Exploring the effects of CTCF binding site mutations on transcriptional regulation in cancer cell lines

JAMES JUSUF

MENTOR DR. MAHMOUD GHANDI

SIXTH ANNUAL PRIMES CONFERENCE, MAY 2016

Outline

- **Introduction**
  - Background
  - Purposes
  - Previous Work
- **Research**
  - Scoring Mutations
  - Monoallelic Expression
- **Future Work**

Background

Genetics of cancer

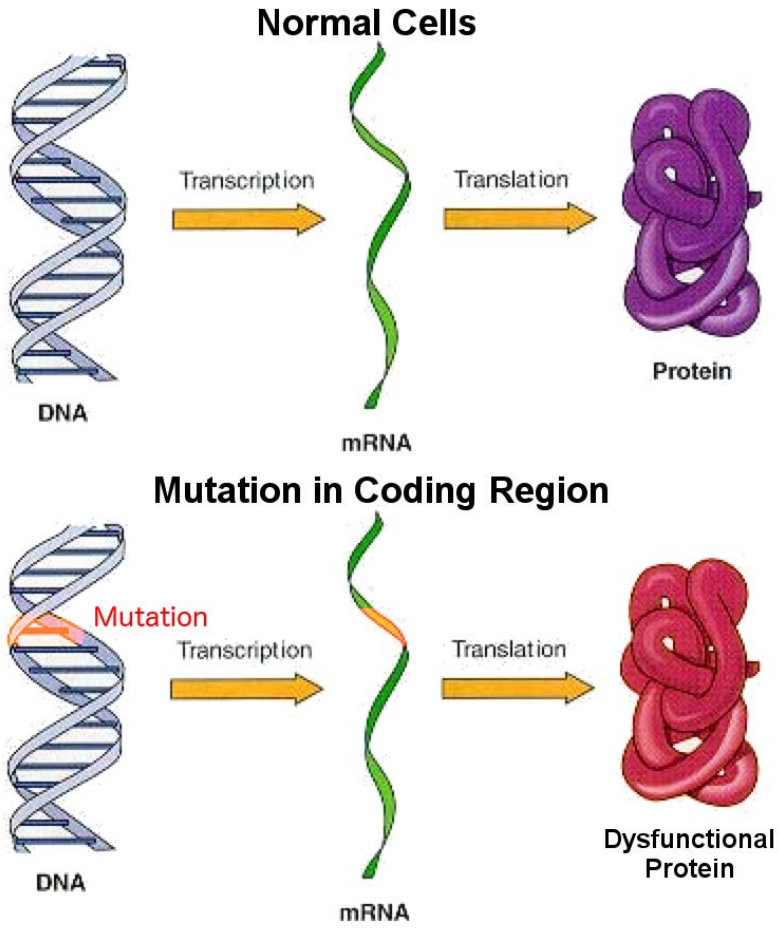


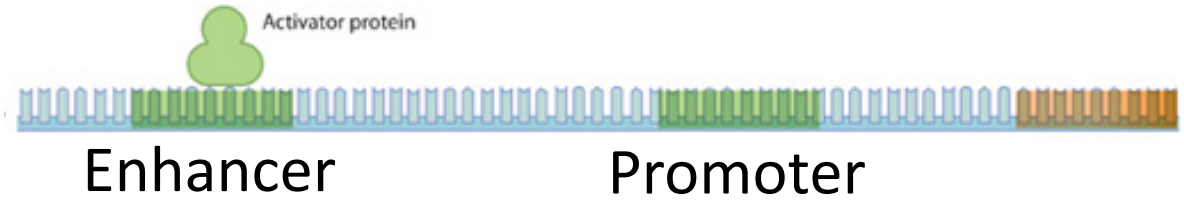
Fig. 1 Effects of mutations in coding regions

- Mutations in coding regions affect:
  - Genes/proteins
- Mutations in noncoding regions affect:
  - Regulation of gene expression

