Expanding the Encyclopedia of DNA Elements (ENCODE) in the Human and Mouse (UM1)

The Jackson Laboratory

Leading the search for tomorrow's cures

ENCODE in the Human and Mouse (UM1)- Mapping Centers

FOA	Deadlines	Funding Level
RFA-HG-16-002	LOI: Feb 21, 2016 App: Mar 21, 2016	Budget: $2-2.5 Million/YR;4 YRS Total : $15.5-20M for 6-8 awards

NHGRI's highest priorities:

- Maps of transcribed regions
- Maps of chromatin accessibility
- Maps of histone marks
- Maps of other relevant chromatin proteins
- Maps of sites of DNA methylation
- **Maps of long range chromatin interactions**

“These centers should employ high-throughput, genome-wide and cost-effective experimental pipelines for a range of genomic assays capable of generating high quality data to map biochemical activities, exhibited by the human and mouse (10%) genomes, that are associated with functional elements.”

“to encourage highly focused research projects and streamline data management, projects are sought that propose the use of only one biochemical assay (e.g., ChIPseq, RNAseq, and variations thereof). An additional 1-2 assay(s) per application may be considered if they are strongly justified in terms of how centralizing data production within one group,”

“new or improved assays may be applied across a relatively small set of common samples previously used within ENCODE (for which significant amounts of ENCODE data already exist)”

Scientific questions (theme):

Structure codes for chromatin topology
and their functions in human genomes

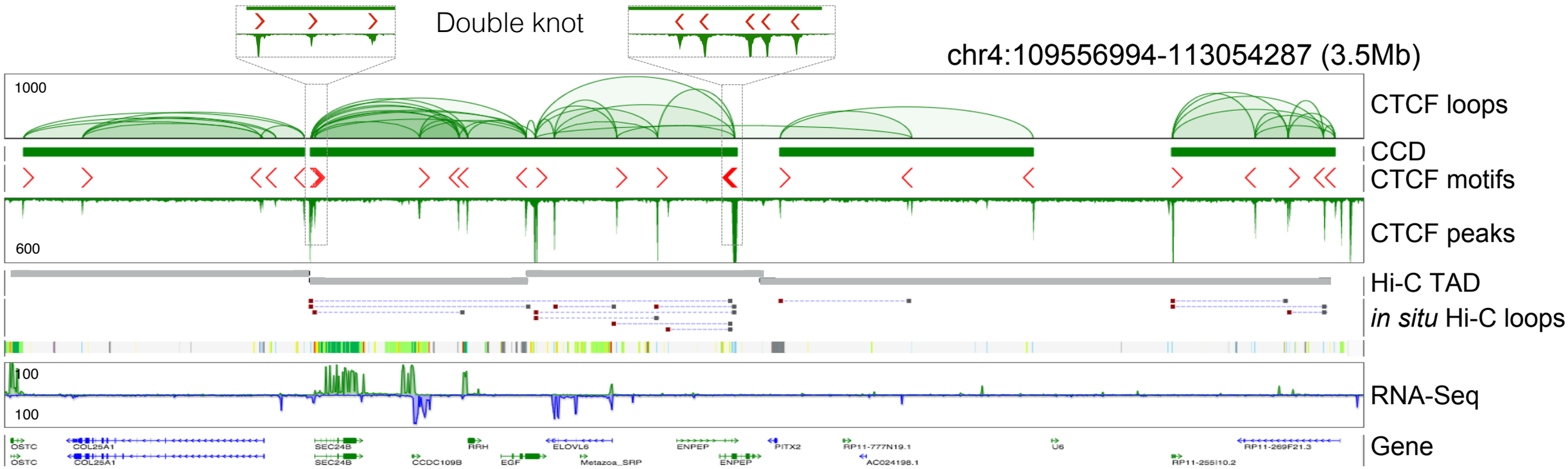
Genetic code:

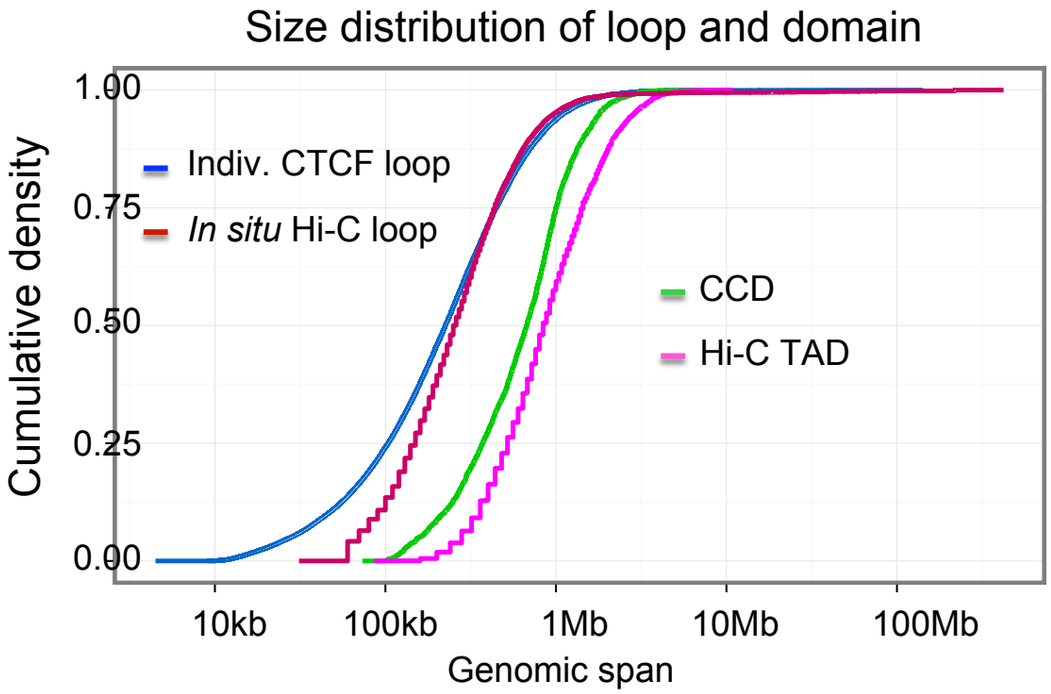
1. Gene codes (gene-centric views)
   protein coding sequences, codon usage
   TSS, exon, intron, splicing site, etc

2. Are there structure codes for genome topology?
   non-coding, distal, regulatory elements
   insulator, enhancer, repressor, etc

CFCT binding motif is a major kind of S-codes
- abundant, genome-wide, chromatin interactions
- Hi-C data showed CTCF associated w/ TAD (80%)
- ChIA-PET showed CTCF define chromatin topology

