Consequences of CTCF H284 mutation – a motif binding analysis using ChIP-Seq data

Kaiqiong Zhao

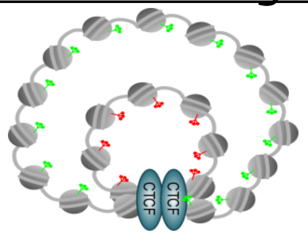
Department of Epidemiology, Biostatistics and Occupational Health

PIs: Dr. Celia Greenwood, Dr. Michael Witcher

Joint work with Benjamin Lebeau and Maïka Jangal

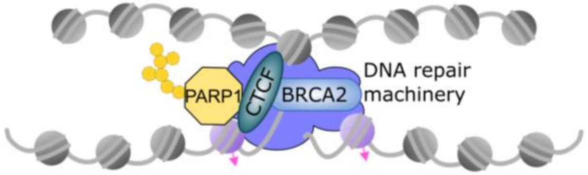
CTCF is a multifunctional epigenetic regulatory protein

Genome organization



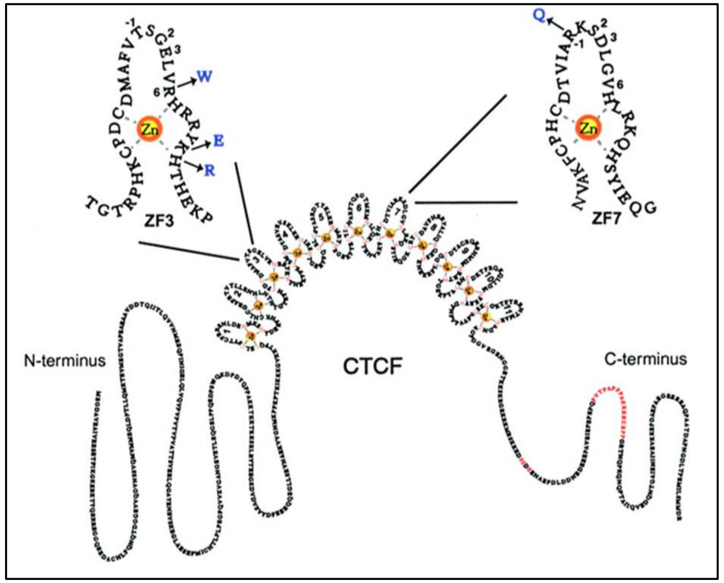
Ong, Chin-Tong and Victor G Corces. (2014) *Nature reviews: Genetics*

Double-Strand Break repair



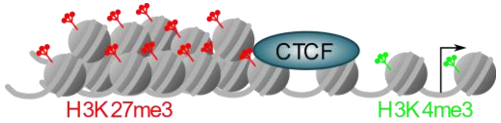
Hilmi, K. et al. (2017) *Science Advances*

CCCTC binding Factor (CTCF)



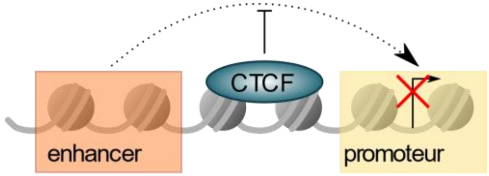
Filippova et al. (2002) *Cancer Res.*

Chromatin boundaries



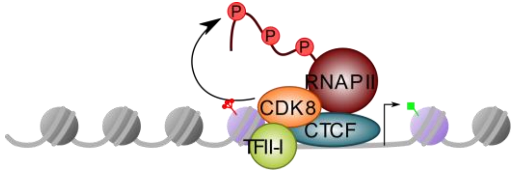
Witcher and Emerson (2009) *Molecular Cell*