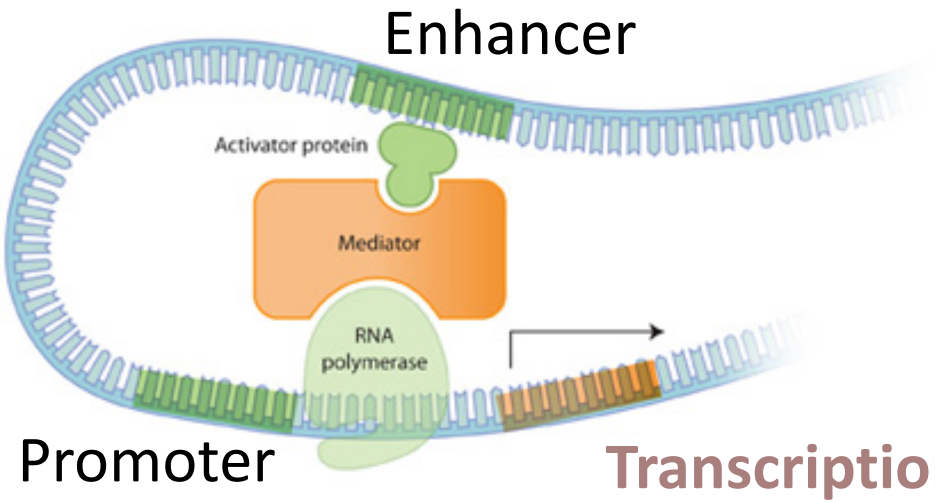
Background

Gene transcription

Fig. 2 Interactions between enhancers and promoters cause transcription



Transcription start site



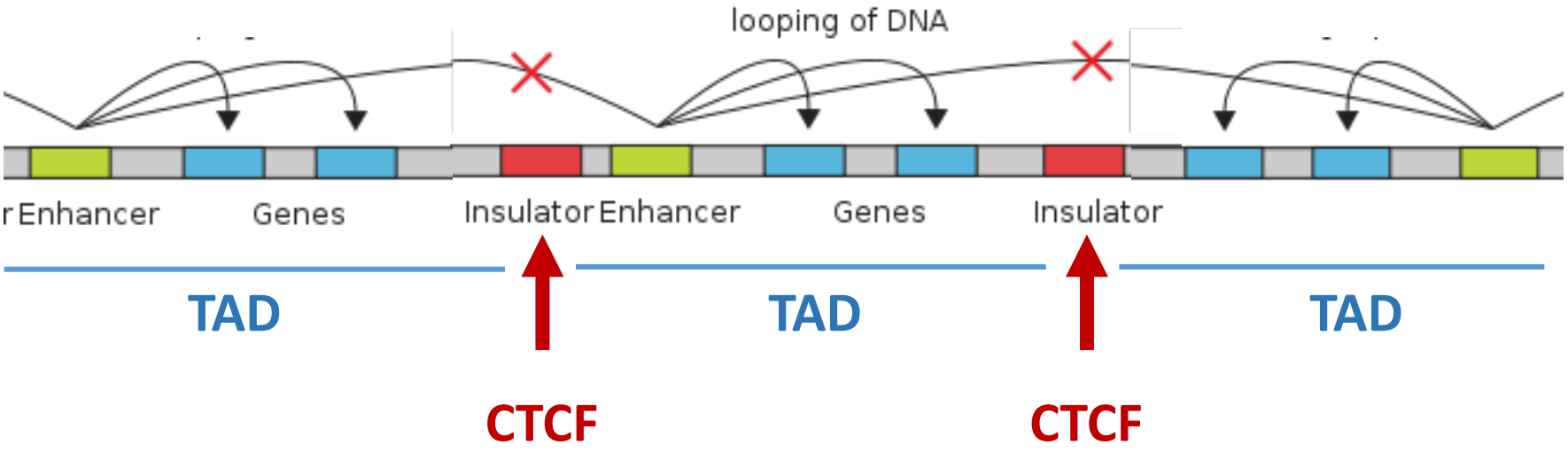
Transcription occurs

Background

What is CTCF?

- CCCTC-binding factor
- Protein that interacts with noncoding regions
- Partitions DNA into topologically associating domains (TADs)

Fig. 3 Interaction of CTCF and DNA



Background

CTCF

- What factors impair CTCF binding activity?
- Hypermethylation of a specific CTCF binding site shown to increase expression of oncogene *PDGFRA*