CTCF binding/looping defines chromatin topology



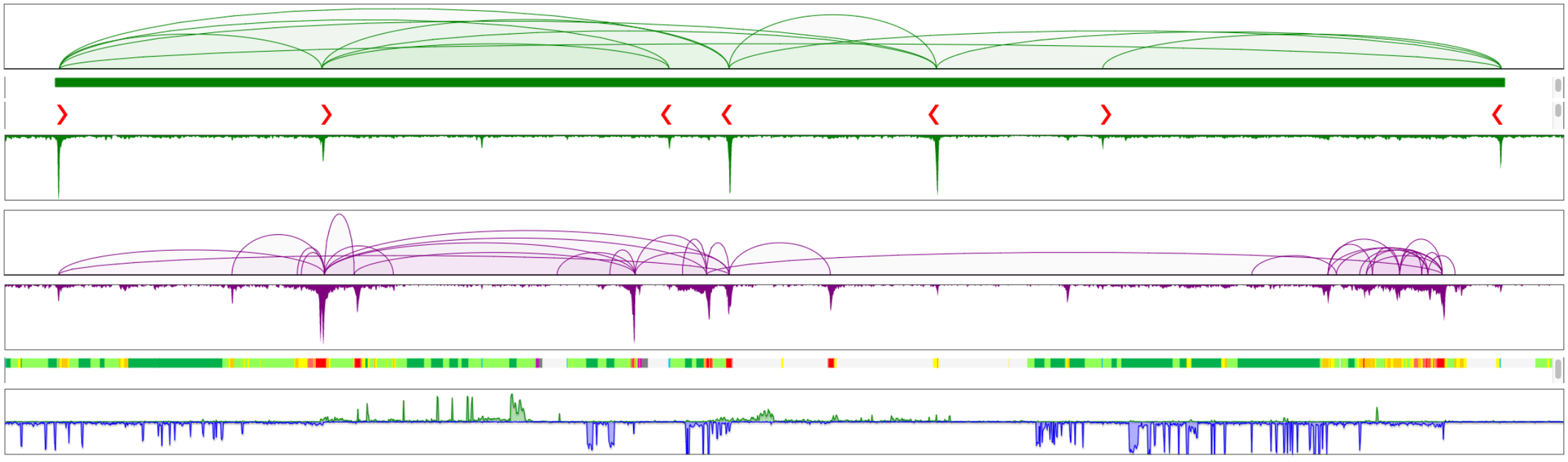


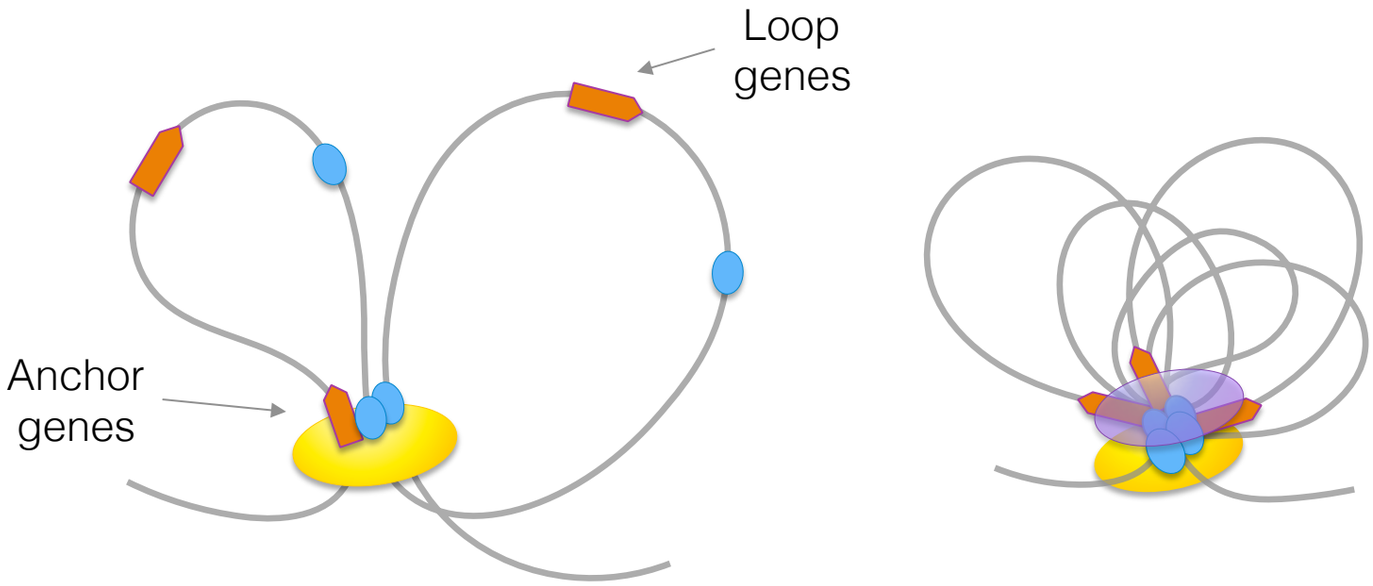
CCD = TAD

CTCF loops define detailed domain and sub-domain structures

CCD

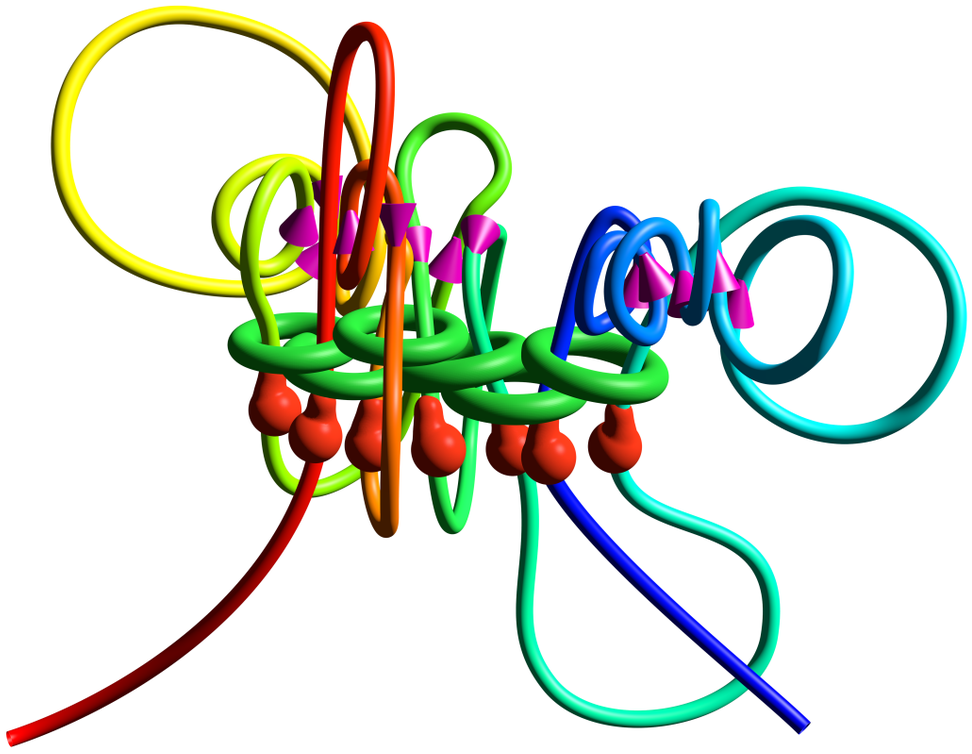
2D mapping data





Gene position/direction (60%)
align w/ CTCF motif orientation

3D simulation



Structure code variation and what affects it ?