Enhancer blocker



Hart, AT. et al. (2000) *Nature*

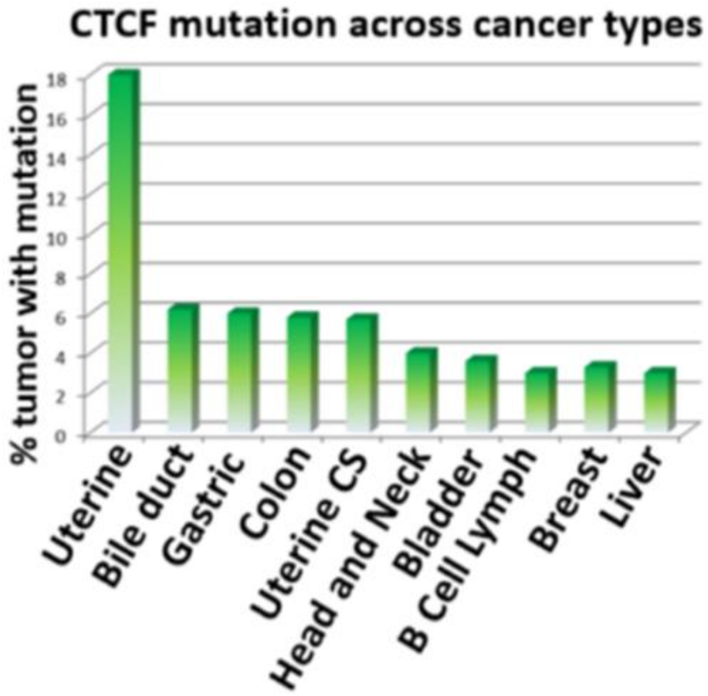
Transcription factor

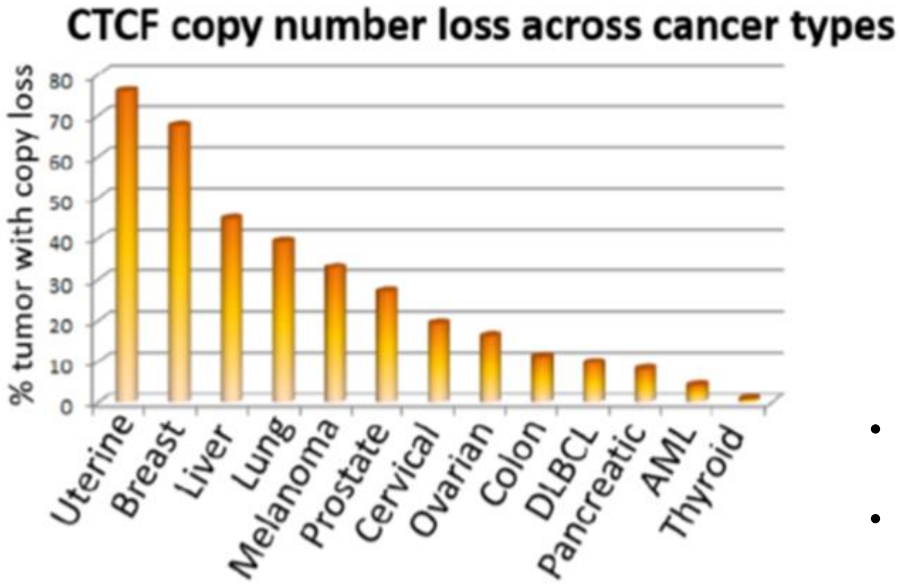


Peña-Hernández, R. et al. (2015) *PNAS*

CTCF alterations in cancer

- CTCF heterozygous mice have increased rate of spontaneous cancer
- In humans, CTCF is found deleted or mutated in a spectrum of tumors

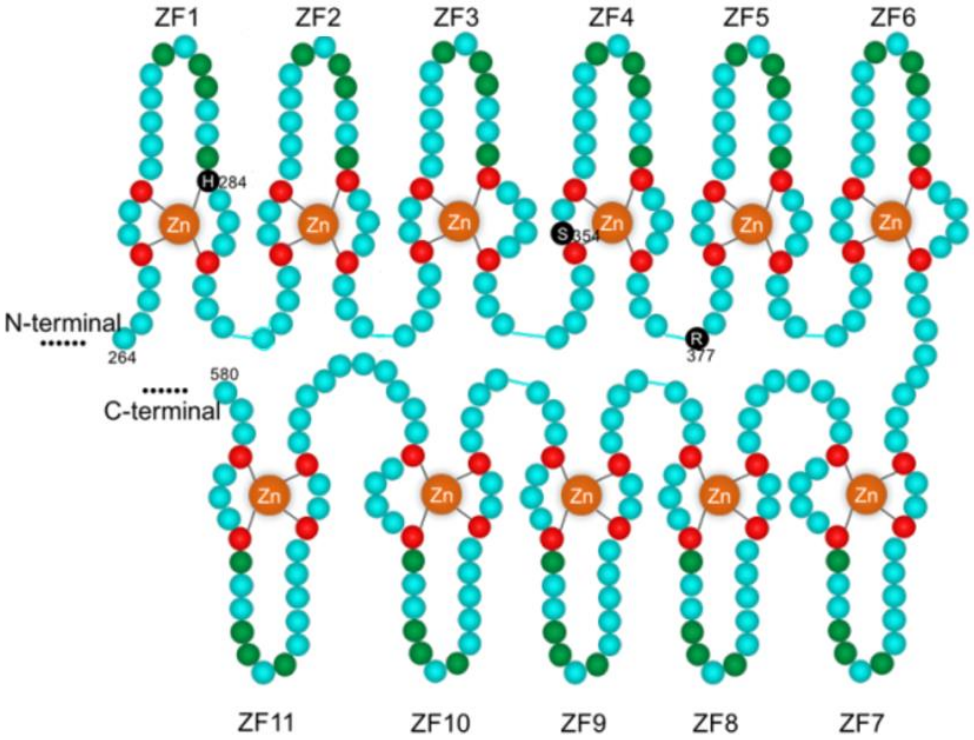

CTCF mutation across cancer types


CTCF copy number loss across cancer types

- Kemp et al. (2014) *Cell Reports*
- Filippova, G.N., et al. (1998) *Genes, chromosomes & cancer*
- Wu, J. et al. (2017) *Oncotarget*

CTCF H284N mutation & breast cancer

- CTCF H284, S354 and R377 are the three most common mutations in cancer
- CTCF H284 mutation is located in the unexplored first zinc-finger of CTCF and is primarily seen in breast cancer
- CTCF mutations are the second most enriched mutations in metastatic vs local breast tumors
- CTCF H284 mutations are found enriched in ER+ tumors resisting hormone therapy