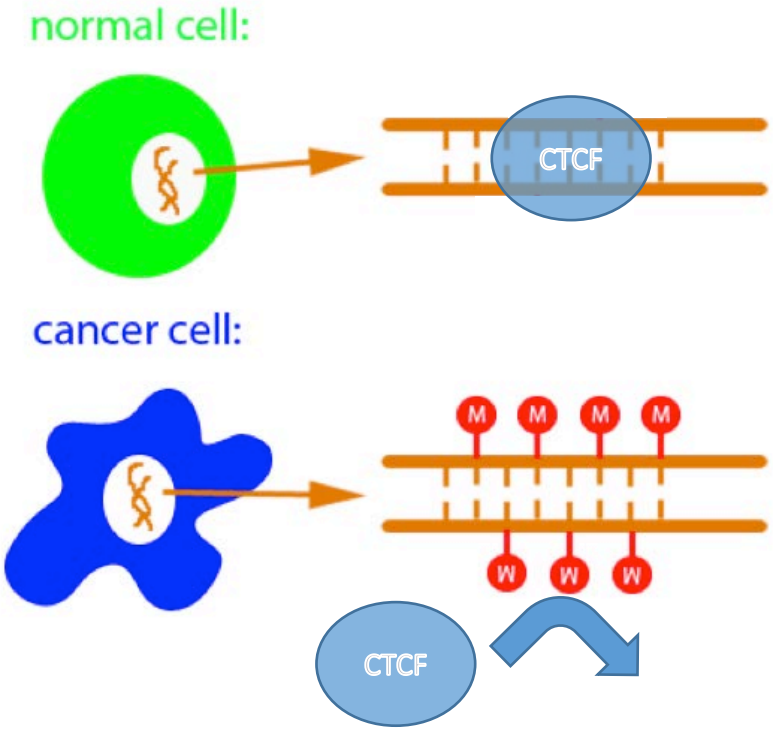


Fig. 5 DNA methylation impairs CTCF binding

Purposes

- To determine the existence of **mutations in CTCF binding sites that significantly affect binding activity**, and the genes and cell lines in which they occur

- To elucidate if **such mutations alter transcriptional activity through the loss of a domain boundary**, possibly leading to cancer
  1. Monoallelic expression
  2. Tumor dependencies

Previous Work

CTCF binding sites

- Region where CTCF protein interacts with DNA
- ≈14-base pair DNA sequence motif
- Can tolerate variations

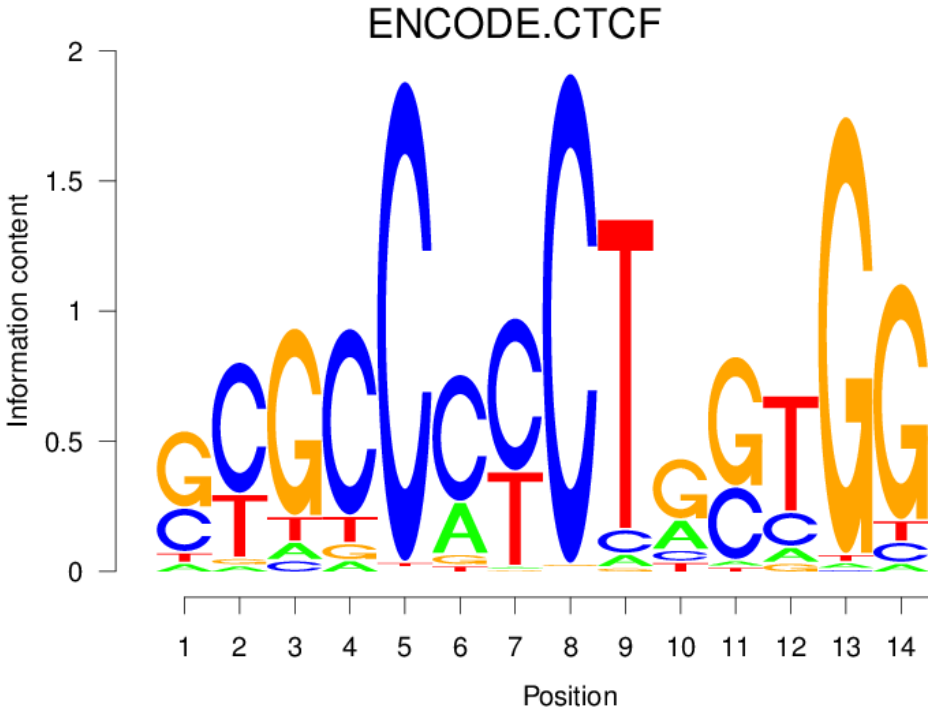


Fig. 4 The CTCF binding motif

Previous Work

How do we quantify CTCF binding activity?

- **Assign numerical scores to CTCF binding sites**
- gkm-SVM
  - Machine learning approach
  - Shown to be highly accurate in predicting CTCF binding sites

Outline

- **Introduction**
  - Background
  - Purposes
  - Previous Work
- **Research**
  - Scoring Mutations
  - Monoallelic Expression
- **Future Work**

Scoring Mutations

How do we quantify the effect of mutations on CTCF binding activity?

- Focus on single nucleotide variants (SNVs)
- Find **the change in CTCF binding score (ΔSVM)** calculated using gkm-SVM

- Example:

GGACATAAGTGCATTCACCT**A**CTGGATGGCGTAAGGGCTGA = 4.349582

GGACATAAGTGCATTCACCT**C**CTGGATGGCGTAAGGGCTGA = 3.270352

ΔSVM = –1.079230

Scoring Mutations

How accurately can gkm-SVM predict the effect of CTCF binding site mutations?