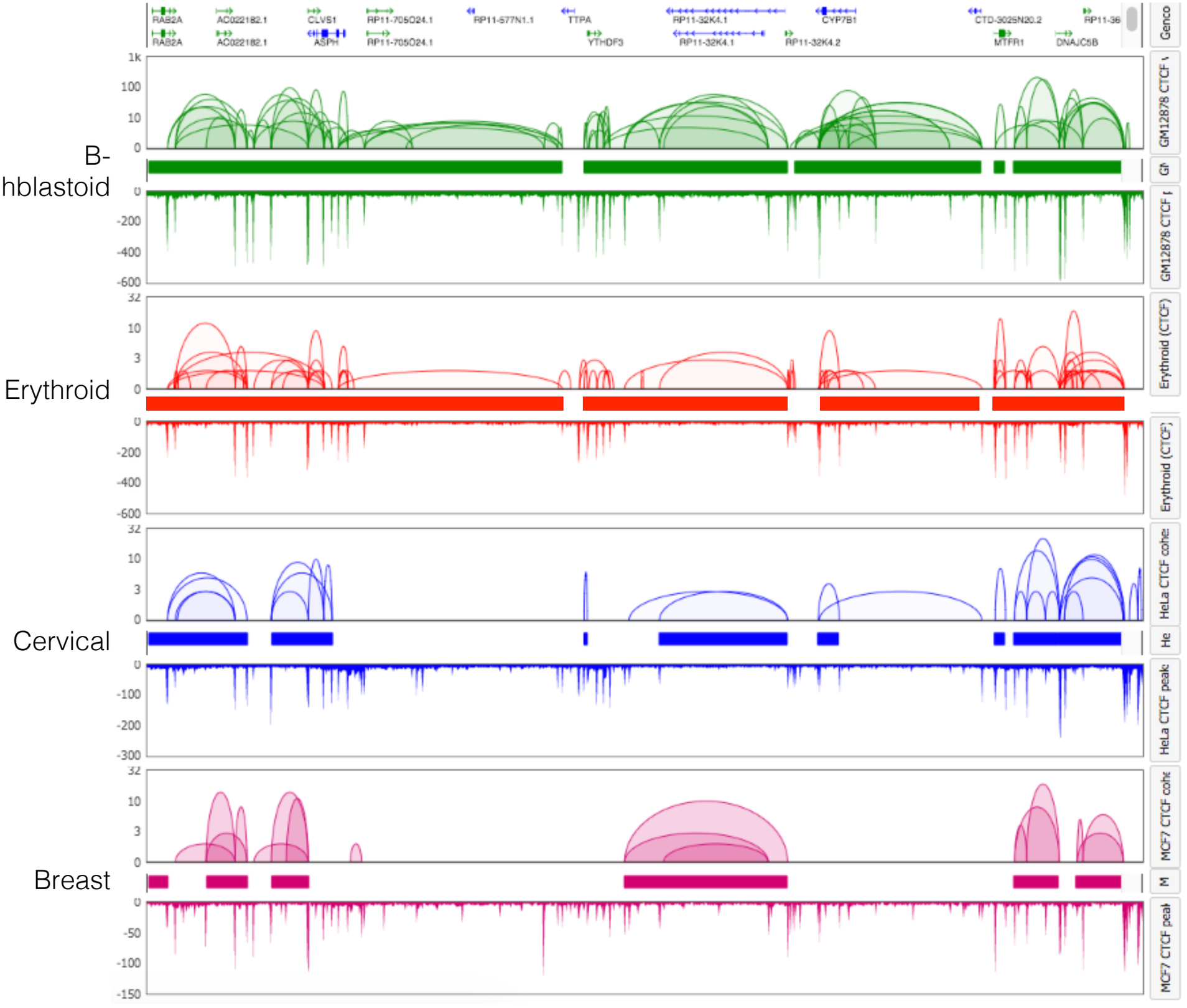
<u>Epigenetic effects</u>,

In same genotype (individual)
Diff. epigenotypes
Diff. cell types

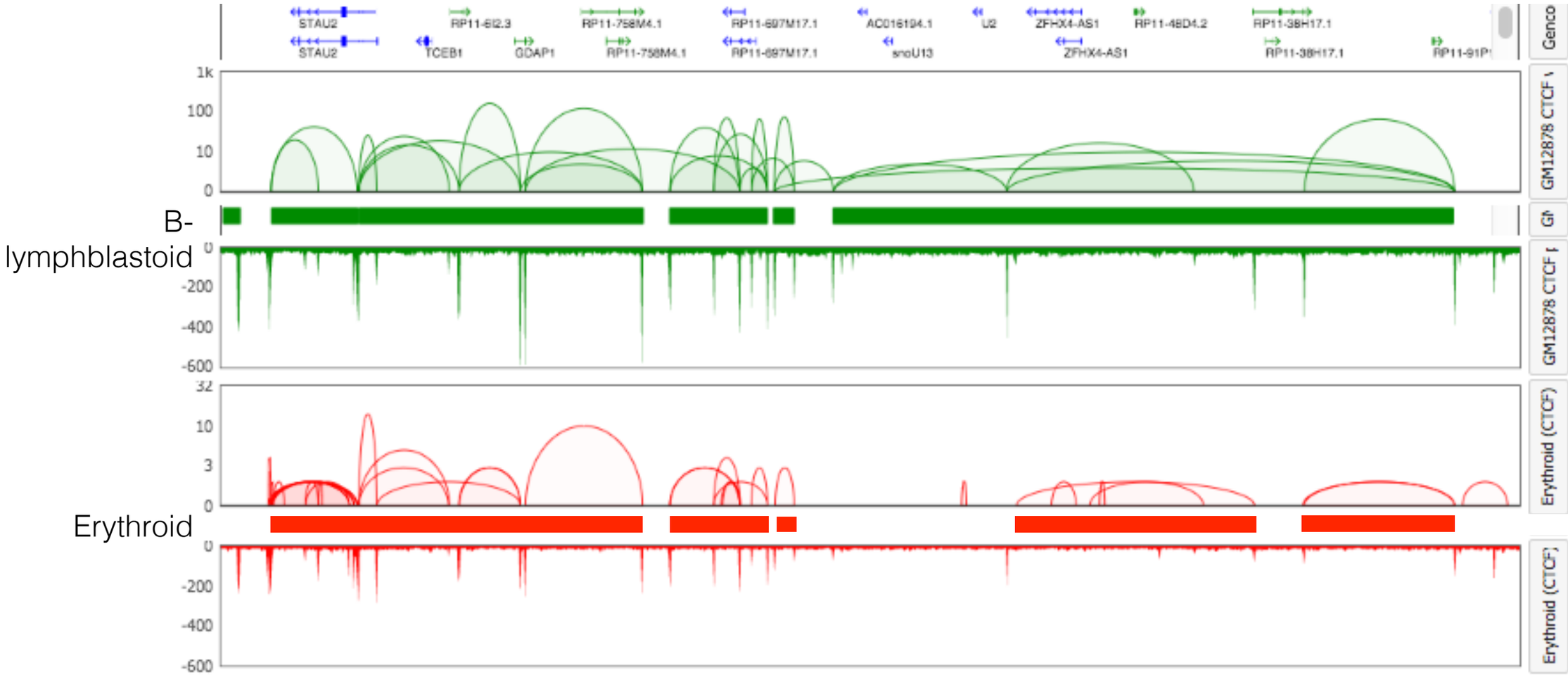
<u>Genetic effects</u>,

In same epigenotype (cell type)
Diff. genotypes
Diff. individuals

Cell type-specific CCD structure

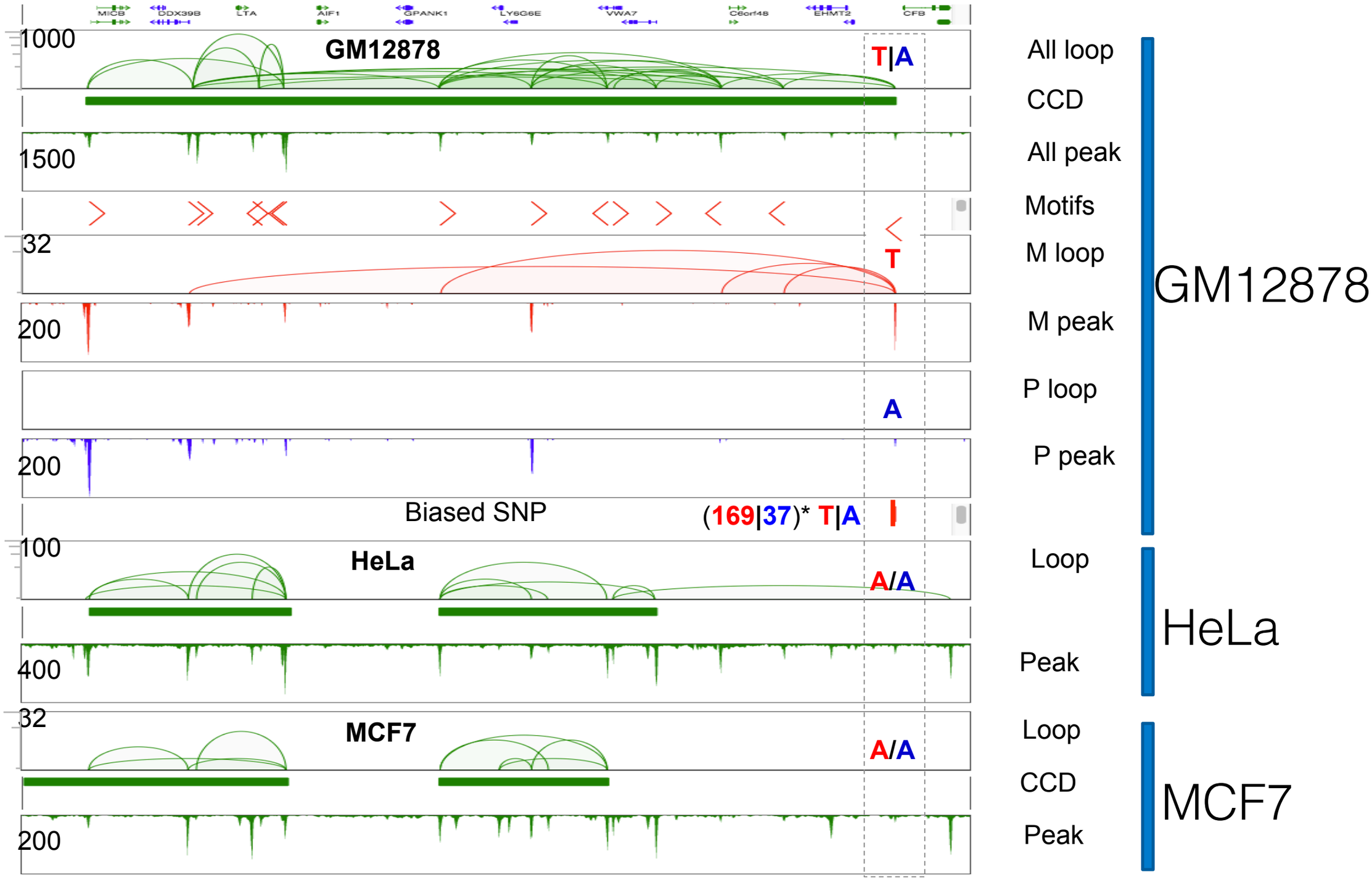


Cell type-specific CCD structure



- 3D genome architecture is dynamic during development and differentiation
- Chromatin topology could be a regulatory mechanism for cell-type specificity

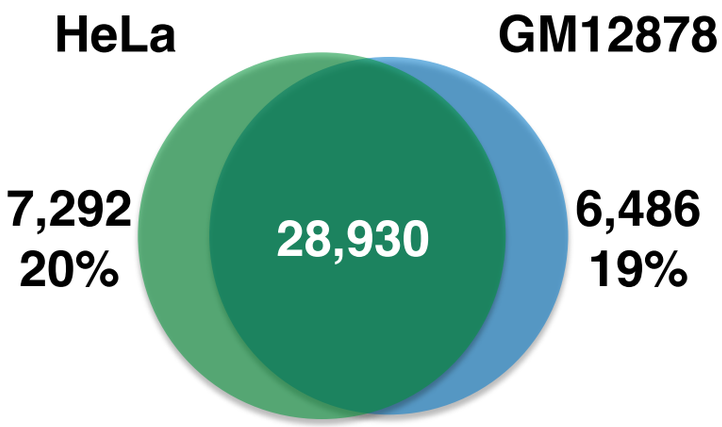
Genetic (SNP) validation of CTCF binding and looping

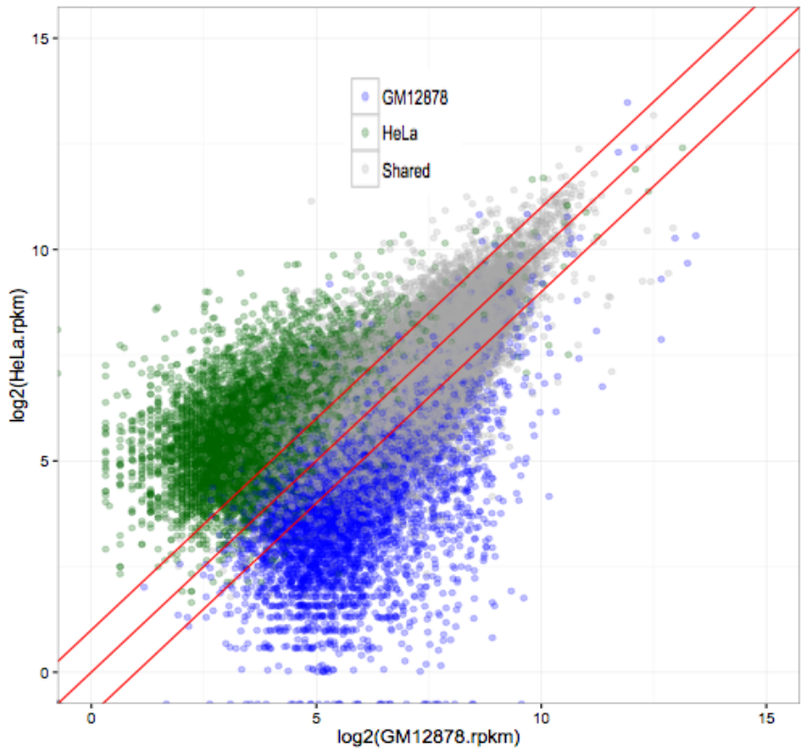


GM12878 and HeLa CTCF binding comparison

~20% CTCF bindings are exclusive to one cell type, two possible causes:

1. Genetic variation
2. Cell-type specificity





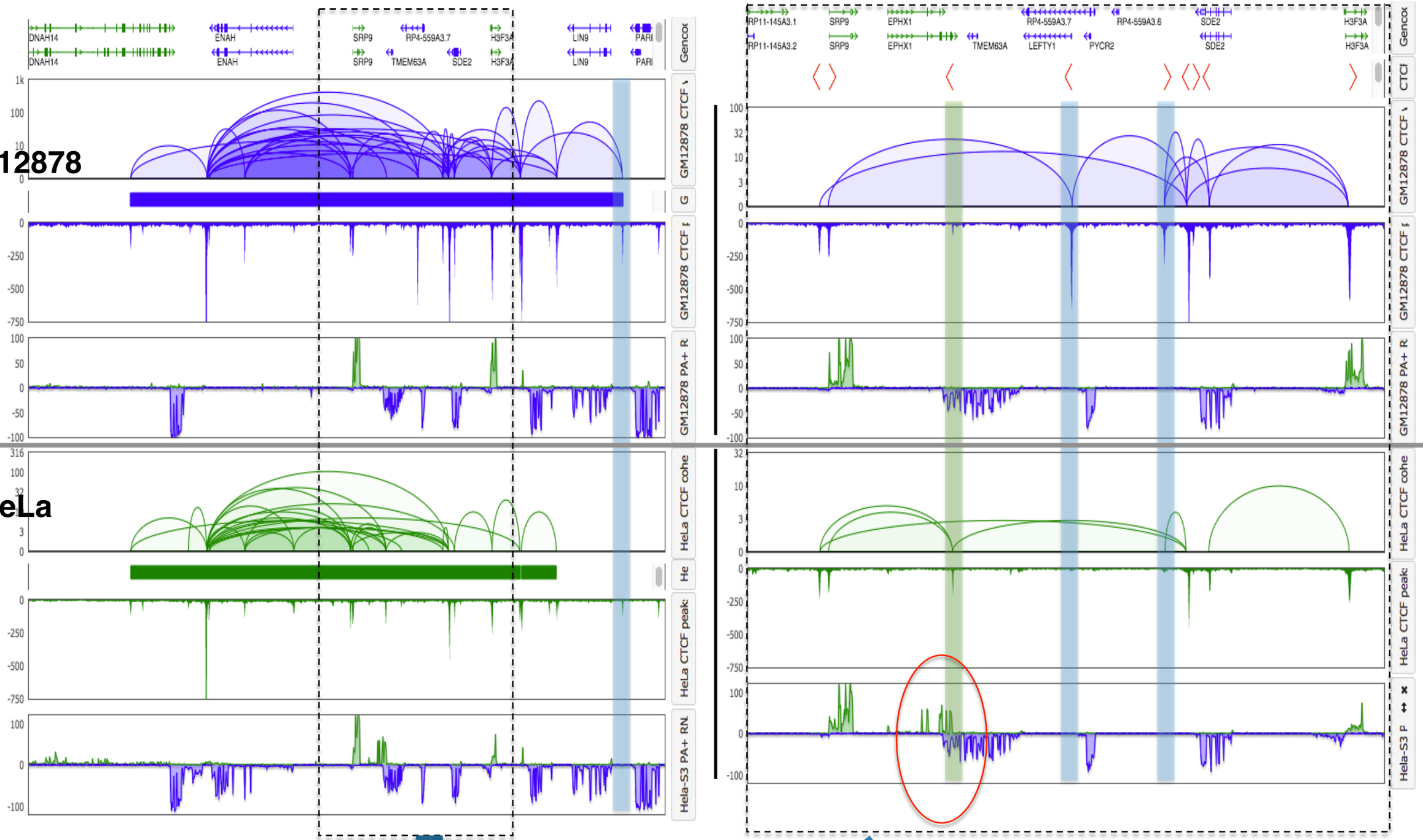
Example showing variation of CTCF binding/looping between GM12878 and HeLa

