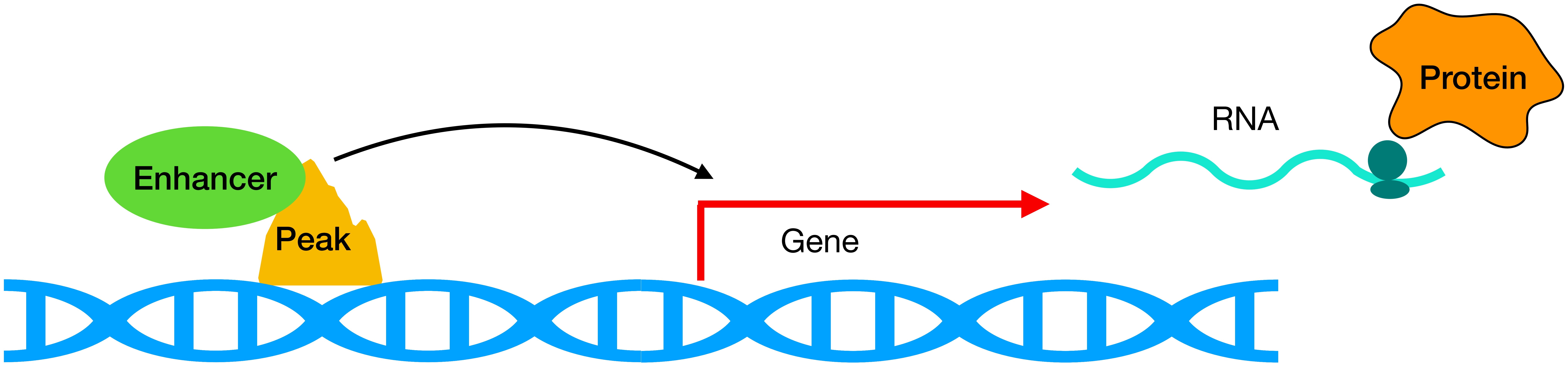
Activity-by-Contact Model to Predict Enhancer-Gene Connections

