Calculated the **p-value** to get the significance

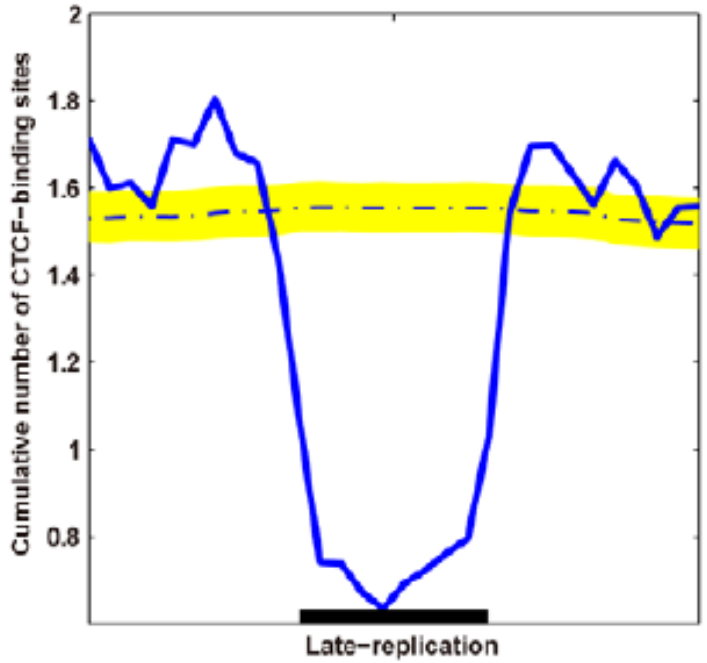
- Intrachromosomal interactions between barrier CTCF-binding sites:
  - **liftOver**
  - **Hi-C method**

CTCF IS INVOLVED IN DNA REPLICATION

- CTCF is closely associated with DNA replication especially during **early- and middle-replication**







- CTCF-binding sites **enriched with replication zones** are **highly cell type-specific**

MATERIALS & METHODS:

- Analysis of replication timing data:
  - **Raw sequencing reads** for replication timing data obtained from **Sequence Read Archive (SRA)**
  - Aligned with human reference genome by using **Bowtie**
  - Tag density of **BrdU-DNA-derived sequence tags** along the genome were calculated for each **cell-cycle fraction**
  - **Average RPKM value** was used as input for **HMMSeg** to identify **replication time domains**

CONCLUSION

- “In conclusion, we have provided a comprehensive and systematic study revealing new functions of cell type-specific and ubiquitous CTCF-binding in the human genome”

- This gives way for further study on the importance of CTCF in many functions such as,
  - Chromatin insulation
  - Gene regulation
  - DNA replication
  - Higher-order chromatin organization

RESEARCH TODAY

Minor mutations in the CTCF zinc finger can lead to tumorigenesis

This affects protein binding, gene regulation, and 3D genome folding

-“Crossed wires: 3D genome misfolding in human disease”, Heidi K. Norton and Jennifer E. Phillips-Cremins, J Cell Biol Aug 2017

REFERENCES

- “Comprehensive Identification and Annotation of Cell Type-Specific and Ubiquitous CTCF-Binding Sites in the Human Genome”

Hebing Chen, Yao Tian, Wenjie Shu, Xiachen Bo, Shengqi Wang. PLoS One, July 2012, Volume 7 , Issue 7

- wikipedia.org