chr1:225296376-226608645

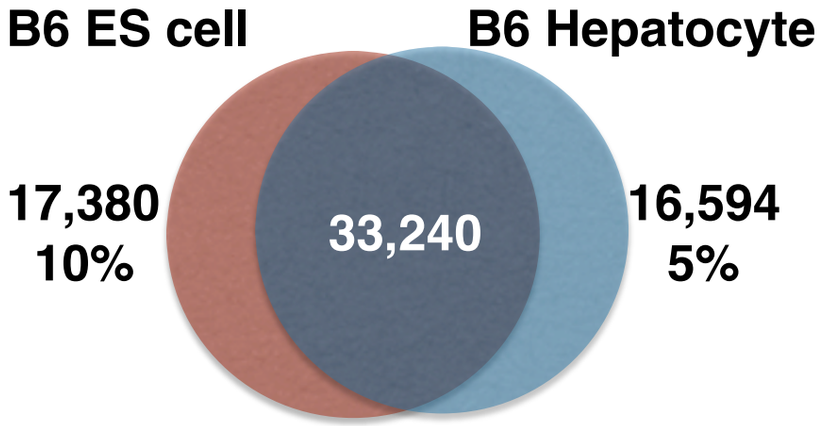
chr1:225920559-226271635

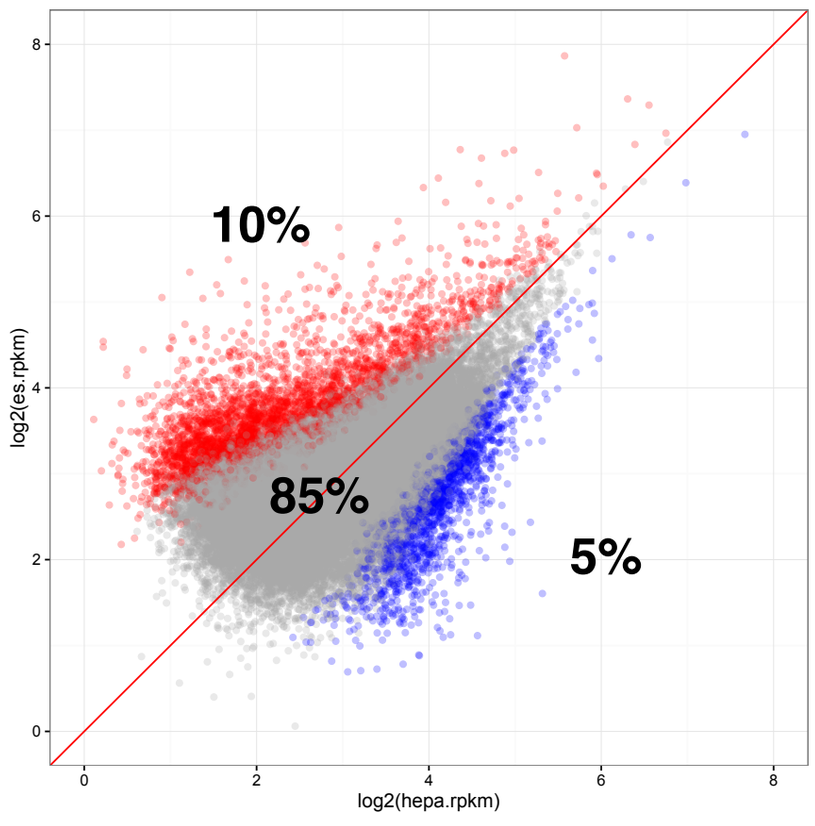
GM12878

HeLa

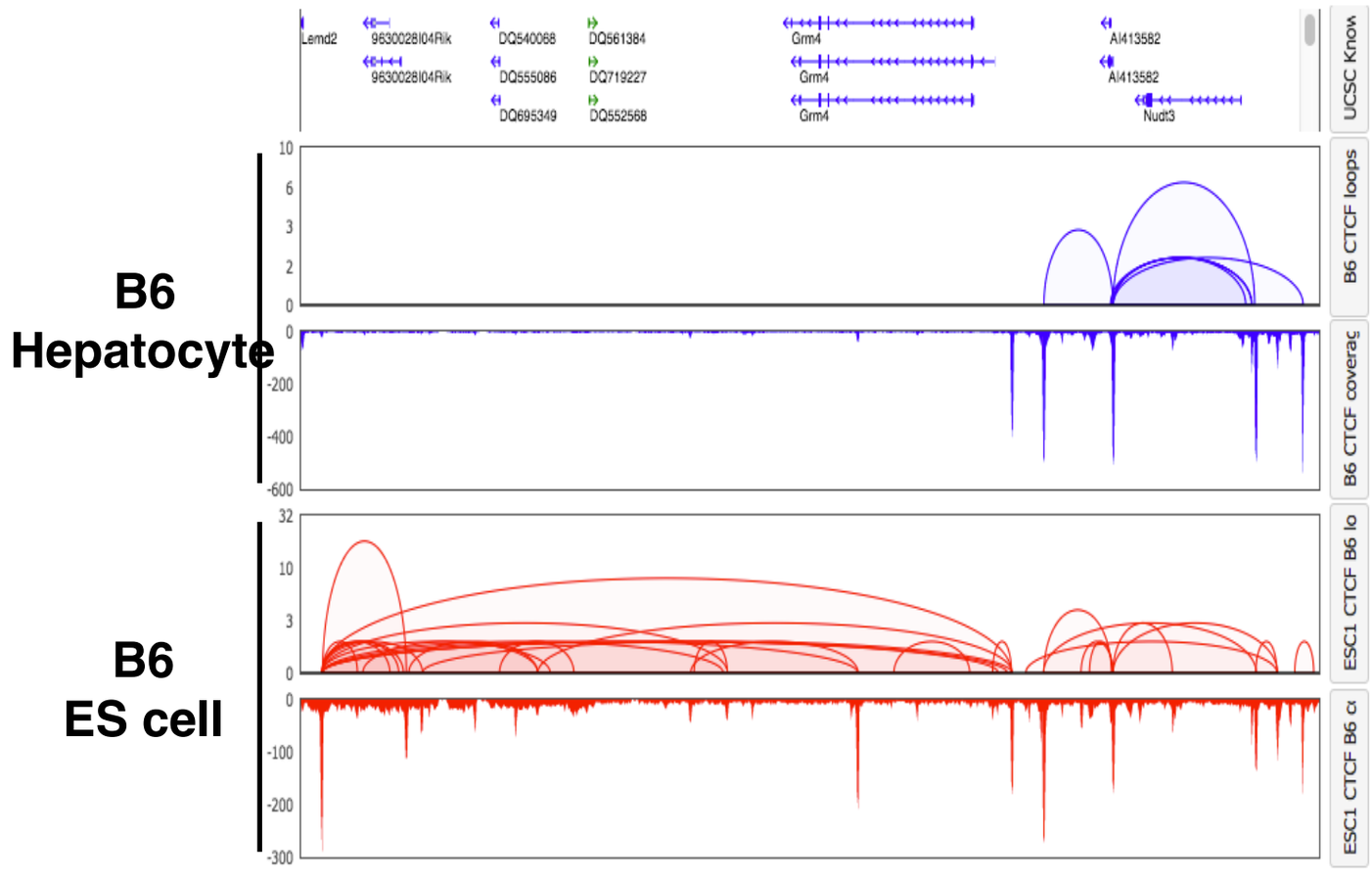


Chromatin topology structure variation in diff. cell types





chr17:27203495-27657812



Haplotype of B6 in ES vs Hepatocyte

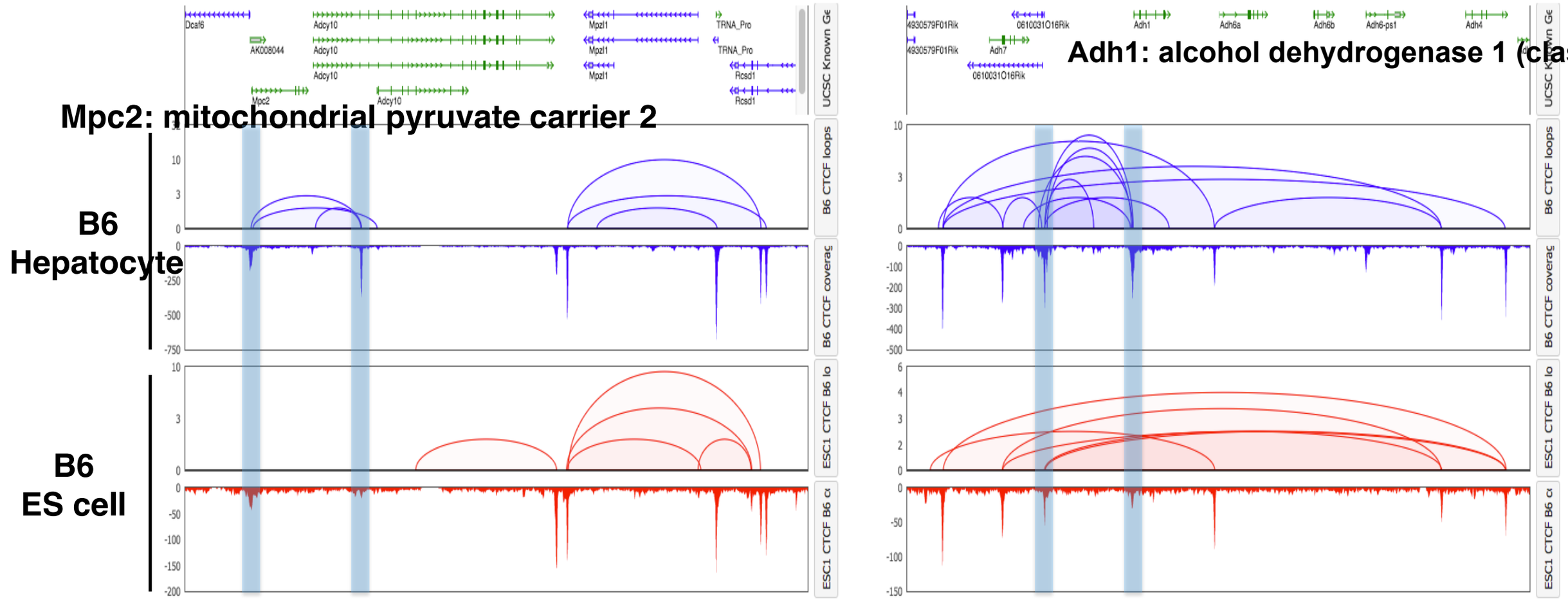
Chromatin topology structure variation in diff. cell types

chr1:165435344-165677124

chr3:138183453-138442028

Mpc2: mitochondrial pyruvate carrier 2

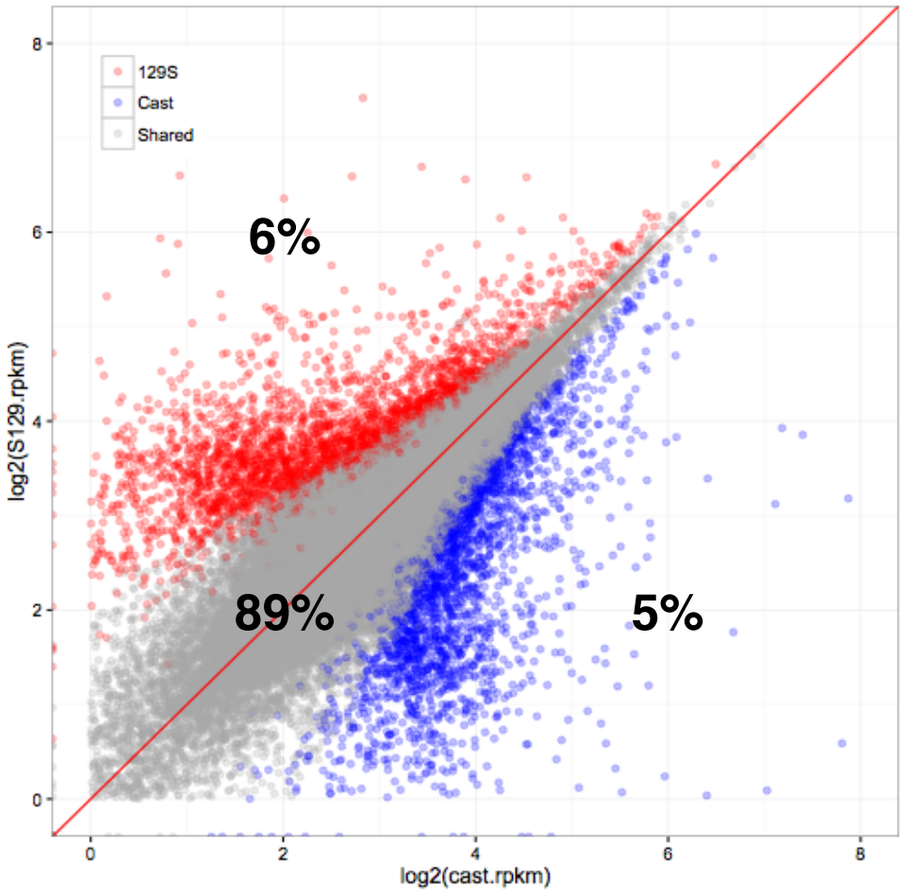
Adh1: alcohol dehydrogenase 1 (class I)

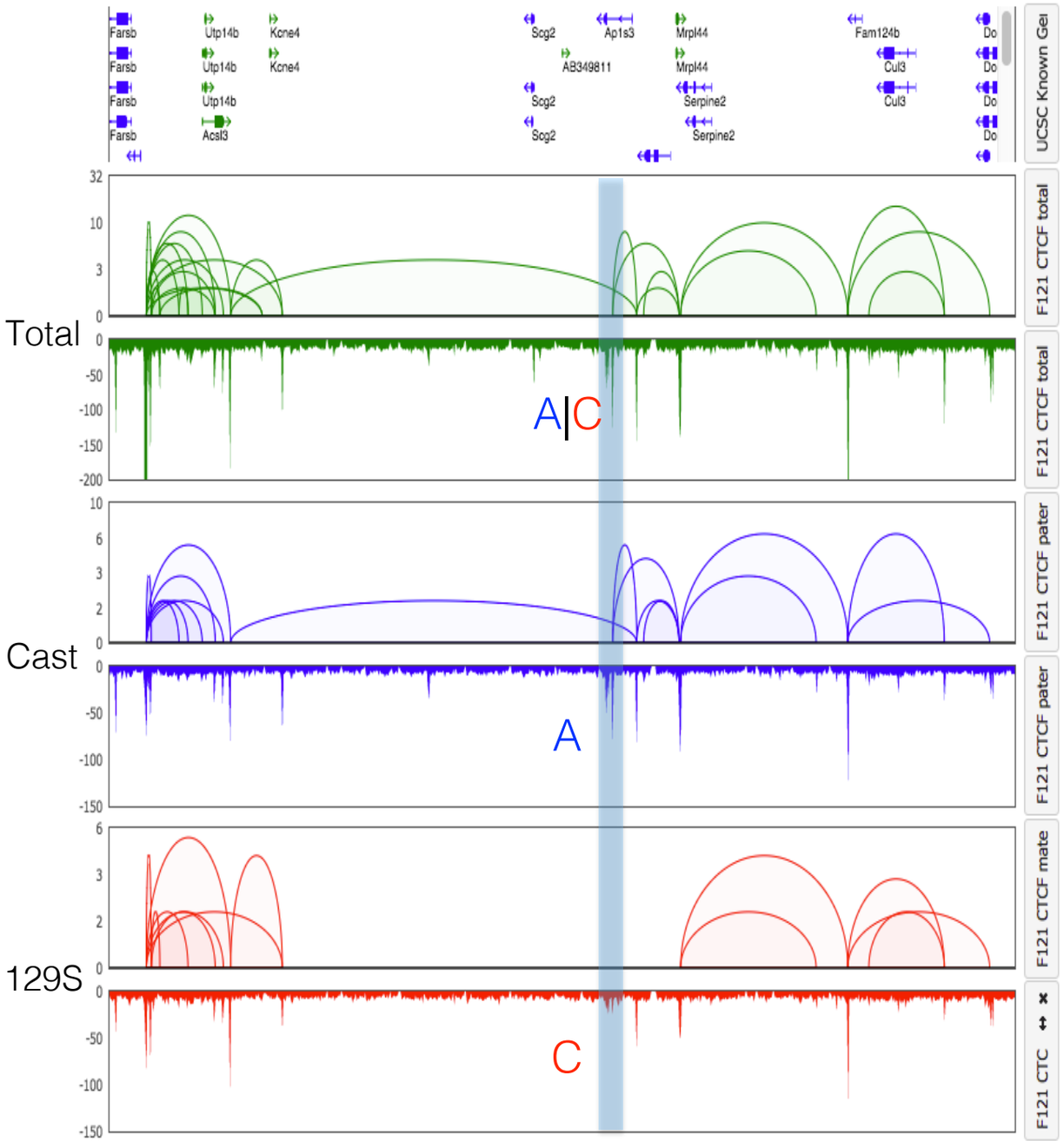


B6 Hepatocyte specific CTCF binding and looping surrounding hepatocyte specific genes

Chromatin topology structure variation in diff. genotypes

Two haplotypes in one cell type





Mouse ES cells of CAST x 129S

...AAGGCCAGAAGAGAGCGCCA...
...AAGGCCAGAAGAGGGCGCCA...