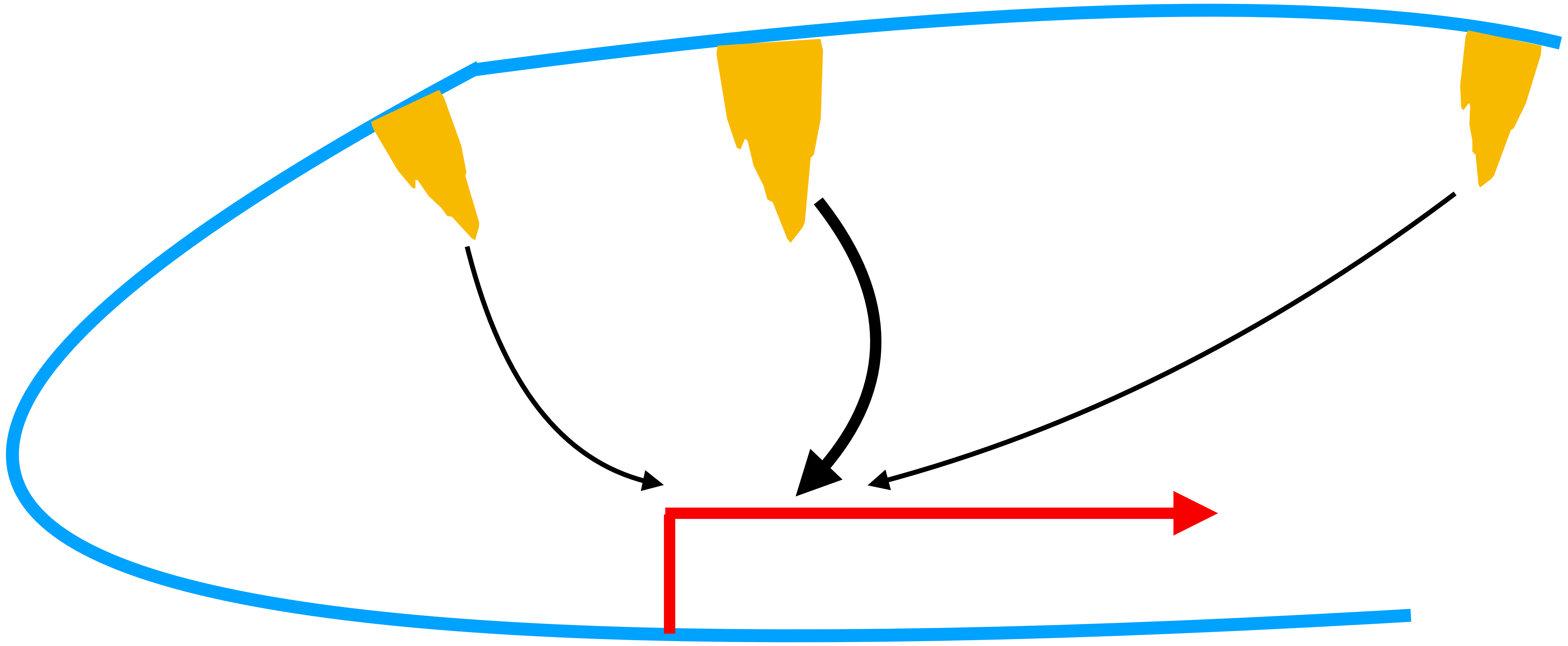
Lillian Petersen

McVicker's Lab

Salk Institute for Biological Sciences

Enhancers Control Gene Expression



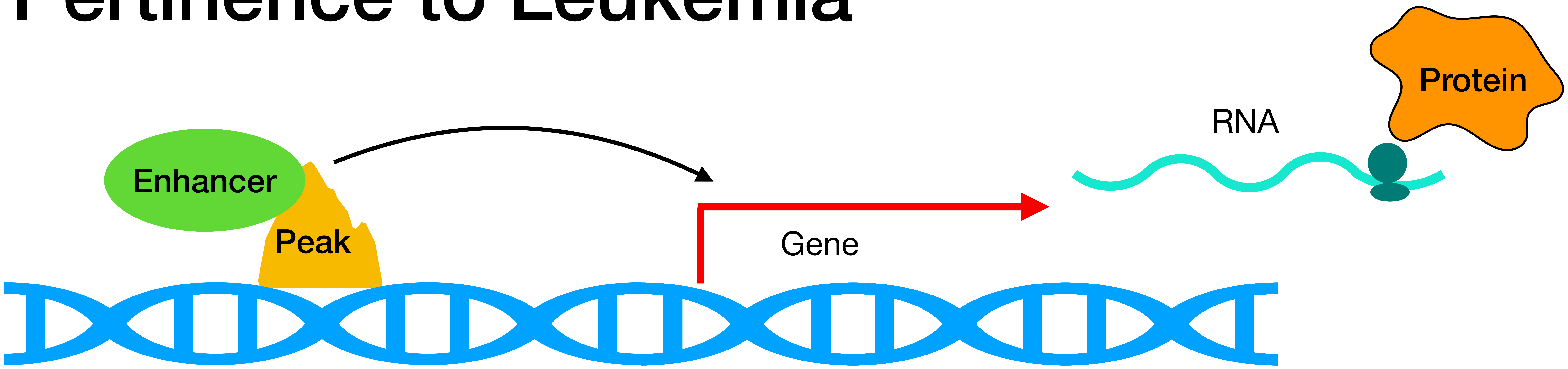


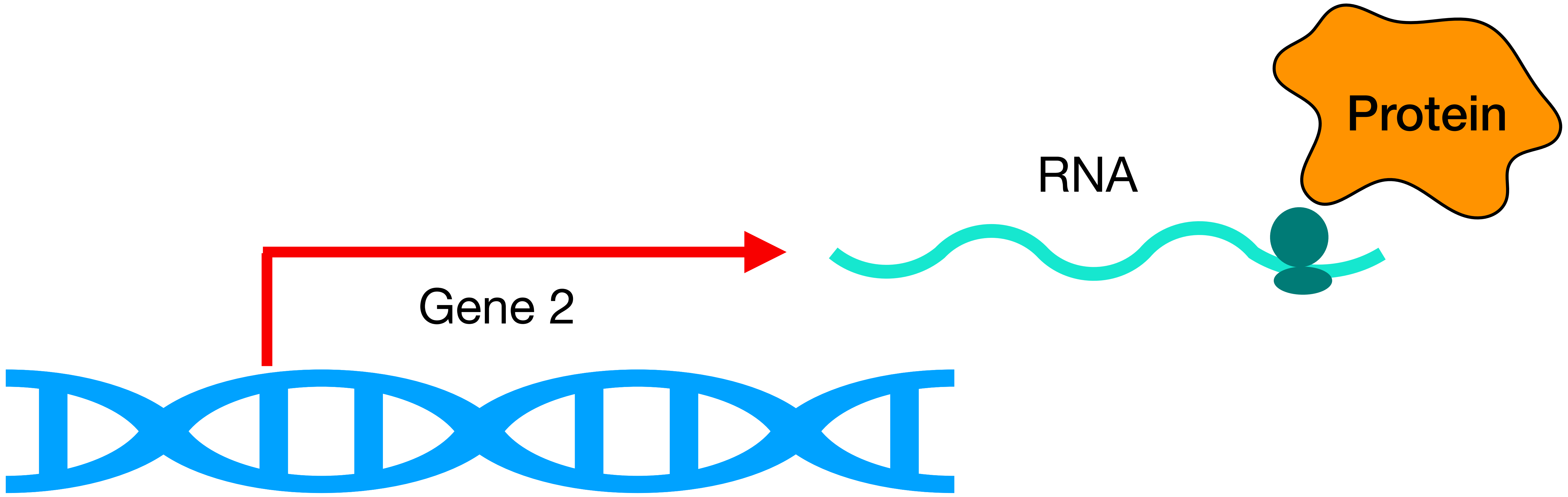
- Multiple enhancers control one gene
- An enhancer may control many genes
- Connections span large genomic distances

Theories:

- Biochemical Specificity
- 3D Architecture (topological domains)

Pertinence to Leukemia





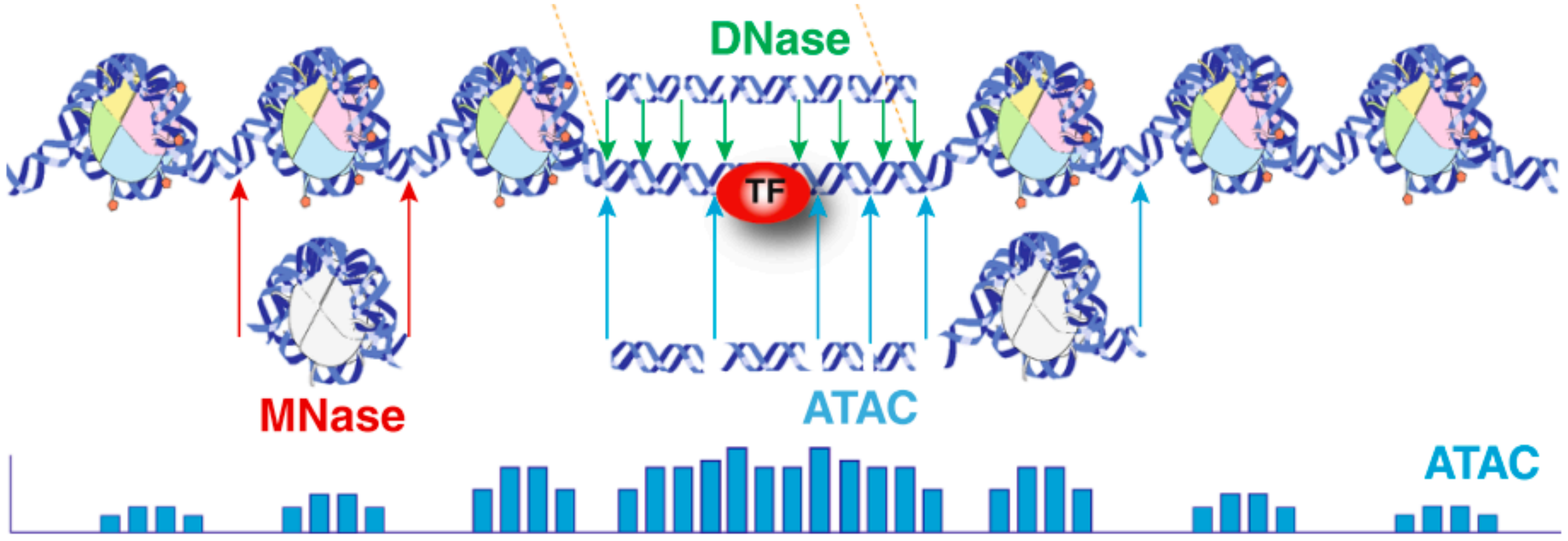
- A mutation or translocation may change the expression of a gene and lead to the creation of an oncogene