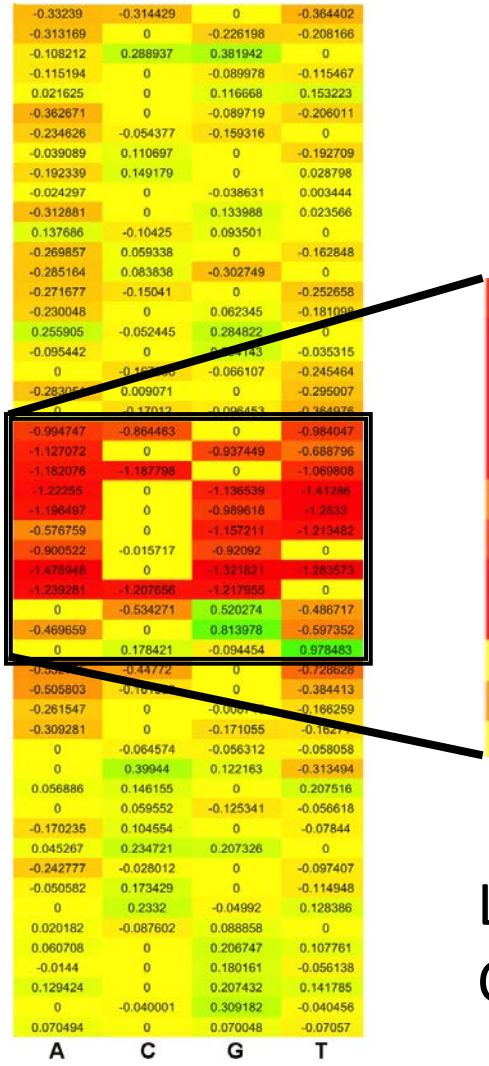
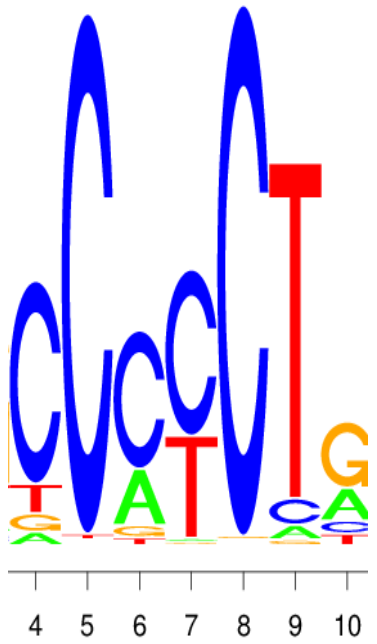


Fig. 8 ΔSVM for all possible SNVs in the vicinity of a CTCF binding site

-0.994747	-0.864463	0	-0.984047
-1.127072	0	-0.937449	-0.688796
-1.182076	-1.187798	0	-1.069808
-1.22255	0	-1.136539	-1.41286
-1.196497	0	-0.989618	-1.2833
-0.576759	0	-1.157211	-1.213482
-0.900522	-0.015717	-0.92092	0
-1.476948	0	-1.321821	-1.283573
-1.239281	-1.207656	-1.217955	0
0	-0.534271	0.520274	-0.486717
-0.469659	0	0.813978	-0.597352
0	0.178421	-0.094454	0.978483

A C G T

Largest changes seen within CTCF binding motif



Scoring Mutations

Large datasets

- Mutations are commonly listed in Mutation Annotation Format (MAF)