chr14:28388460-29367455

chr17:33966846-34101821

CAST x 129S

Tota

CAS

129

C|A

C

A

A|G

A

G

CAST x B6

Tota

CAS

B6

C|A

C

A

A|G

A

G

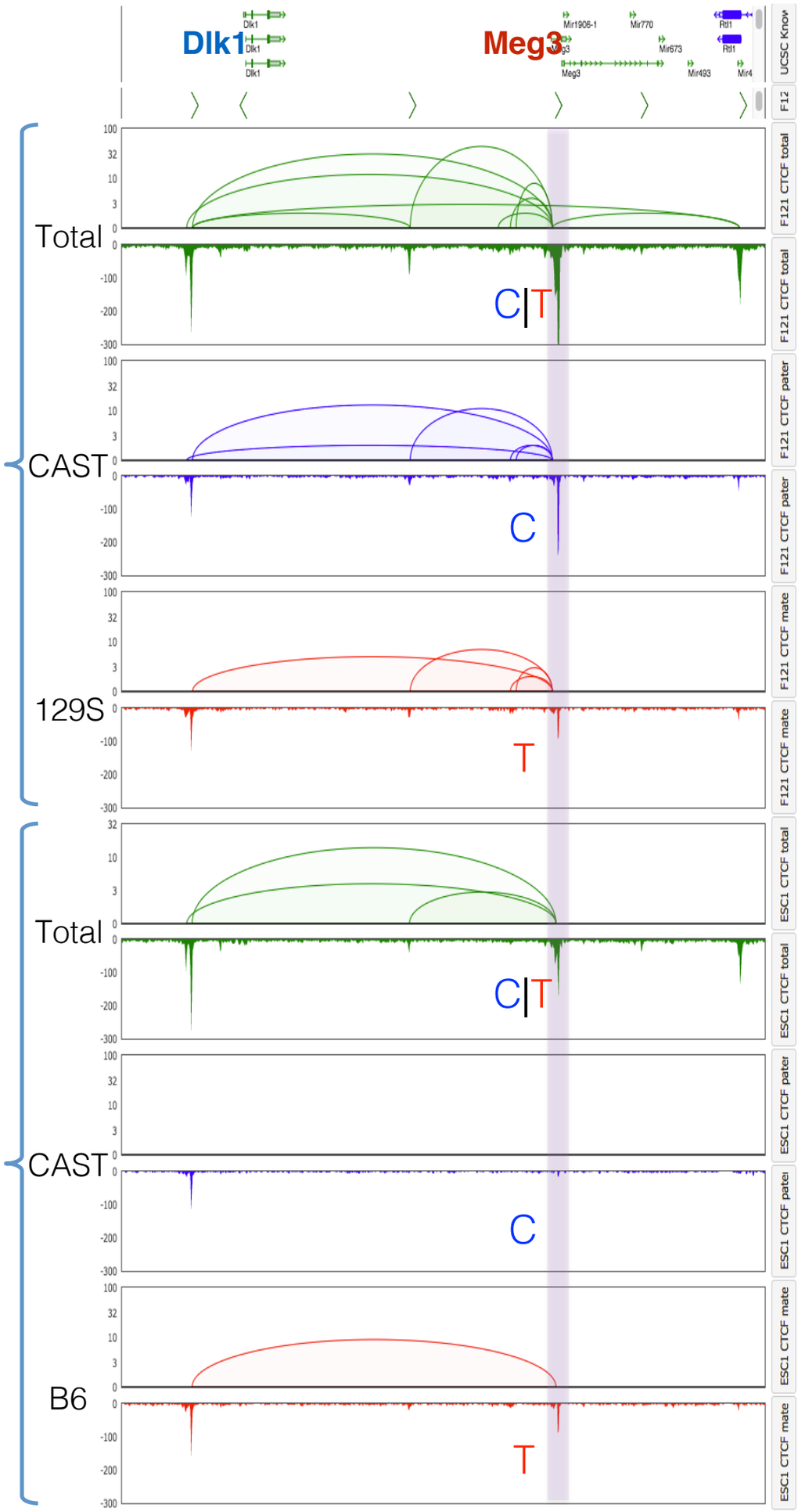
Genetic variant affects “weak/strong” binding/looping.

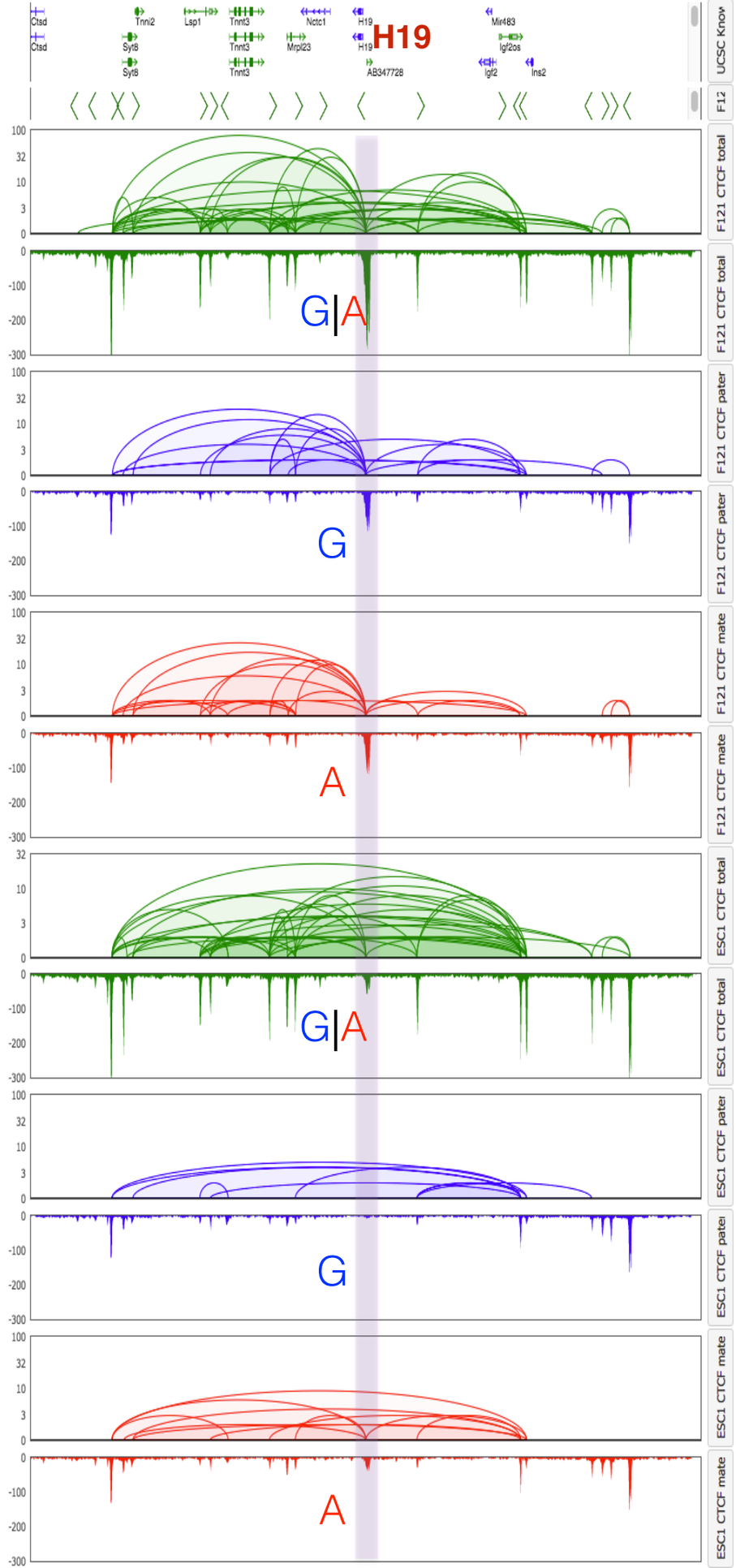
Genetic variant affects “on/off” binding/looping.

CAST x 129S

Same allele could behave differently due to unknown trans-acting effect. (Epistasis)