Connecting Enhancers to Genes: Importance

- Better understanding of the activation of oncogenes
- Identify transcription factor binding sites
- Possibly identify kinases for drug targets

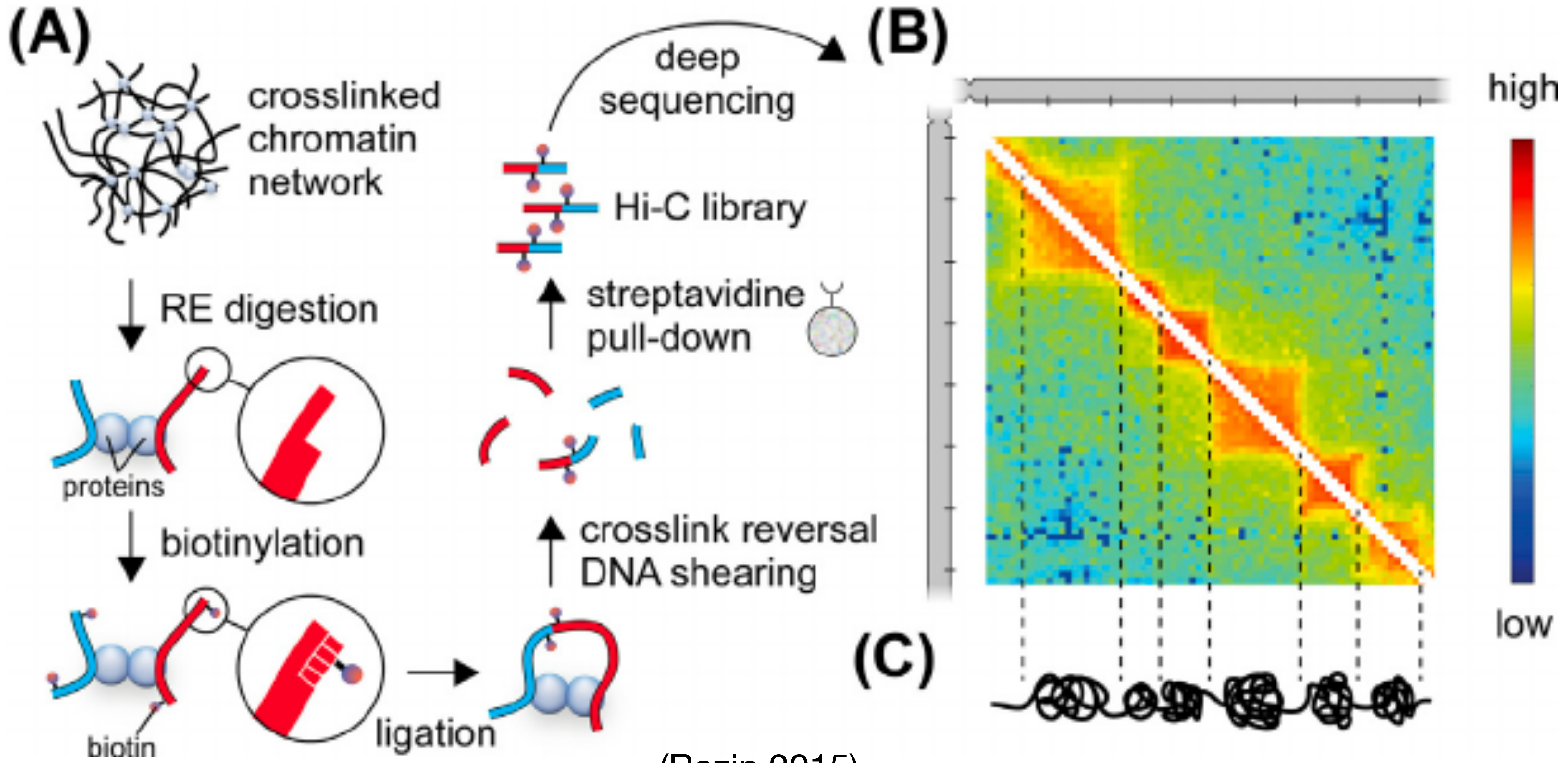
Sequencing Methods