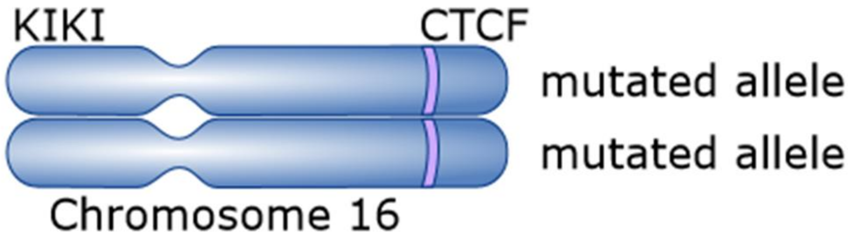


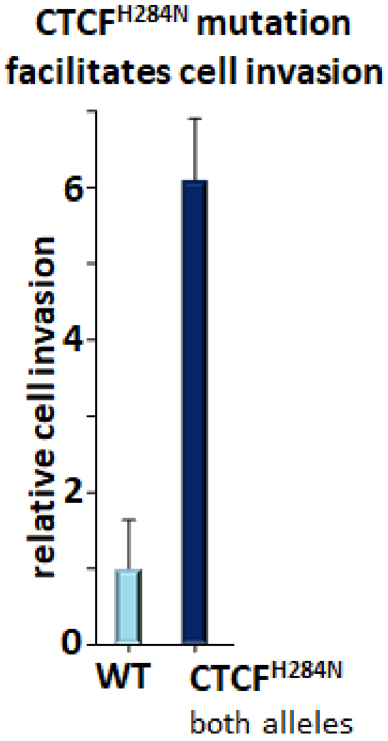
Chenxi Zhang. (2017)

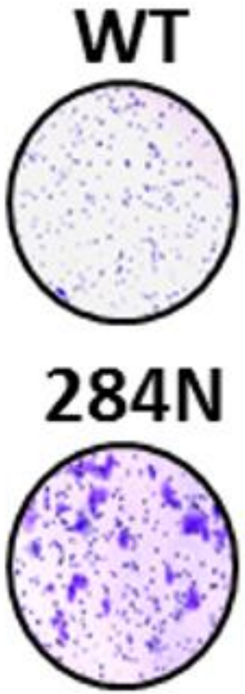
- Razavi et al. (2018) *Cancer Cell*
- Rinaldi et al. (2020) *PLOS One*

Experiment & oncogenic phenotypes

- Introduction of CTCF H284N mutation in both alleles of CTCF in MCF10A (immortalized mammary cells) cell line by CRISPR/Cas9







- ChIP-seq data for three samples:
  - one wild type
  - two mutant cell lines (KIKI)

- Shows a more regressive phenotype
- Mechanism through epigenetic changes? e.g. DNA binding motif changes?

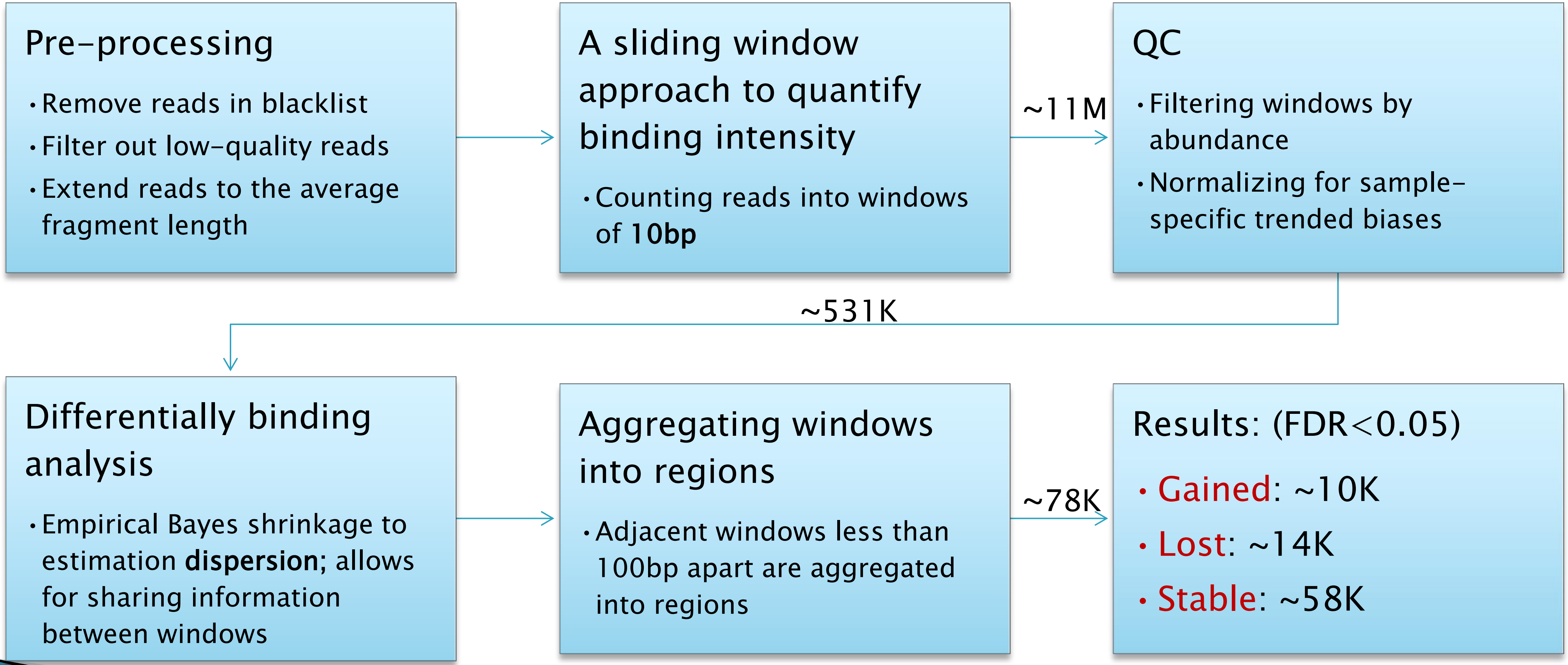
Analysis Goals:

- What is the consequence of CTCF H284N mutation on its binding profile?
  - Locate the gained and lost CTCF binding sites/regions

- How to precisely define the motif consensus underlying those gained and lost sites?
  - What are the common sequence patterns?
  - What are the differences? Is there a small sequence, or single base pair that disrupts or enhances CTCF binding?