CAST x B6

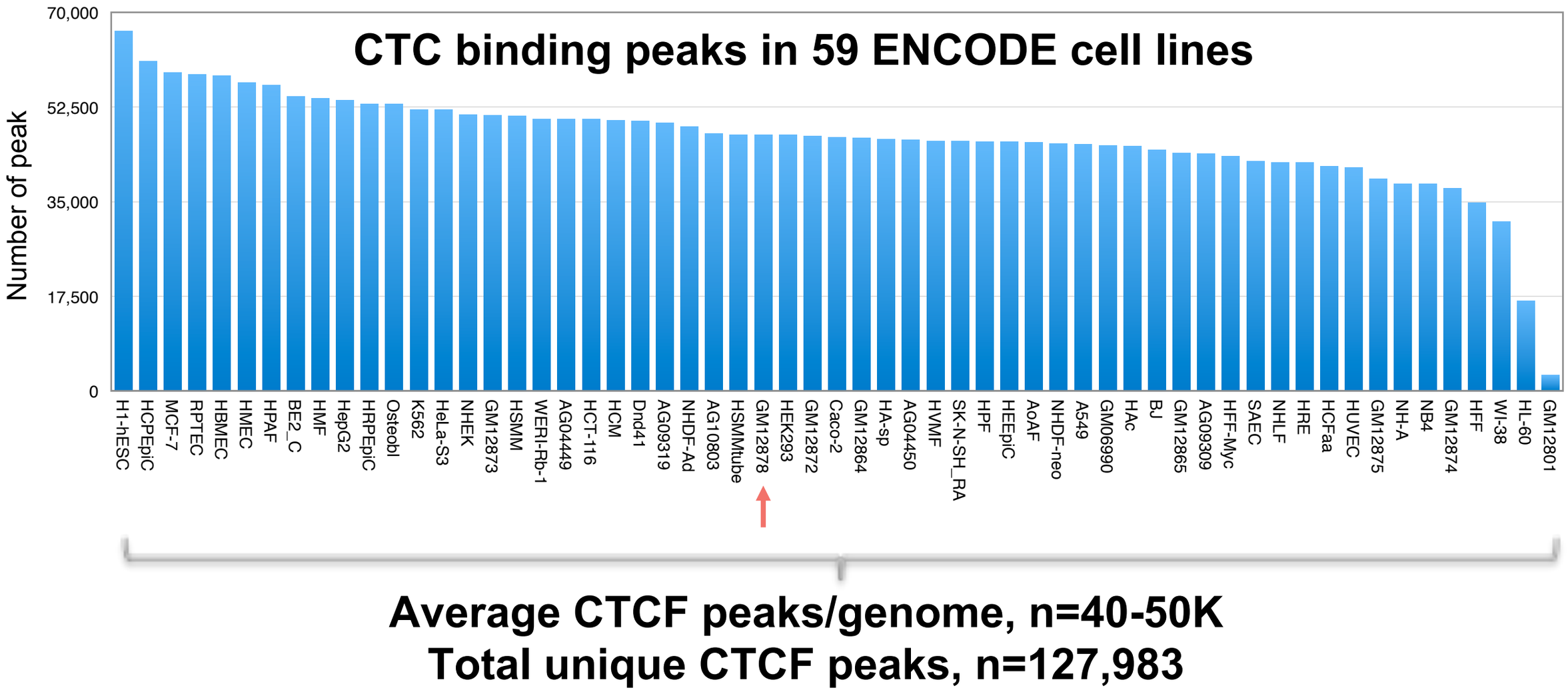




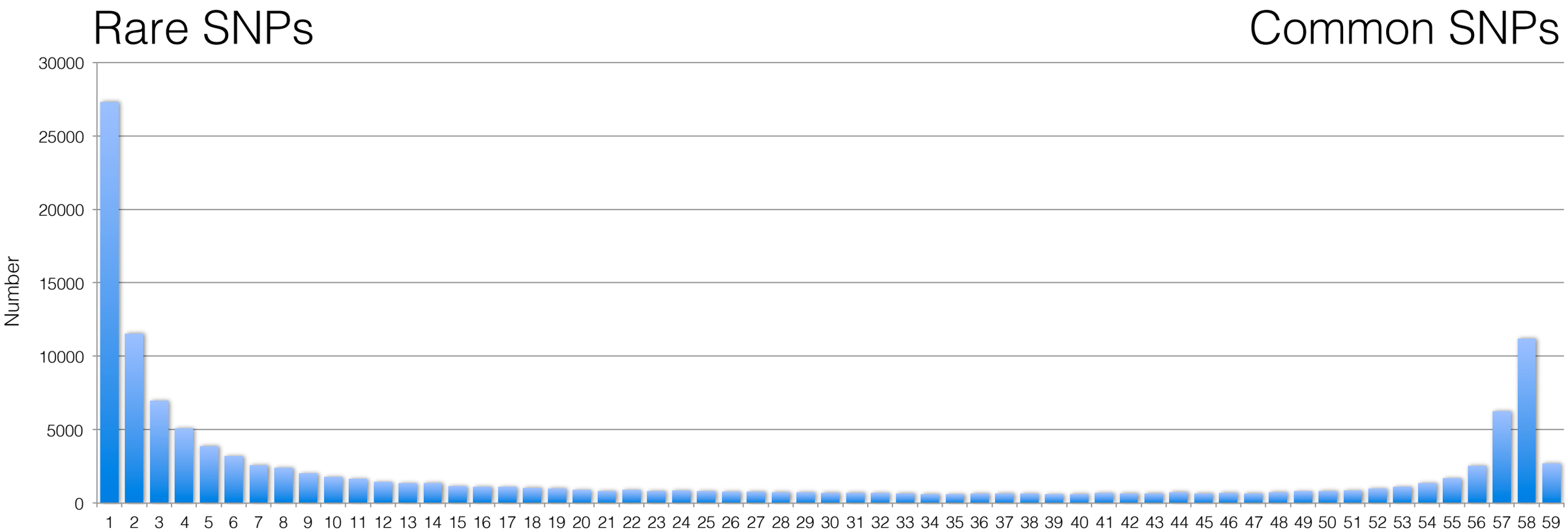
Total capacity of structure codes in human genome?

Possible CTCF motifs in a given genome, ~15 millions

(by scan the genome for motifs)



Average CTCF peaks/genome, n=40-50K
Total unique CTCF peaks, n=127,983



CTCF binding peak shared in different cell lines

Our strategy to study structure codes of chromatin topology

Vertical approach (epigenetic):

Same individual, many different cell types

Horizontal approach (genetic):

Same cell type, many different individuals

Comprehensive Mapping and Elucidating the Structure Codes in Human and Mouse Genomes

Aim 1. Chromatin topology and transcription regulation in ENCODE cells
(tie 1 & 2+ cells, ≈ 20-30 cell lines)

Aim 2. Mapping structure code in human hematopoietic cells
(vertical epigenetic approach, many blood cells from same individuals)

Aim 3. Mapping structure code in 1000 human population
(horizontal genetic approach, one cell type, 2500 individuals)

Aim 4. Mapping structure code in mouse models
(vertical & horizontal approach, 8 founder lines, 200s DO hybrids)

Aim 5. Mapping structure code in human disease populations
(100s lupus patient-derived b-cells, 100s T1D patient primary T-cells)

Experimental approaches

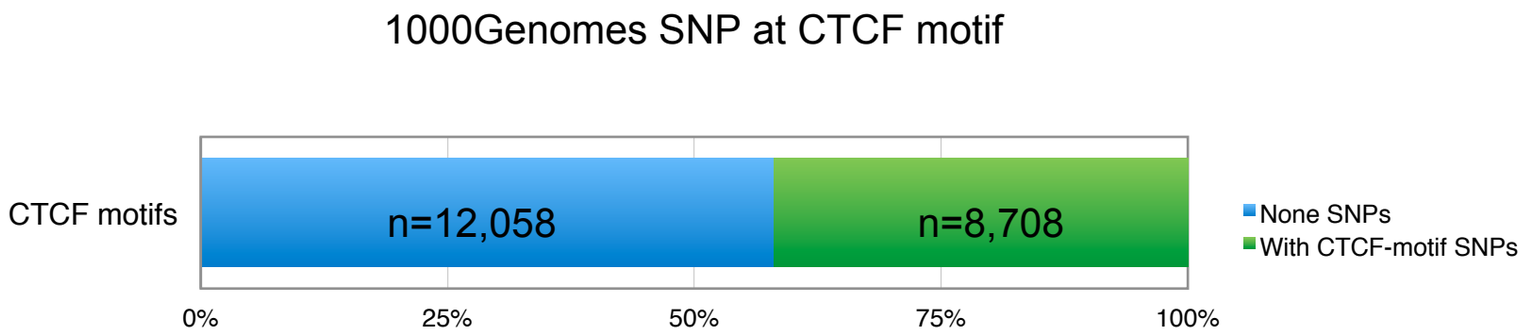
Multiplex ChIA-PET, 10s-100s (8-16 format)

CTCF, RNAPII, cell-specific TFs, RNA-Seq

Multiplex ChIP-Seq, 100s-1000s (96 format)

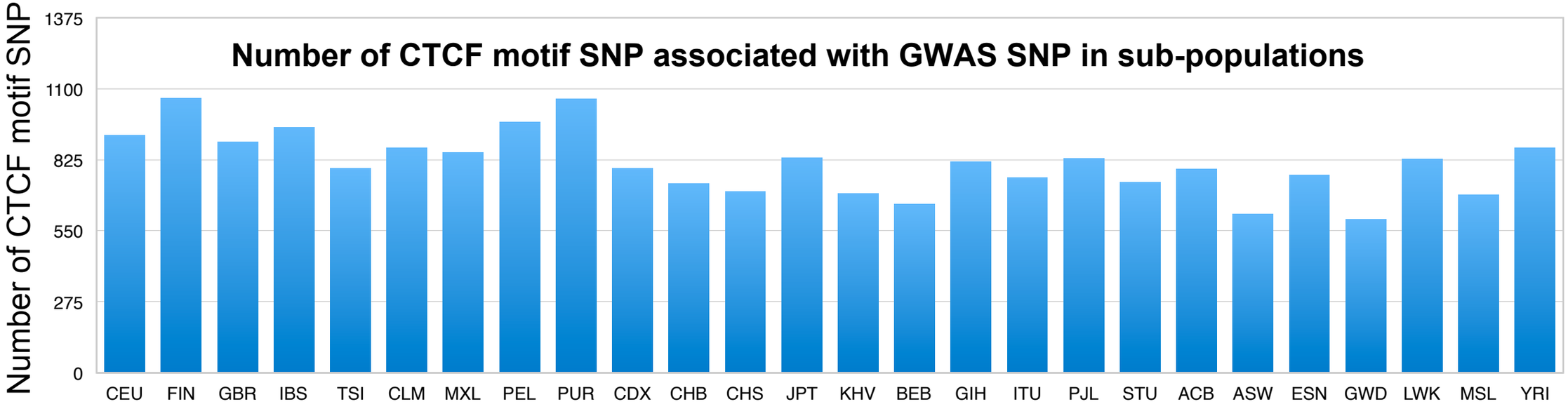
CTCF, RNA-Seq

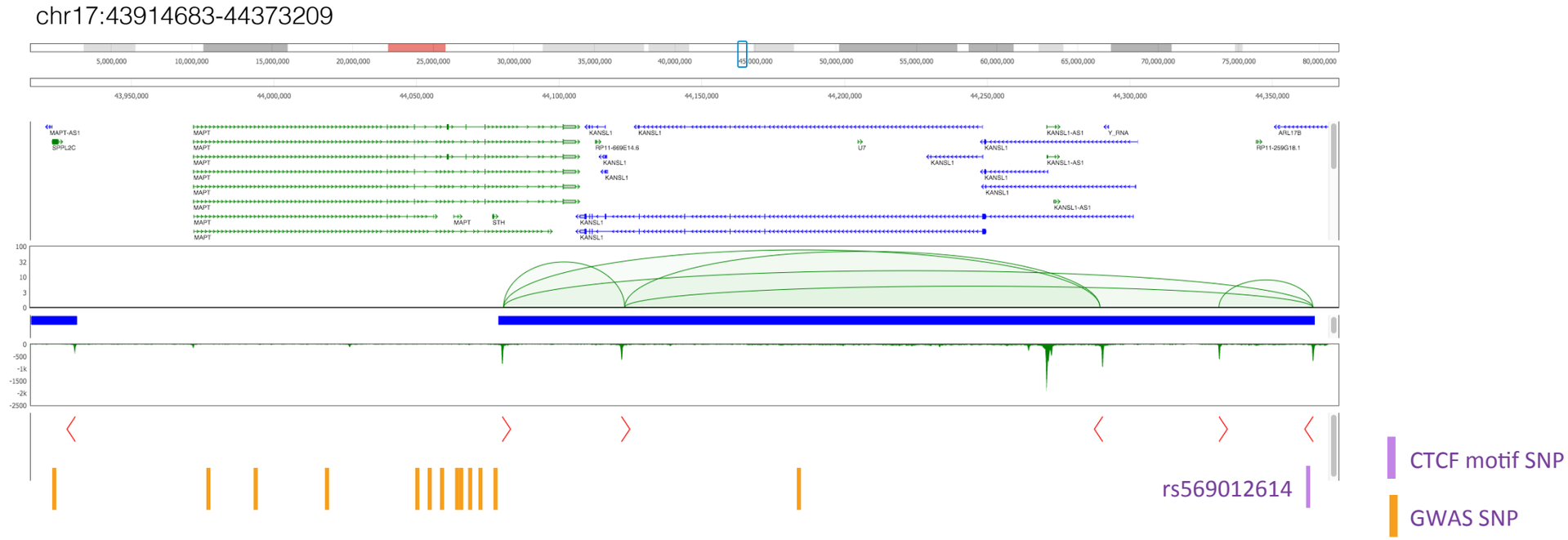
Preliminary assessment of the 1000 genomes