ATAC-seq: Chromatin Accessibility



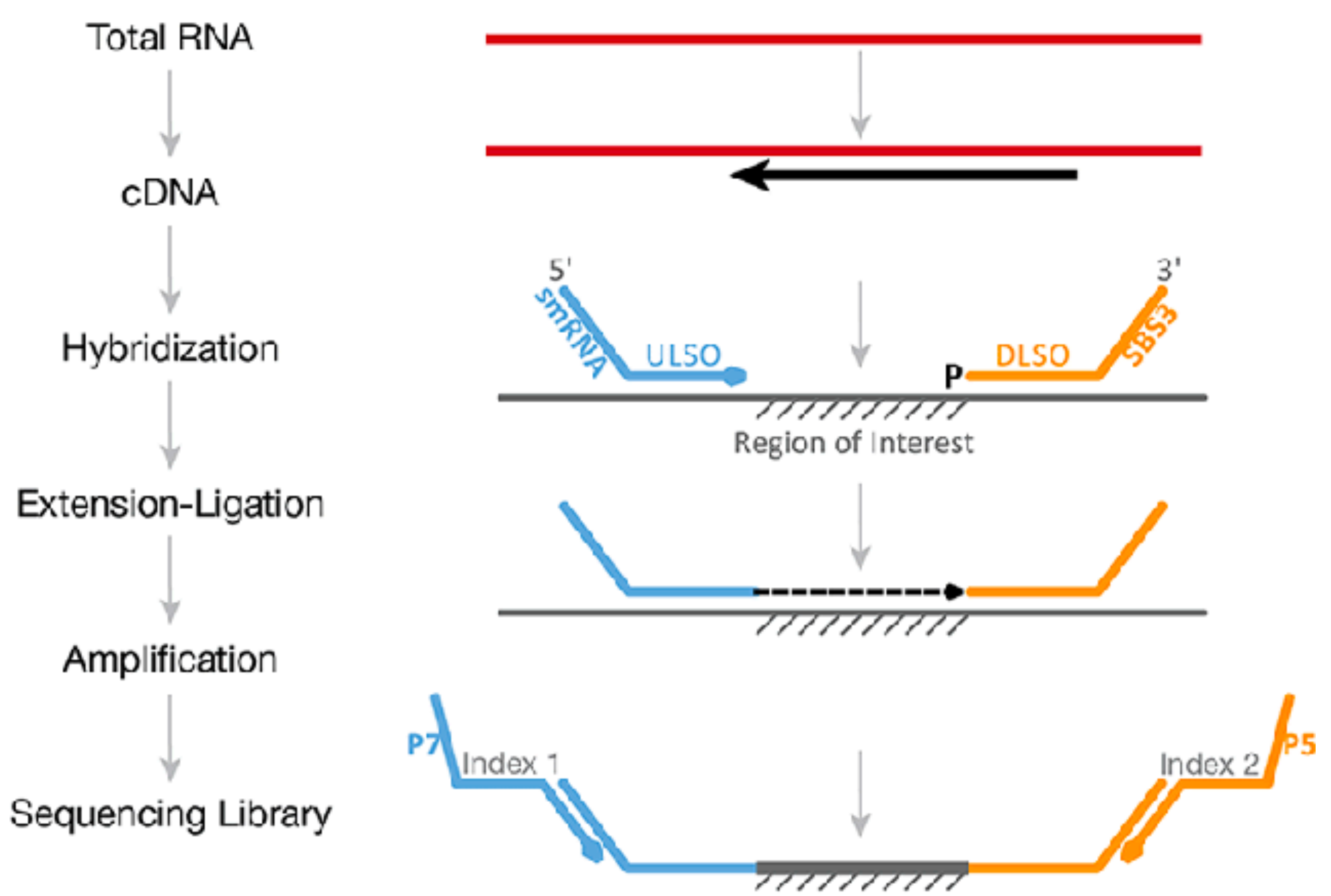
(Yiwei 2019)

HiC: Contact Frequency



(Razin 2015)

RNA-seq: Gene Expression



Equation

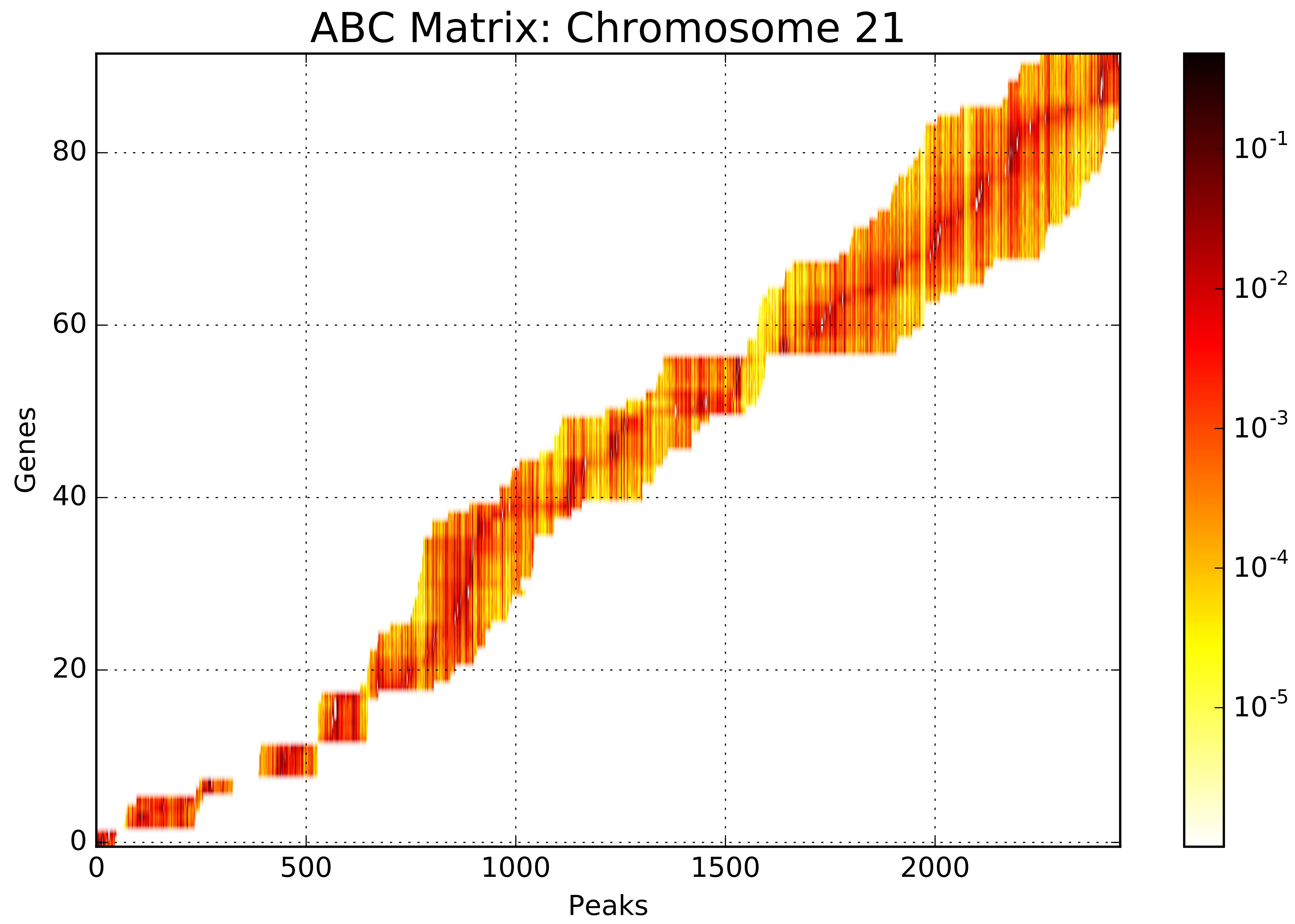
Activity-by-Contact (ABC) = Activity (ATAC) x Contact (HiC)