Differentially binding peaks

(KIKI vs. WT)



Pre-processing

- Remove reads in blacklist
- Filter out low-quality reads
- Extend reads to the average fragment length

A sliding window approach to quantify binding intensity

- Counting reads into windows of **10bp**

~11M

QC

- Filtering windows by abundance
- Normalizing for sample-specific trended biases

~531K

Differentially binding analysis

- Empirical Bayes shrinkage to estimation **dispersion**; allows for sharing information between windows

Aggregating windows into regions

- Adjacent windows less than 100bp apart are aggregated into regions

~78K

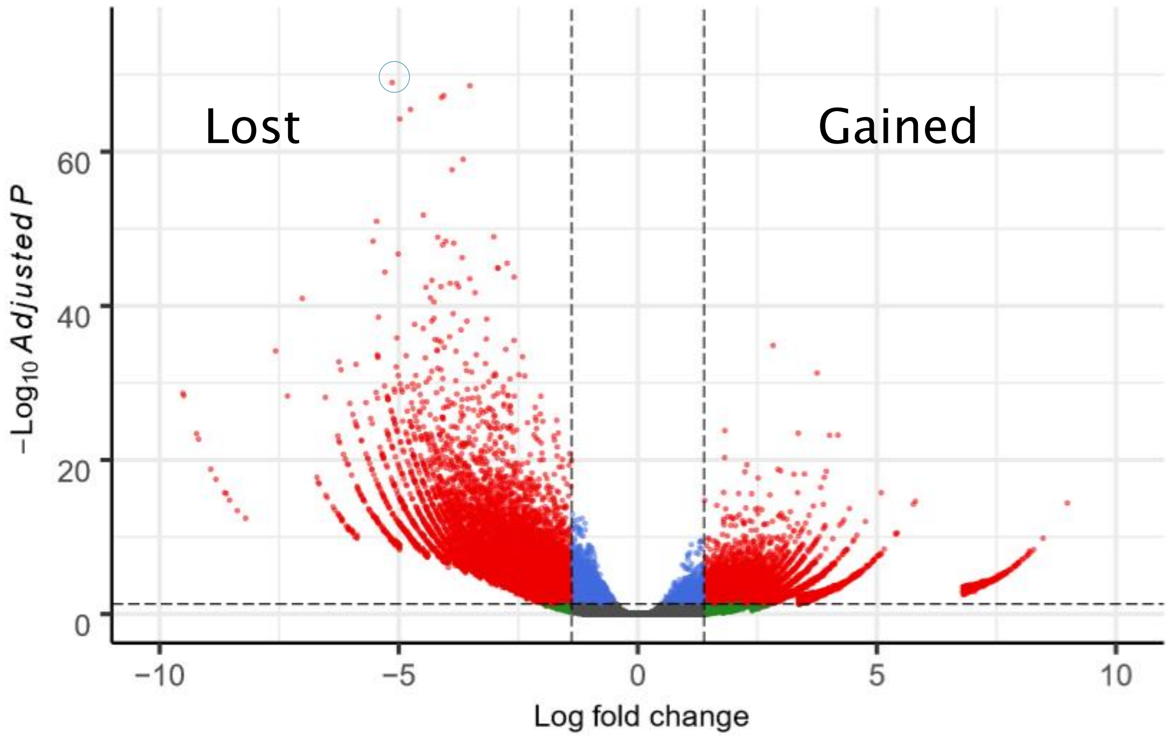
Results: (FDR<0.05)

- Gained: ~10K
- Lost: ~14K
- Stable: ~58K

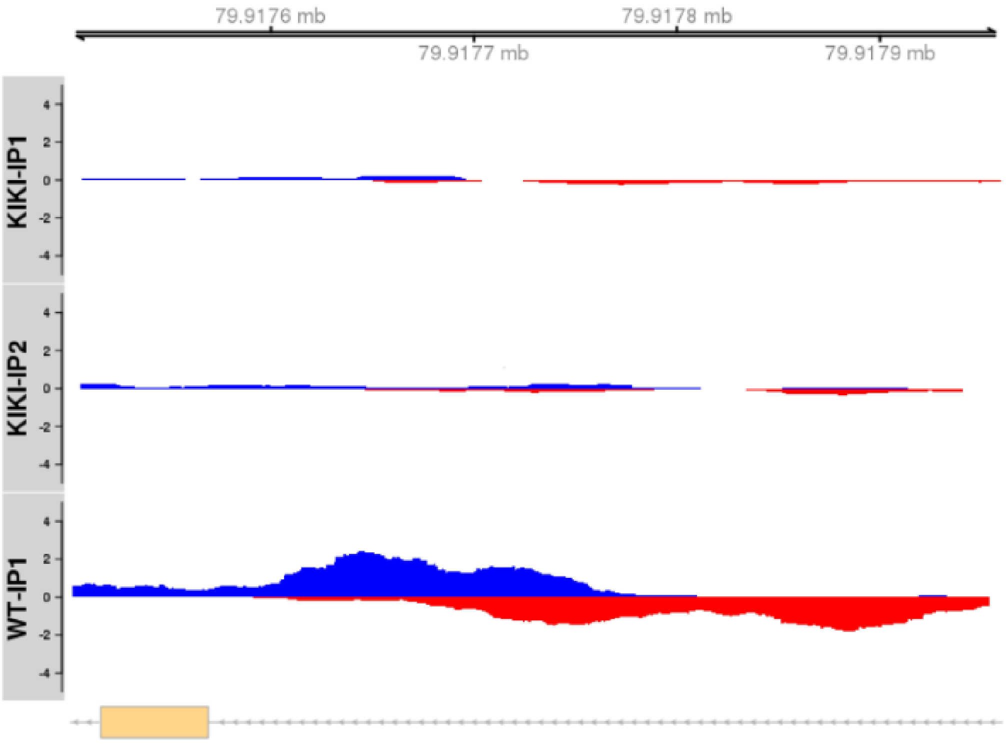
R package ‘csaw’