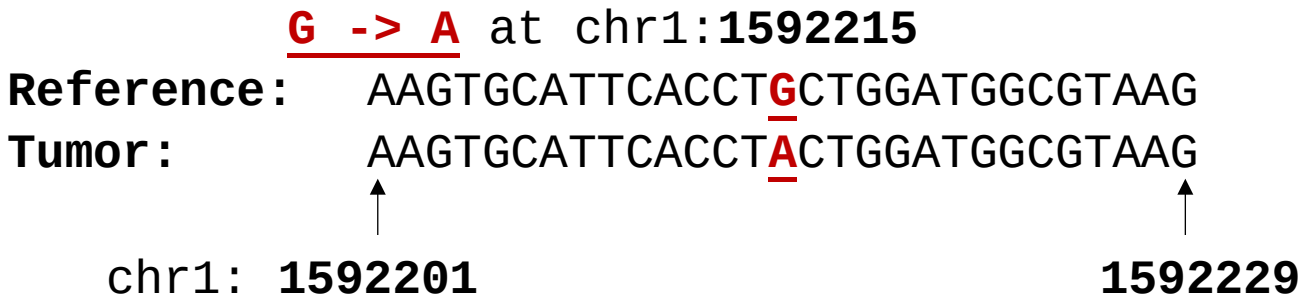
Table 1 Example MAF file

Hugo_Symbol	Entrez_Gene	Center	Chr	Position	Str	Reference_A	Tumor_Seq_
SAMD11	148398	broad.mit.edu	1	876826	+	G	A
SKI	6497	broad.mit.edu	1	2169428	+	C	G
NOL9	79707	broad.mit.edu	1	6607435	+	C	A
Unknown	0	broad.mit.edu	1	11925732	+	C	T
Unknown	0	broad.mit.edu	1	14155426	+	C	T
NBL1	4681	broad.mit.edu	1	19969708	+	C	T
SNHG3-RCC1	751867	broad.mit.edu	1	28833143	+	G	A
Unknown	0	broad.mit.edu	1	31402468	+	G	A
Unknown	0	broad.mit.edu	1	32915081	+	C	T
Unknown	0	broad.mit.edu	1	48175635	+	C	T
Unknown	0	broad.mit.edu	1	59118254	+	A	C

Scoring Mutations

Large datasets

1. Filter irrelevant mutations that appear far from CTCF binding sites
   - ChIP-seq data shows general regions of high CTCF activity
2. Gather information about the sequence surrounding each mutation
   - Lookup surrounding sequence in reference genome (+/- 14 base pairs)
   - Example:



Scoring Mutations

Large datasets

New columns