$$\text{ABC score}_{E\text{-}G} = \frac{A_E \times C_{E\text{-}G}}{\sum_{e \text{ within } 5\text{ Mb}} A_e \times C_{e\text{-}G}}$$

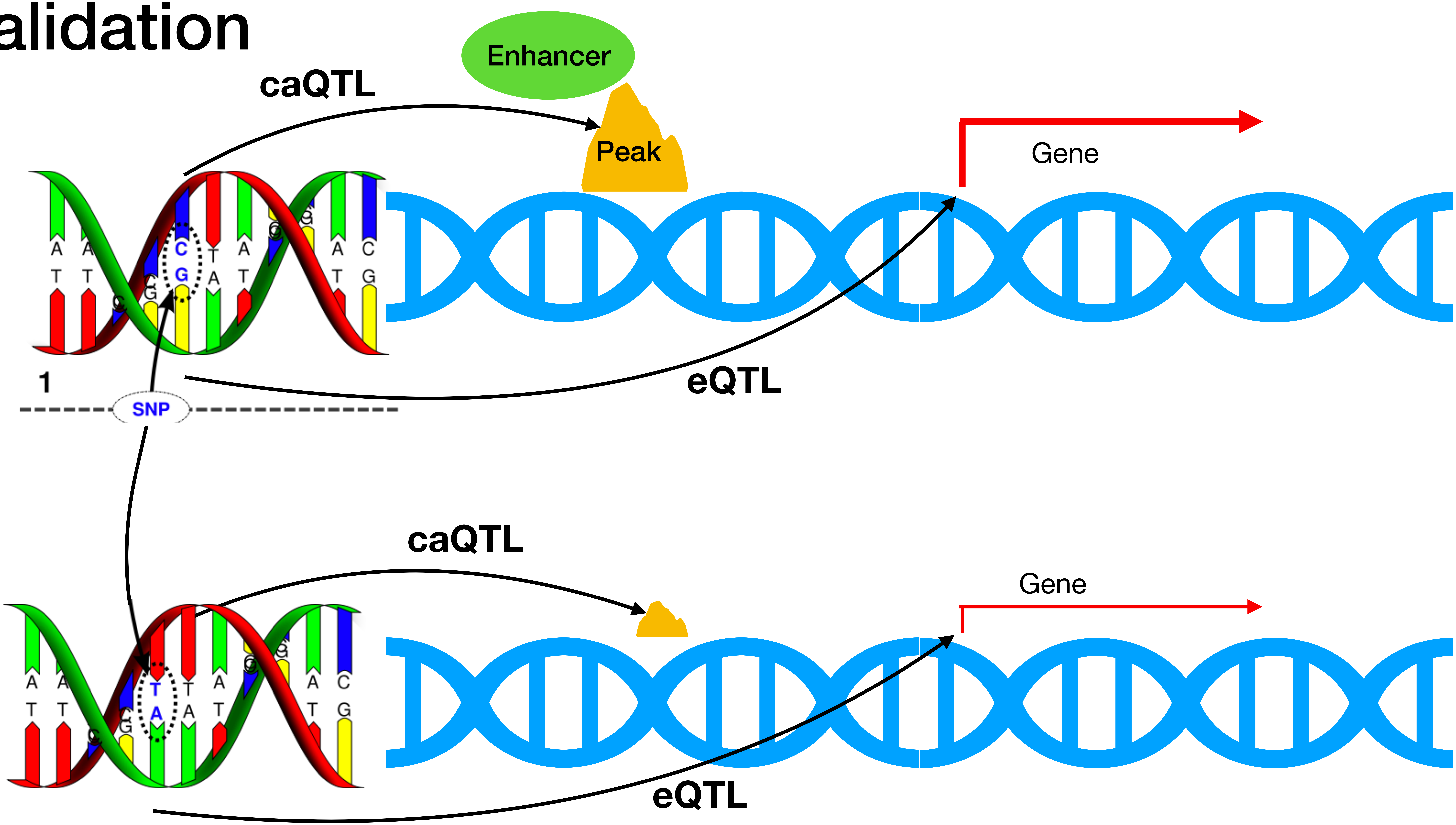
- Identify peaks that are likely regulatory elements for specific genes

Data and Computation

- 16 Leukemia B ALL samples with ATAC-seq, HiC, and RNA-seq
  - HiC matrices: merged to a single, high-resolution matrix
  - 10,000 protein-coding genes and 120,000 peaks
- Computed ABC score for every possible peak-gene connection within 1.5 Mb of TSS
- Computed other indices to predict connections
  - Distance from peak—gene
  - Correlation between peak intensity (ATAC) and gene expression