Differentially binding peaks

(KIKI vs. WT)



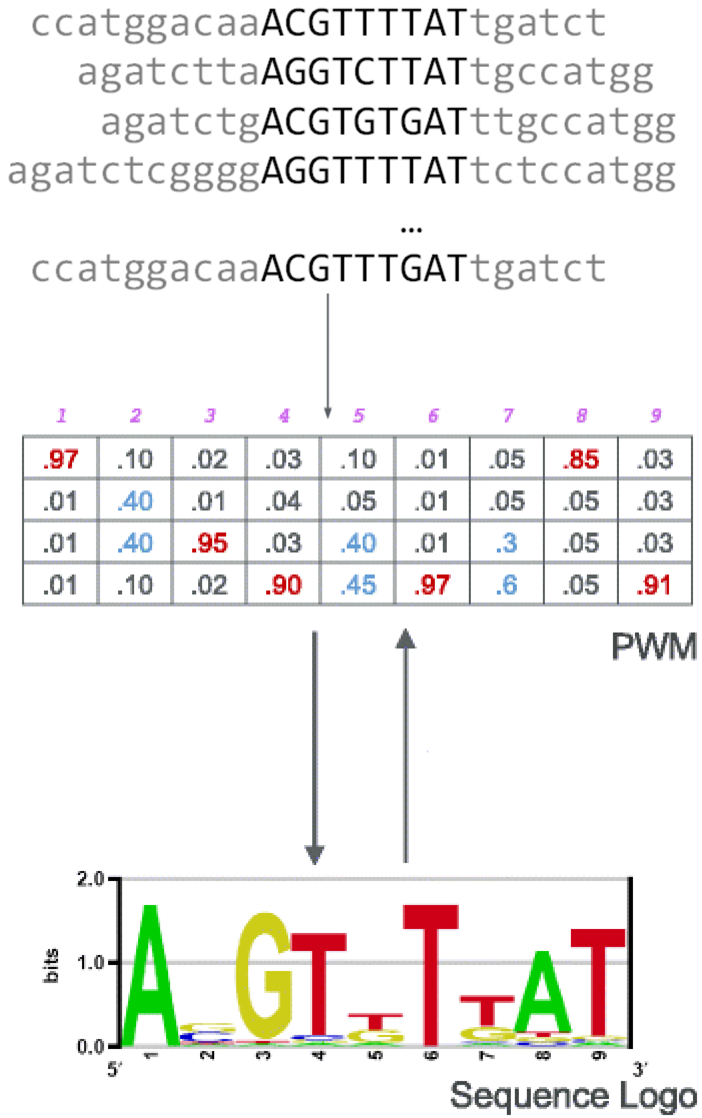
One mutation-induced lost binding peak:



NOTUM

Motif model learning

- Given:
  - a set of sequences of varying length (10-4000bp with mean 300bp) from the Gained, Lost or Stable cluster.
- Tasks:
  - Infer a model for the motif in each cluster
  - Identify motif patterns unique to individual clusters
    - Could be a small sequence or single base pair within a canonical motif model



ccatggacaaACGTTTATtgatct
agatcttaAGGTCTTATtgccatgg
agatctgACGTGTGATttgccatgg
agatctcggggAGGTTTATtctccatgg
…
ccatggacaaACGTTTGATtgatct

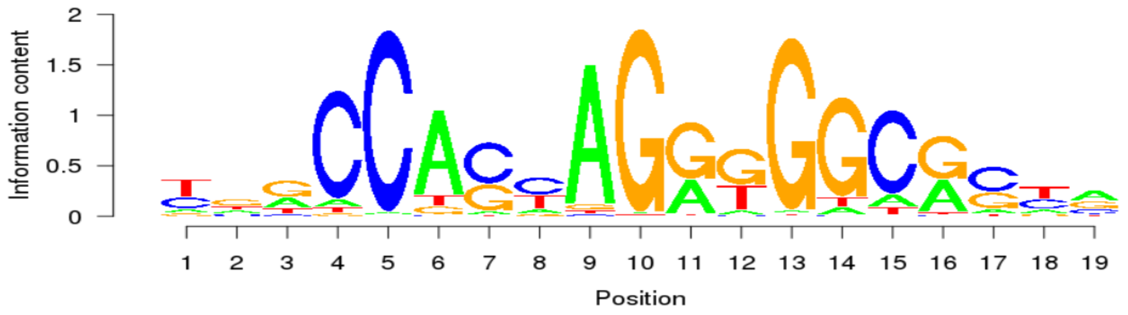
	1	2	3	4	5	6	7	8	9
A	.97	.10	.02	.03	.10	.01	.05	.85	.03
C	.01	.40	.01	.04	.05	.01	.05	.05	.03
G	.01	.40	.95	.03	.40	.01	.3	.05	.03
T	.01	.10	.02	.90	.45	.97	.6	.05	.91