Table 2 Example MAF file after scoring and filtering

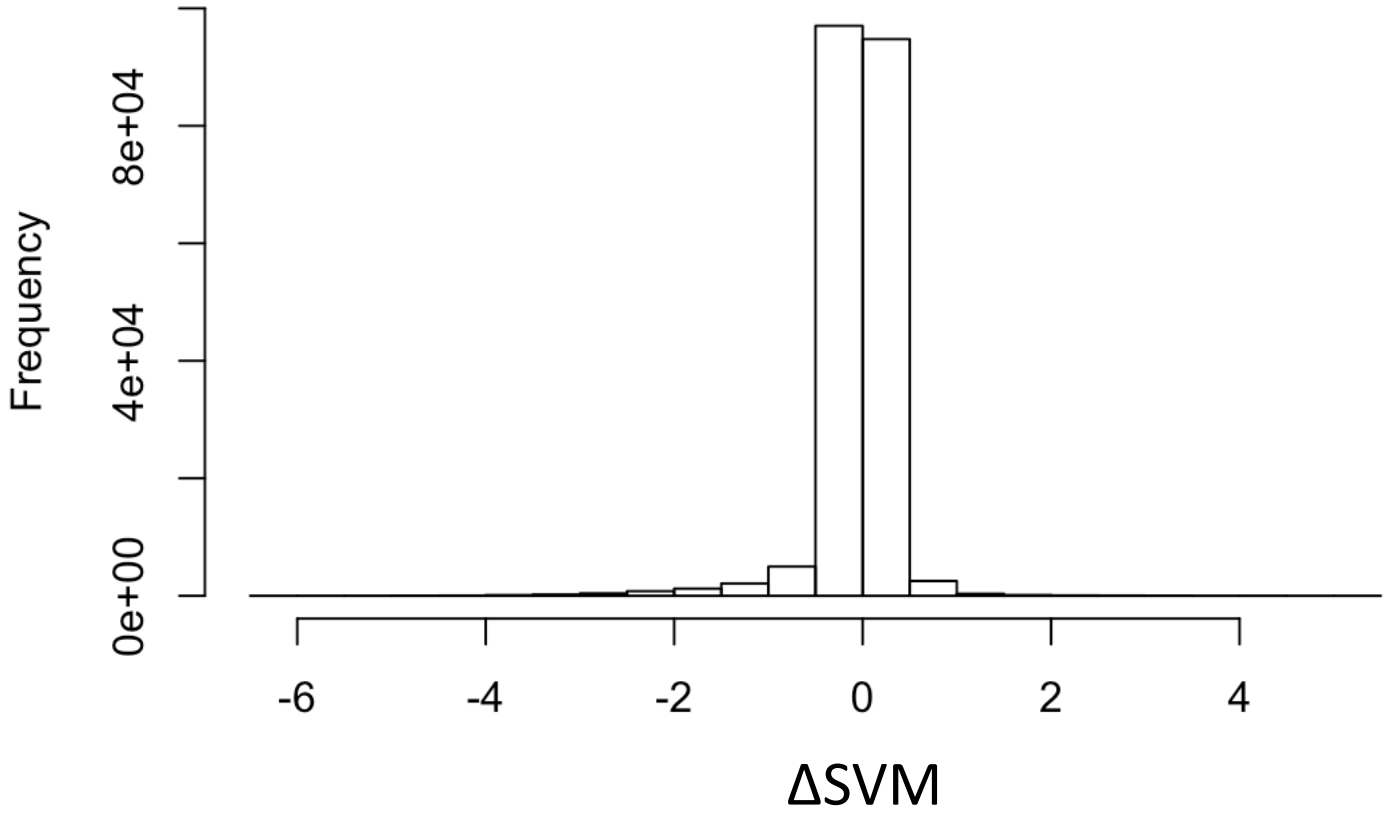
Hugo_Symbo	Entrez_Gene	Center	Chr	Position	Str	Reference_A	Tumor_Seq_	Original score	Mutated score	ΔSVM
SAMD11	148398	broad.mit.edu	1	876826	+	G	A	0.909494	0.776297	-0.133197
SKI	6497	broad.mit.edu	1	2169428	+	C	G	1.752897	1.908221	0.155324
NOL9	79707	broad.mit.edu	1	6607435	+	C	A	1.099167	1.273119	0.173952
Unknown	0	broad.mit.edu	1	11925732	+	C	T	0.683691	0.562857	-0.120834
Unknown	0	broad.mit.edu	1	14155426	+	C	T	0.492041	0.255654	-0.236387
NBL1	4681	broad.mit.edu	1	19969708	+	C	T	2.500658	2.281291	-0.219367
SNHG3-RCC1	751867	broad.mit.edu	1	28833143	+	G	A	1.645073	1.613121	-0.031952
Unknown	0	broad.mit.edu	1	31402468	+	G	A	0.409811	0.211928	-0.197883
Unknown	0	broad.mit.edu	1	32915081	+	C	T	0.241013	0.106942	-0.134071
Unknown	0	broad.mit.edu	1	48175635	+	C	T	1.506331	1.379986	-0.126345
Unknown	0	broad.mit.edu	1	59118254	+	A	C	-0.406675	-0.223868	0.182807