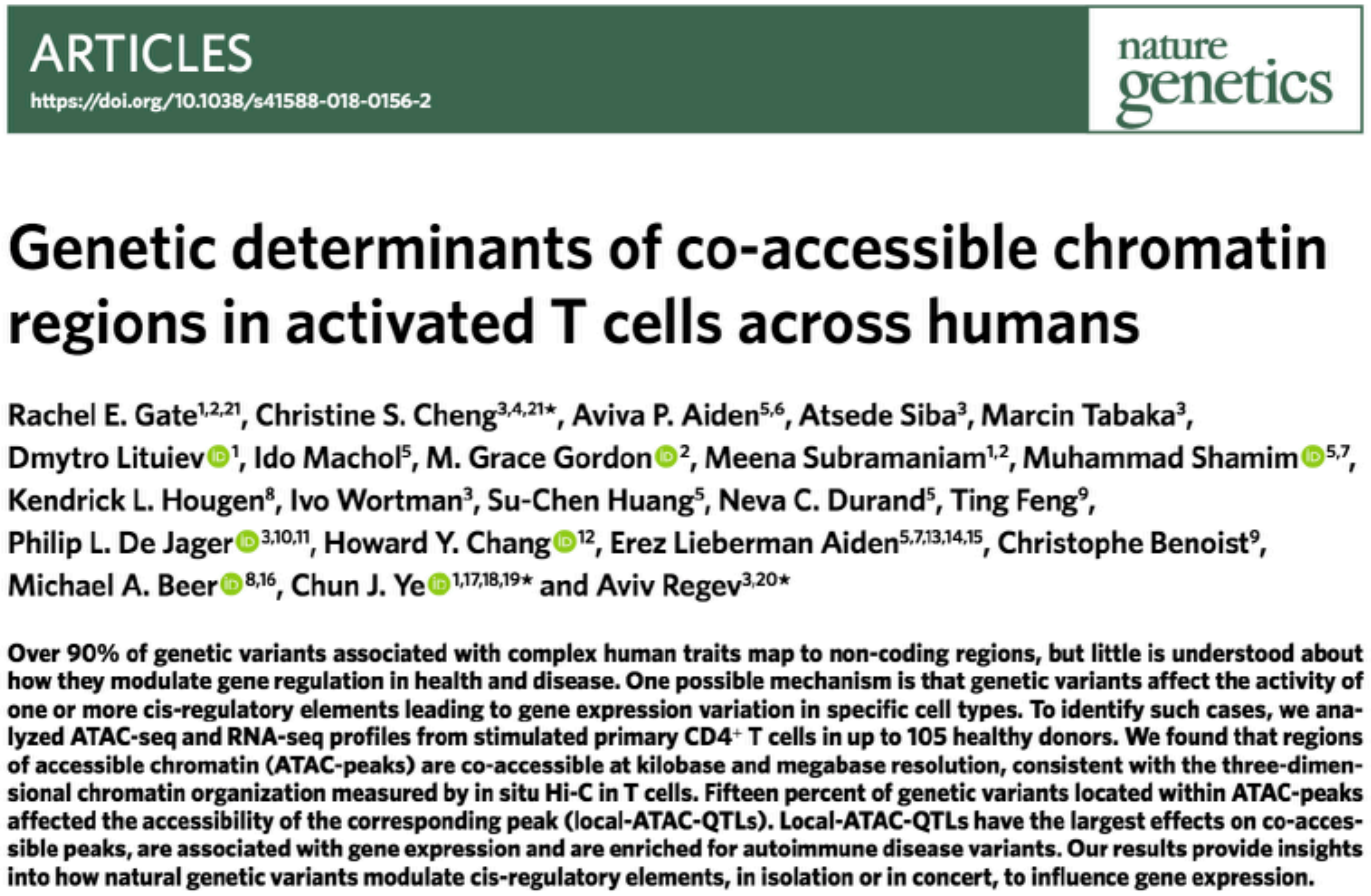


Validation


Enhancer
caQTL
Peak
Gene
A
A
C
T
A
A
C
T
T
G
A
T
T
G
G
G
G
G
1
eQTL
SNP
caQTL
Gene
A
A
T
T
A
A
C
T
T
A
A
T
T
G
G
G
G
eQTL

caQTL Publication

- Gate et al. (2018) identified caQTLs and eQTLs in T cells
- We can use this data to gain a list of known connections to examine the efficacy of the ABC model



ARTICLES

https://doi.org/10.1038/s41588-018-0156-2

nature genetics

Genetic determinants of co-accessible chromatin regions in activated T cells across humans