PWM

Sequence Logo

Gao (2017) *BMC Genomics*

Challenges



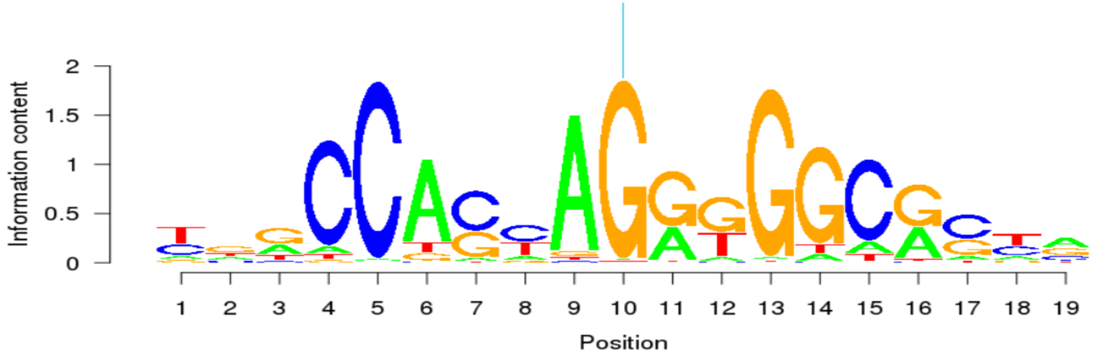
- CTCF binding sites are large, and highly variable in nature

- Identifications of subtle differences requires aligning input sequences to the canonical CTCF model with allowances for mismatches

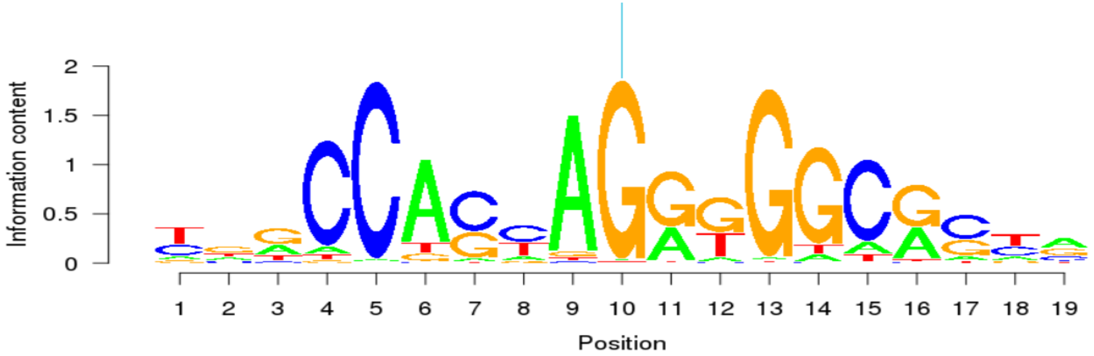
- Existing software packages, e.g. “MEME”, “DREME”, “HOMER, “GADEM” and “DeepBind”, lack the capacity to identify small variations in complex motif model
  - report the canonical CTCF binding motif as a perfect consensus for all the 3 clusters

Our solution

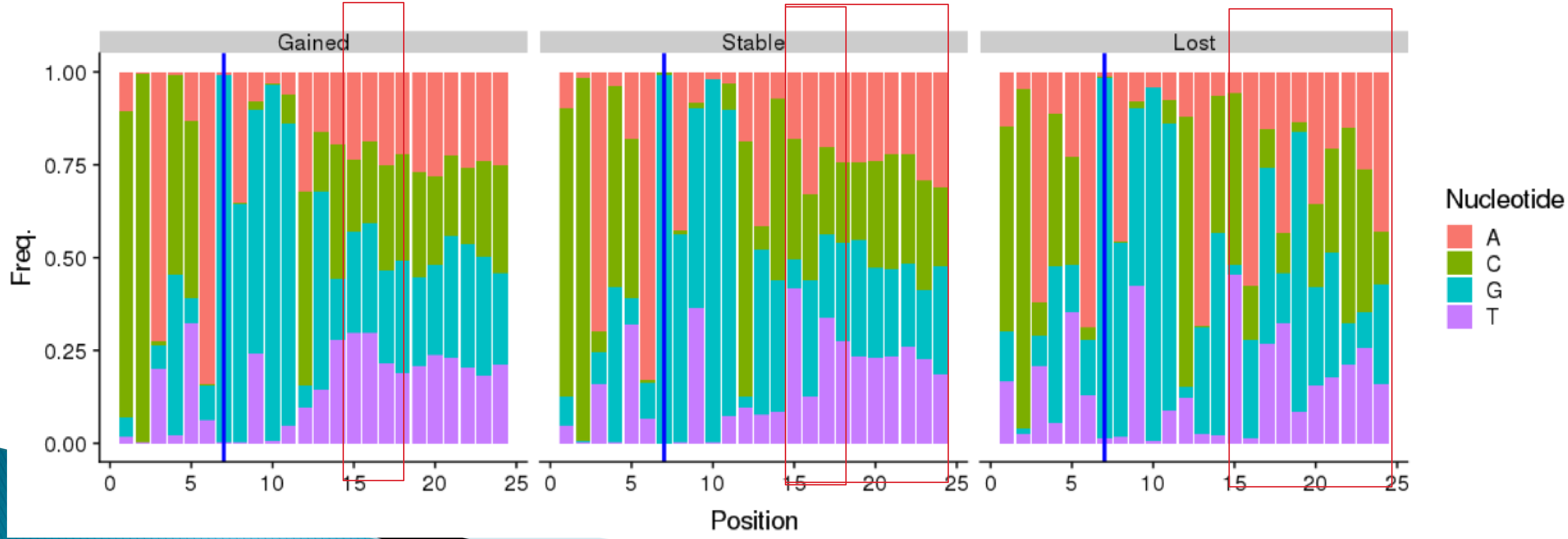
- Identify the locations of the CTCF-like consensus in the given sequences for each cluster
  - ‘GADEM’: word enumeration + EM algorithm for pattern matching
- Align those identified (short) CTCF-like sequences and **extend** on each side by more base pairs
- Compare the nucleotide distributions in the three cluster
  - within a window of different lengths (11, 21, 41 or 61bp) centered at midpoint of the canonical CTCF motif