- Find mutations that significantly alter CTCF binding activity
- Filter MAF file again based on ΔSVM

Scoring Mutations

Large datasets

Fig. 9a Distribution of ΔSVM

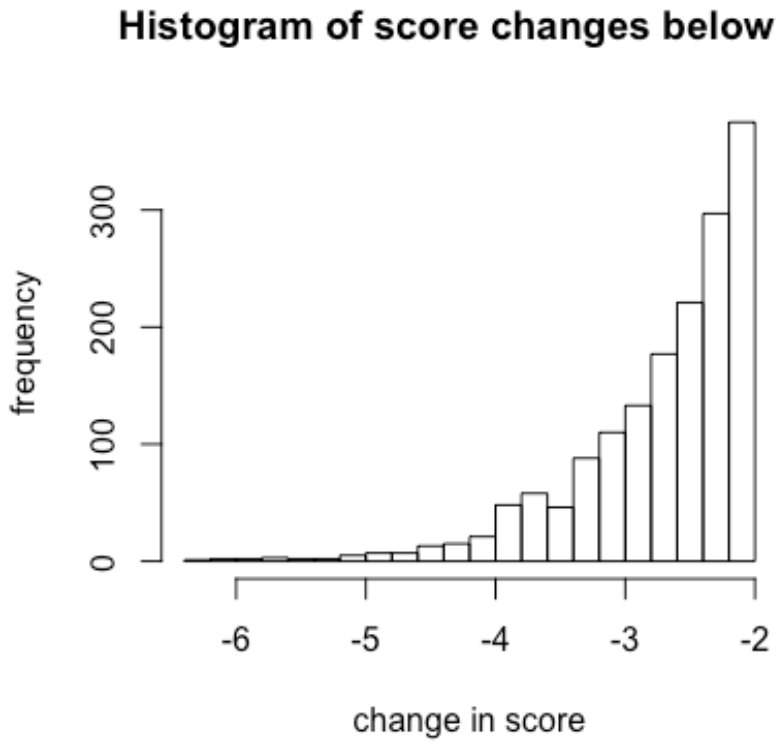


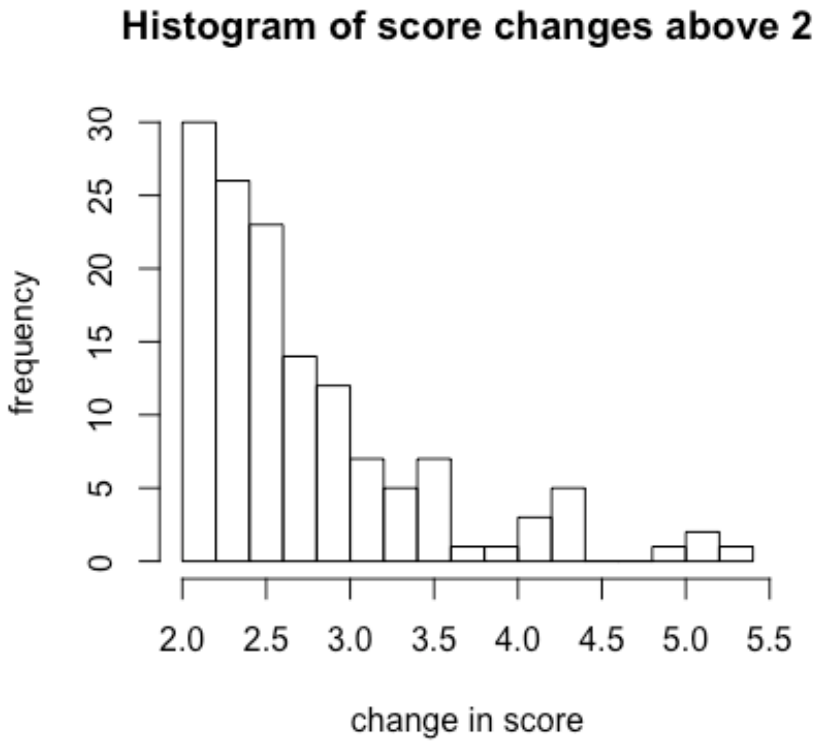
Scoring Mutations

Large datasets

Fig. 9b

Threshold value: ±2 (0.6% of all mutations)





Result: A list of all mutations that significantly alter CTCF binding activity

Monoallelic Expression

Background

- A phenomenon in which one allele is expressed while the other is not
- An indicator of mutations that alter transcriptional activity

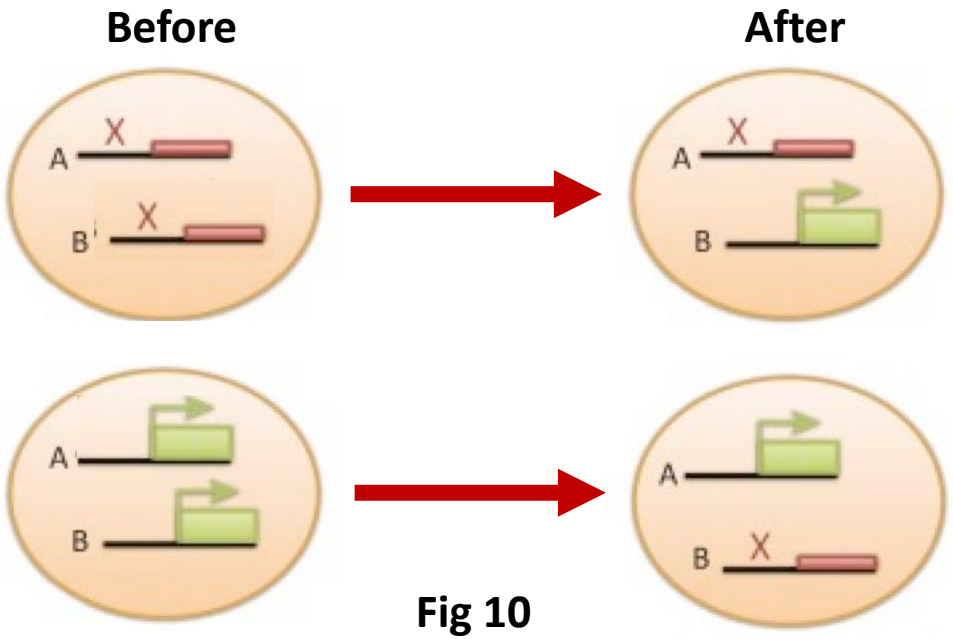


Fig 10

A mutation in one homolog may cause monoallelic expression by activating or inhibiting one copy of a nearby gene.

Monoallelic Expression

Dataset

Table 3 Sample of monoallelic expression data

Genes → . . . 15729

↓ Cell Lines

	NOC2L	KLHL17	PLEKHN1	HES4	ISG15	ATAD3B	ATAD3A
22RV1_PROSTATE	0	0	1	0	0	0	0
2313287_STOMACH	NA	0	0	NA	0	NA	NA
5637_URINARY_TRACT	NA	NA	0	NA	1	0	NA
59M_OVARY	NA	NA	NA	NA	NA	NA	0
769P_KIDNEY	NA	NA	NA	NA	NA	NA	NA
786O_KIDNEY	0	0	0	NA	0	0	1
8305C_THYROID	NA	0	0	NA	NA	1	0
A101D_SKIN	0	0	0	NA	0	1	0
A204_SOFT_TISSUE	0	0	NA	NA	NA	NA	0
A2058_SKIN	0	0	NA	NA	NA	0	0
A2780_OVARY	0	0	0	NA	0	0	0
A375_SKIN	NA	NA	0	NA	NA	0	0