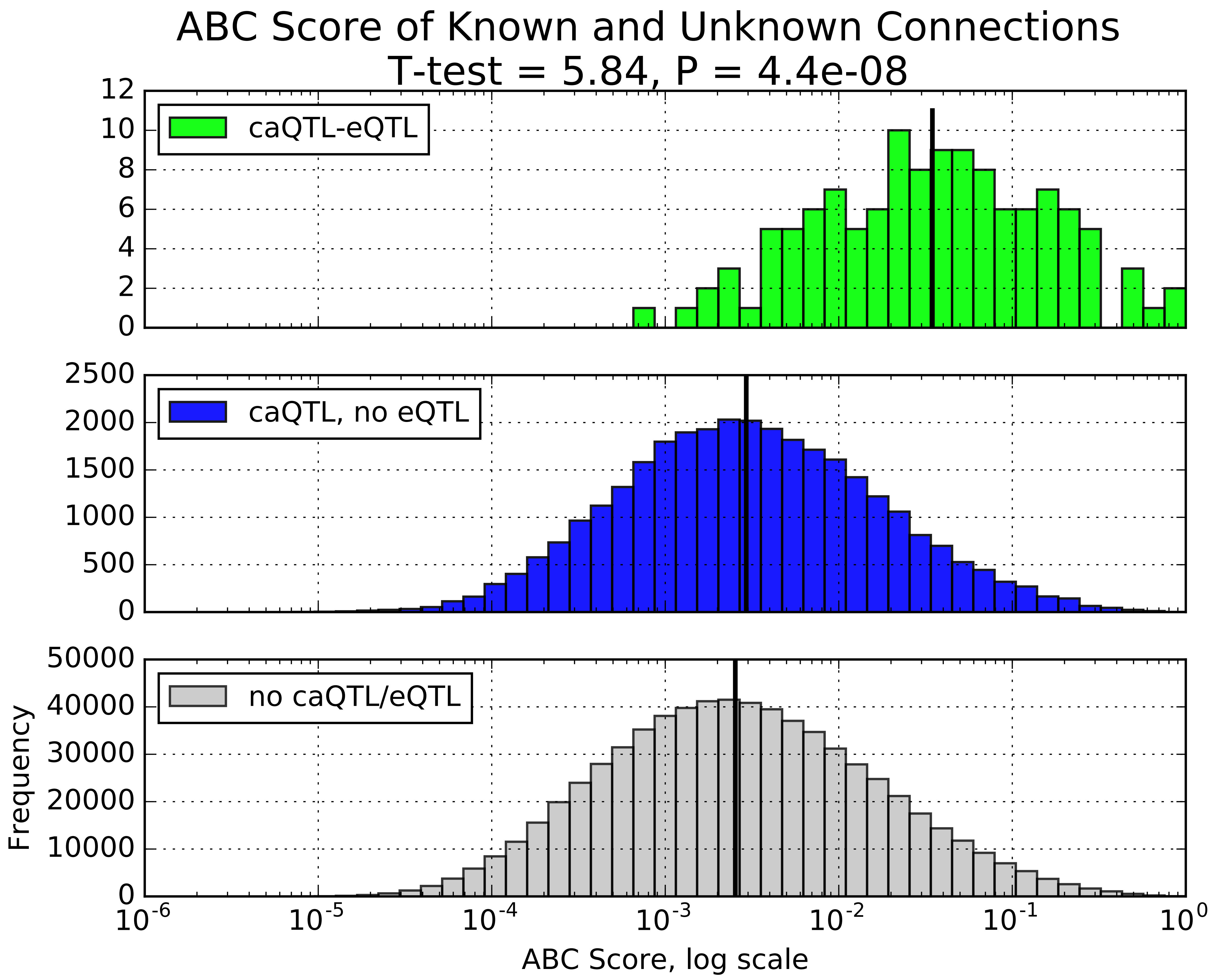
Rachel E. Gate[1,2,21], Christine S. Cheng[3,4,21*], Aviva P. Aiden[5,6], Atsede Siba[3], Marcin Tabaka[3], Dmytro Lituiev[1], Ido Machol[5], M. Grace Gordon[2], Meena Subramaniam[1,2], Muhammad Shamim[5,7], Kendrick L. Hougen[8], Ivo Wortman[3], Su-Chen Huang[5], Neva C. Durand[5], Ting Feng[9], Philip L. De Jager[3,10,11], Howard Y. Chang[12], Erez Lieberman Aiden[5,7,13,14,15], Christophe Benoist[9], Michael A. Beer[8,16], Chun J. Ye[1,17,18,19*] and Aviv Regev[3,20*]

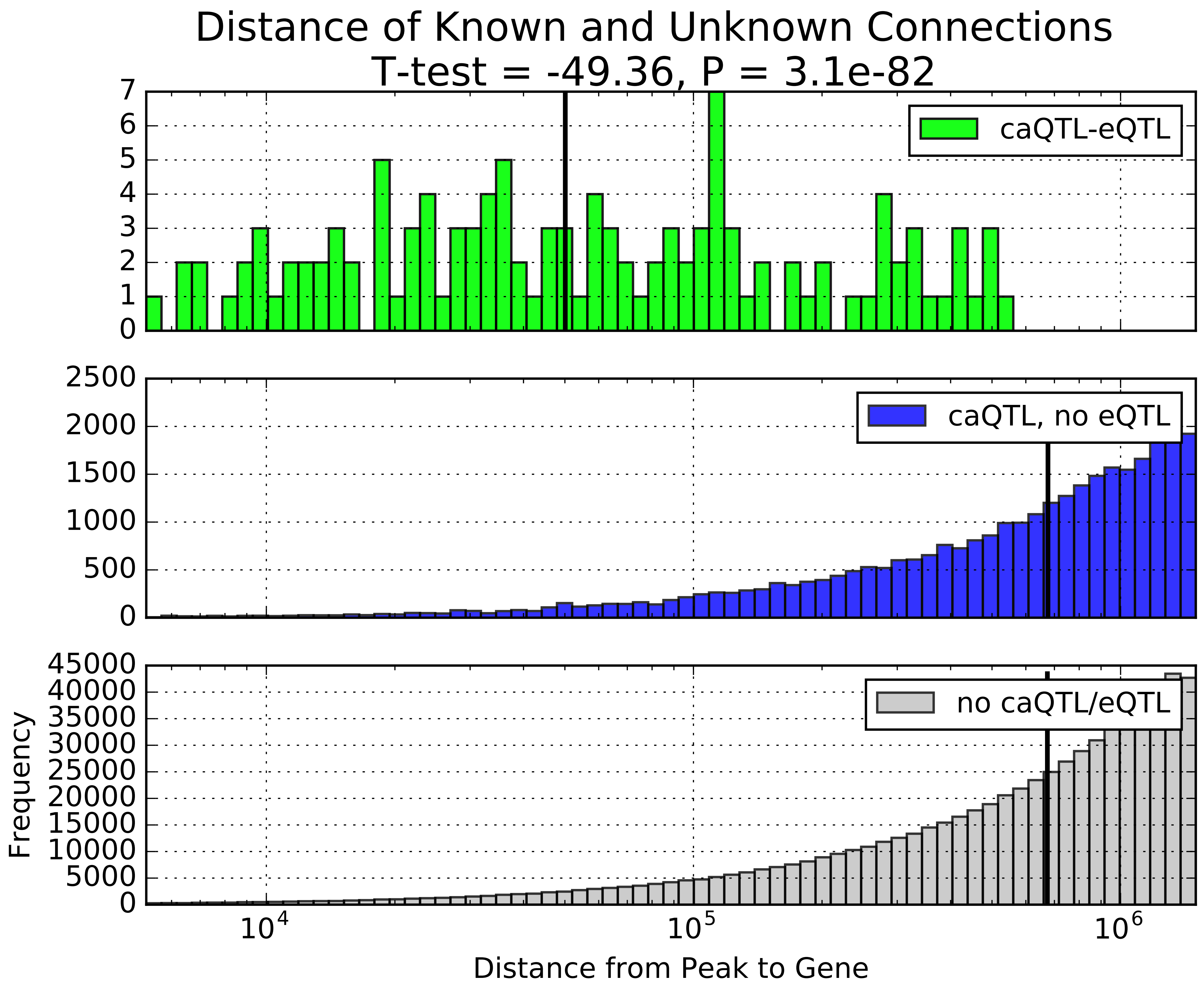
Over 90% of genetic variants associated with complex human traits map to non-coding regions, but little is understood about how they modulate gene regulation in health and disease. One possible mechanism is that genetic variants affect the activity of one or more cis-regulatory elements leading to gene expression variation in specific cell types. To identify such cases, we analyzed ATAC-seq and RNA-seq profiles from stimulated primary CD4$^+$ T cells in up to 105 healthy donors. We found that regions of accessible chromatin (ATAC-peaks) are co-accessible at kilobase and megabase resolution, consistent with the three-dimensional chromatin organization measured by in situ Hi-C in T cells. Fifteen percent of genetic variants located within ATAC-peaks affected the accessibility of the corresponding peak (local-ATAC-QTLs). Local-ATAC-QTLs have the largest effects on co-accessible peaks, are associated with gene expression and are enriched for autoimmune disease variants. Our results provide insights into how natural genetic variants modulate cis-regulatory elements, in isolation or in concert, to influence gene expression.