CTCF motif prohibits SNP in human genome

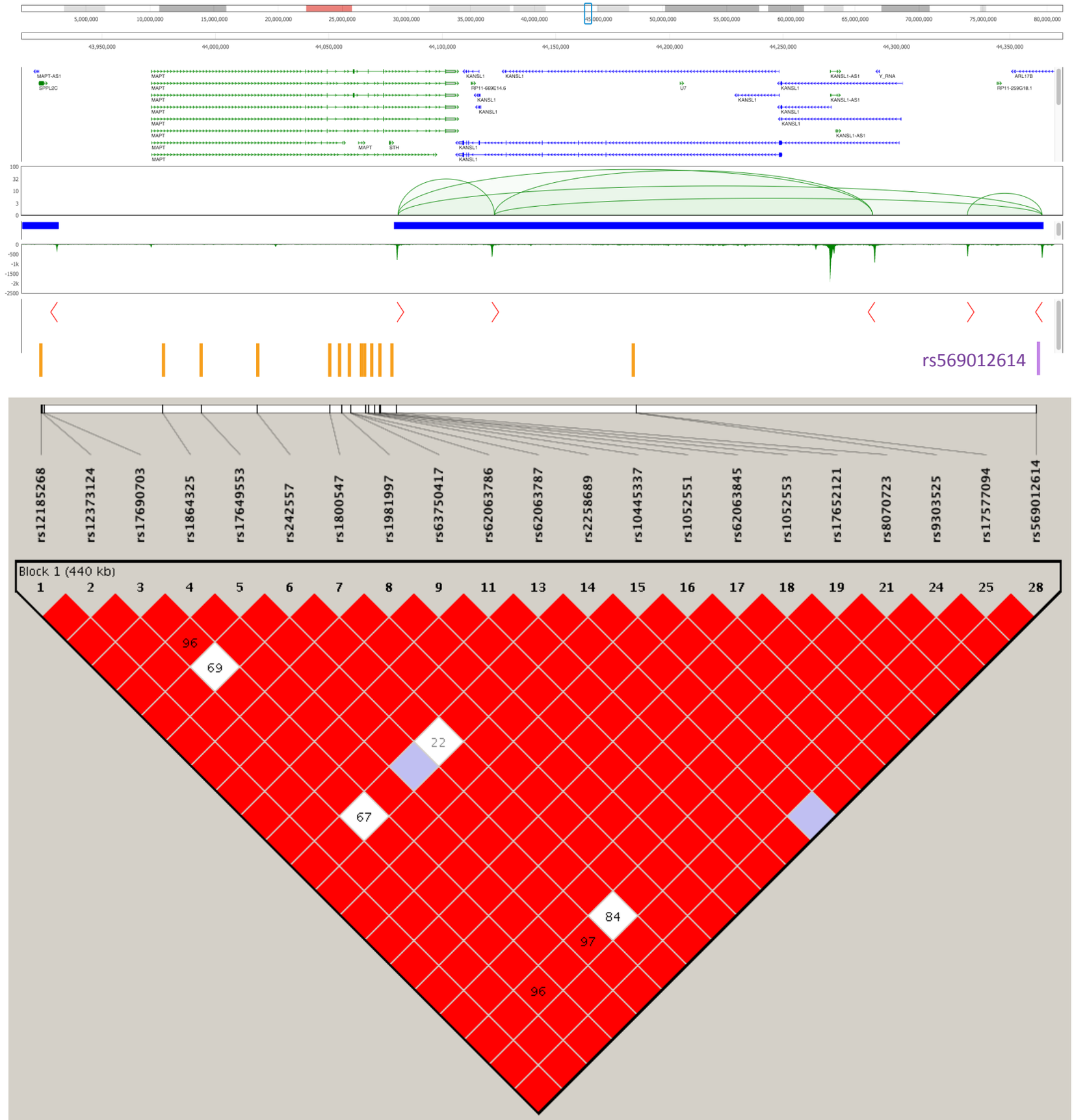
CTCF motifs	None SNPs	With CTCF-motif SNPs	Chi-Square Test
CTCF motifs	12058	8708	$p < 0.00001$
Random	10317	10449	
Gene coding regions	?	?	





CTCF_motif SNP	Location	Functional_SNP	Functional_SNP Types	D-prime	LOD	r-square	SNP function
rs569012614	CTCFboundary	rs11012	GWAS	0.882	15.12	0.654	
rs569012614	CTCFboundary	rs17631303	GWAS	0.919	16.19	0.685	
rs569012614	CTCFboundary	rs2942168	GWAS	1	26.25	0.909	Parkinson disease
rs569012614	CTCFboundary	rs393152	OMIM_GWAS	1	26.25	0.909	Parkinson disease
rs569012614	CTCFboundary	rs12185268	GWAS	1	26.25	0.909	Parkinson disease
rs569012614	CTCFboundary	rs12373124	GWAS	1	26.25	0.909	
rs569012614	CTCFboundary	rs17690703	GWAS	1	20.18	0.722	
rs569012614	CTCFboundary	rs1864325	GWAS	1	26.25	0.909	
rs569012614	CTCFboundary	rs17649553	GWAS	1	25.17	0.882	Parkinson disease
rs569012614	CTCFboundary	rs1800547	OMIM	1	26.25	0.909	Parkinson disease
rs569012614	CTCFboundary	rs1981997	GWAS	1	26.25	0.909	
rs569012614	CTCFboundary	rs63750417	clinVar	1	26.25	0.909	
rs569012614	CTCFboundary	rs62063786	clinVar	1	26.25	0.909	
rs569012614	CTCFboundary	rs62063787	clinVar	1	26.25	0.909	
rs569012614	CTCFboundary	rs10445337	clinVar	1	26.25	0.909	
rs569012614	CTCFboundary	rs1052551	clinVar	1	26.25	0.909	
rs569012614	CTCFboundary	rs62063845	clinVar	1	26.25	0.909	
rs569012614	CTCFboundary	rs1052553	clinVar	1	26.25	0.909	
rs569012614	CTCFboundary	rs17652121	clinVar	1	26.25	0.909	
rs569012614	CTCFboundary	rs8070723	GWAS	1	26.25	0.909	
rs569012614	CTCFboundary	rs9303525	GWAS	1	25.17	0.882	
rs569012614	CTCFboundary	rs17577094	GWAS	1	26.25	0.909	Parkinson disease
rs569012614	CTCFboundary	rs183211	GWAS	1	22.76	0.807	
rs569012614	CTCFboundary	rs199533	GWAS	1	28.89	0.968	Parkinson disease
rs569012614	CTCFboundary	rs199515	GWAS	0.967	26.57	0.936	Parkinson disease
rs569012614	CTCFboundary	rs415430	GWAS	0.966	24.85	0.903	Parkinson disease

chr17:43914683-44373209



CTCF motif SNP

GWAS SNP

Project Schedule (Proposed)

Grant Submission Timeline	Due Date	Days to Complete	Status Comment
Submit LOI	2/21/2016	25	Yijun/Jo Anne
Final Draft Review	3/14/2016	47	Red Team (JAX Peers)
Submit	3/17/2016	50	OSP
Narrative Preparation Timeline	**Due Date**	**Days to Complete**	**Status Comment**
Team Meeting (BH)	1/28/2016	1	
NIH Meeting (Elise Feingold, Mike Pazen)	2/5/2016	9	Jo Anne to organize
Budget	2/11/2016	15	Yijun, Team, Jon Maslow
First Complete Draft - ALL Sections	SEE BELOW		
Overall Goals: 6 pages	2/4/2016	8	Yijun
Experimental Assay Section: 12 pages	2/18/2016	22	Yijun, Greg, Laura (mouse)
Selection of Biological Samples Section: 6 pages	2/18/2016	22	Yijun, JB/VP, Greg/Laura
Data Management Plan: 6 pages	2/18/2016	22	Yijun, Greg, Mark
Project Management Plan: 6 pages	2/18/2016	22	Yijun
1000 genomes			Yijun
Mouse DO/CC			Greg and Laura
Disease- Lupus			JB and VP
Disease- T1D			Derya, Dave
Functional validation			Laura, Albert, Haoyi
Second Draft- ALL SECTIONS- Red team review	2/25/2016	29	
and REVIEW FOR INTEGRATION	3/7/2016	40	
Final Drafts- ALL SECTIONS	3/14/2016	47	
Final Production	**Due Date**	**Days to Complete**	**Status Comment**
Forms Package	3/17/2016	50	

GM12878 SNP in CTCF motifs

Heterozygous SNPs
n=1,411,442

Homozygous SNPs
n=2,378,153

CTCF motif at CTCF anchor
n=20,766

CTCF motif with heterozygous SNP
n=173

CTCF motif with homozygous SNP
n=286

Haplotype specific binding bias
N=70

Article

December 17, 2015

Cell

CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription

Zhonghui Tang,[1,12] Os
Przemyslaw Szalaj,[4,5,6]
Emaly Piecuch,[1,3] Ping
Xiaoan Ruan,[1] Chia-Li
and Yijun Ruan[1,2,3,*]
[1]The Jackson Laboratory
[2]National Key Laboratory
Hubei 430070, China

hu,[1,3]
rcki,[7] Paul Michalski,[1]
Laurent M. Sachs,[9]
liang Li,[2,11]

ltural University, Wuhan,