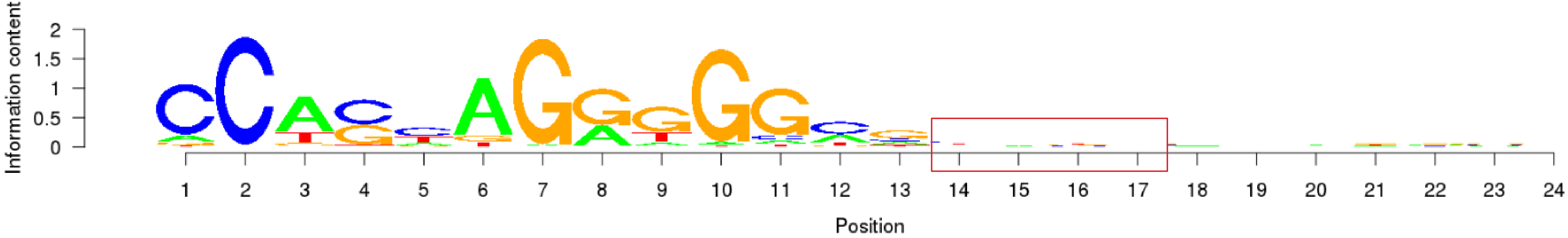
Results



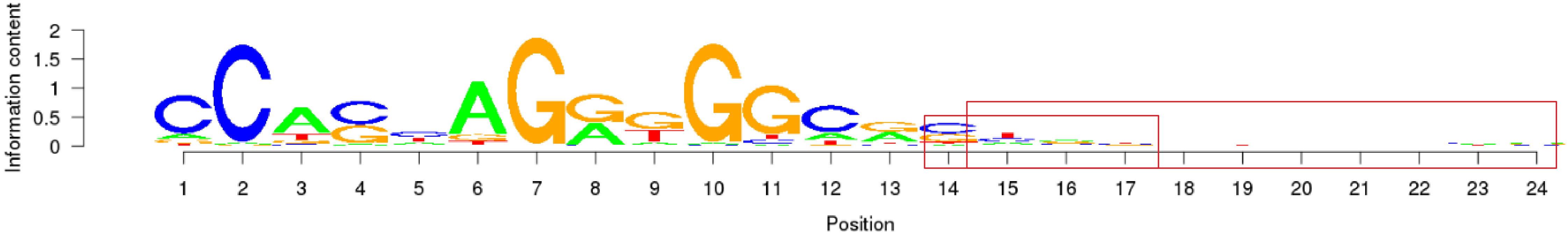
Nucleotide frequencies in a 24 bp window with freq. differences greater than 10%



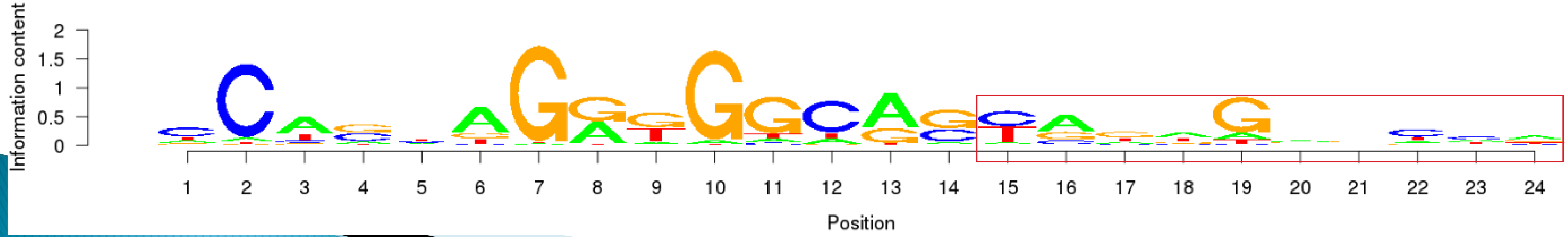
Gained



Stable



Lost



Summary

- We have shown a way to precisely define motif consensuses, which is sensitive to small variation in complex motif model

- Analytical results show that mutant cell lines tend to have less capacity to binding to longer CTCF motifs

Next-step plans

- Use permutation to assess significance

- Build a prediction model using our aligned nucleotide sequences
  - Flexible feature space: single nucleotides, nucleotide pairs or k-mers, at differing distances from the peak centers
  - Models allowing for different ways of interactions

- Investigate sequence-independent factors that could alter CTCF binding to DNA,
  - e.g. DNA methylation, non-coding RNA, or protein cofactors

Thank you!