. . . 329

Monoallelic Expression

Intersecting datasets

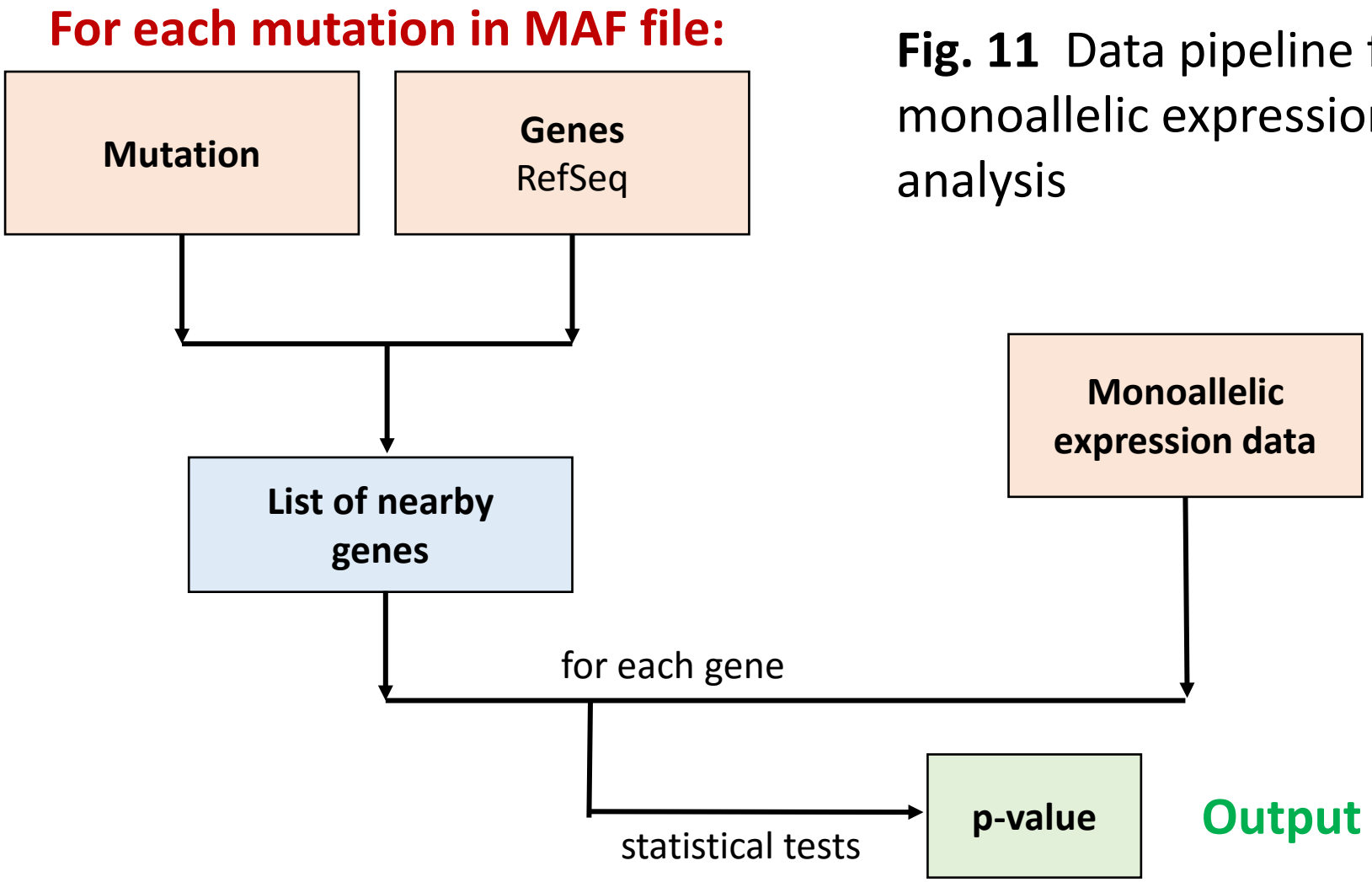


Fig. 11 Data pipeline for monoallelic expression analysis

Outline

- **Introduction**
  - Background
  - Purposes
  - Previous Work
- **Research**
  - Scoring Mutations
  - Monoallelic Expression
- **Future Work**

Future Work

- Complete analysis of **monoallelic expression**
- **Another dataset to analyze**
  - Achilles Project: catalog of tumor dependencies
- **Verifying results in the laboratory**
  - Use CRISPR Cas9 to artificially induce specific CTCF binding site mutations in a controlled experiment

Acknowledgements

I would like to thank…

- **MIT PRIMES** for making this research possible
- My mentor **Dr. Mahmoud Ghandi** for his guidance
- My parents for their support and encouragement