Questions or Comments

Acknowledgment

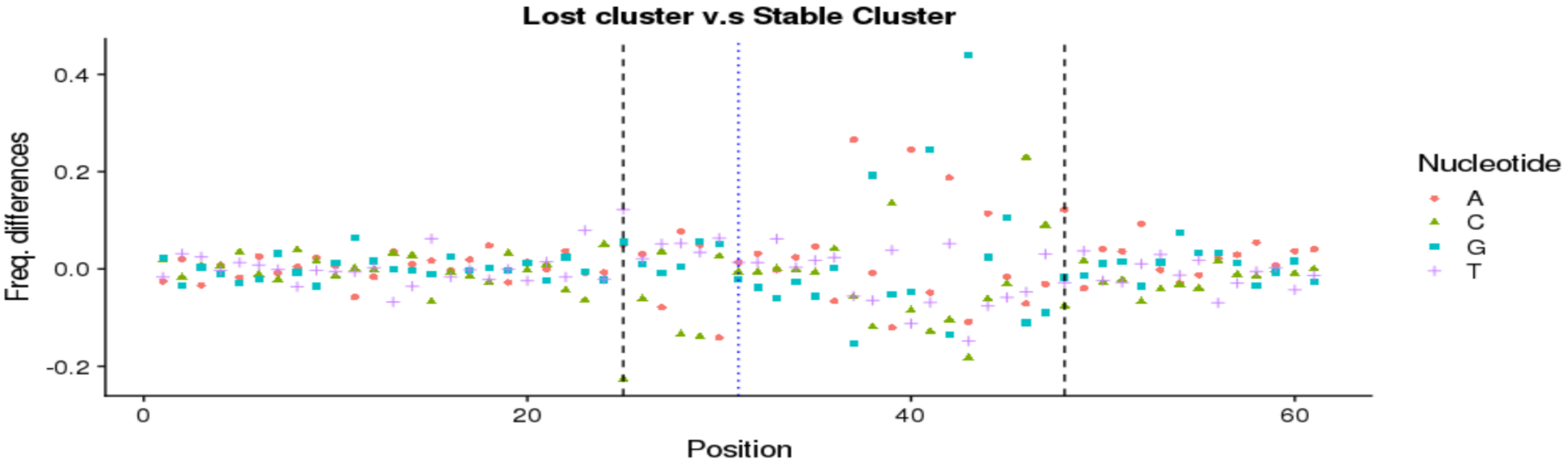
Principle Investigators:

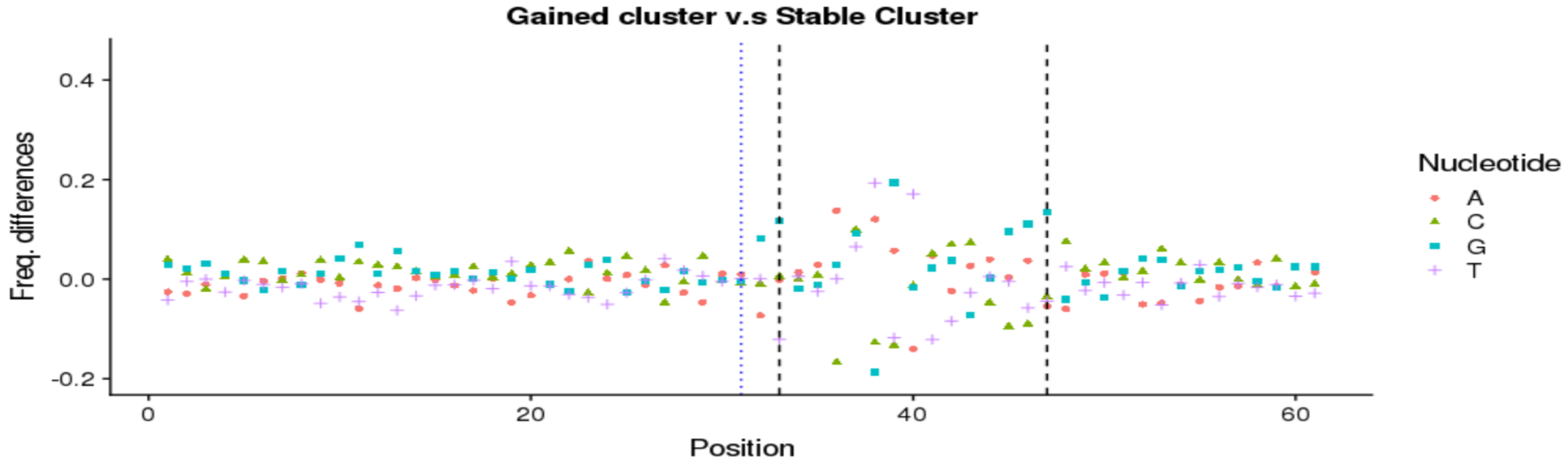
Dr. Michael Witcher
Dr. Celia Greenwood

Dr. Witcher's lab

Maïka Jangal
Benjamin Lebeau

Frequency differences


Lost cluster v.s Stable Cluster


Gained cluster v.s Stable Cluster