Validation: Process

- Downloaded ATACseq, RNAseq data from GEO for 95 individuals in their study
- Our merged HiC matrix
- Identified peak-gene connections from caQTLs and eQTLs
    - 130 connections in total
- Computed ABC score, correlation between ATAC peaks and expression, distance from peak to tss

Distributions by ABC Score and Distance

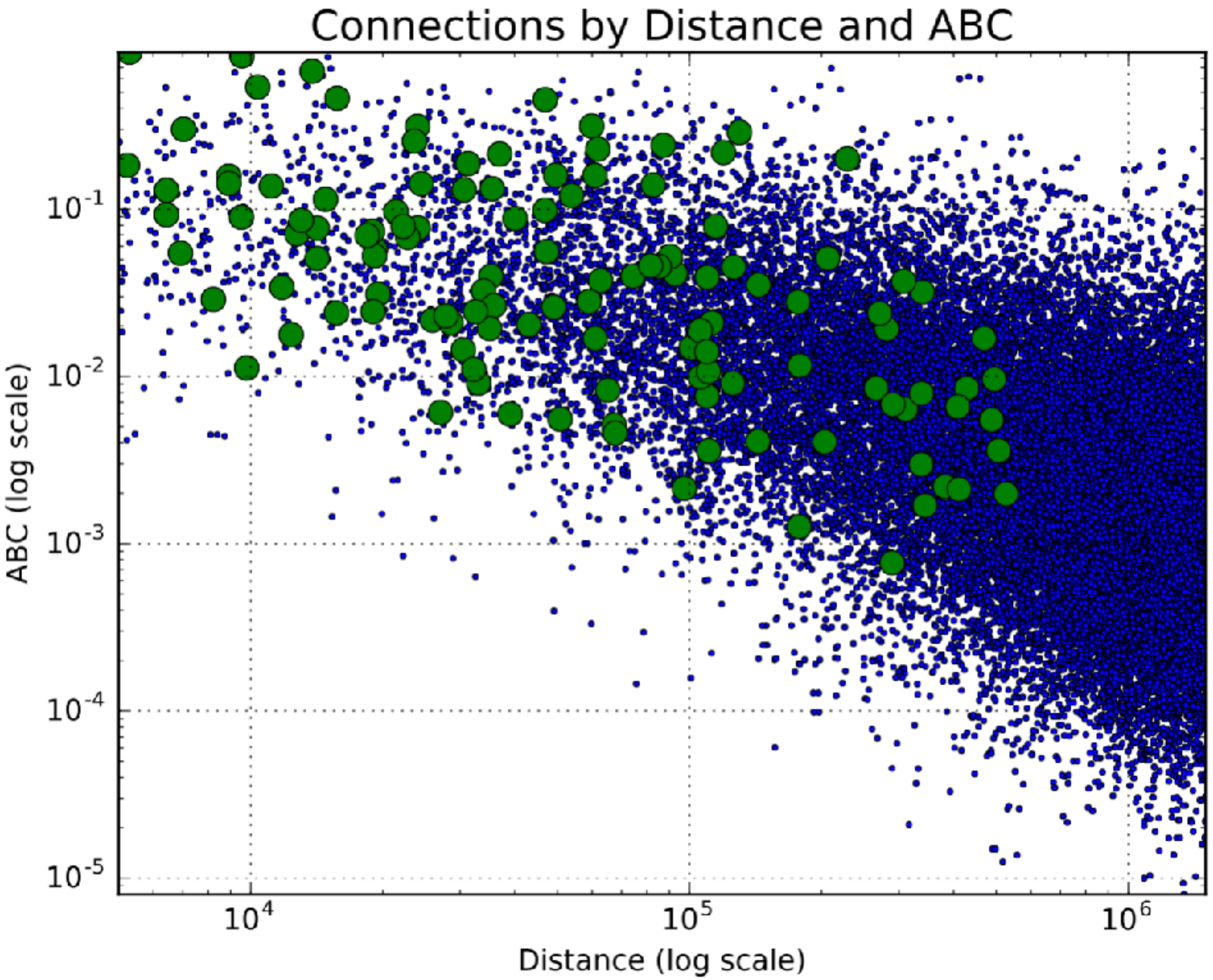

ABC Score of Known and Unknown Connections
T-test = 5.84, P = 4.4e-08


Distance of Known and Unknown Connections
T-test = -49.36, P = 3.1e-82

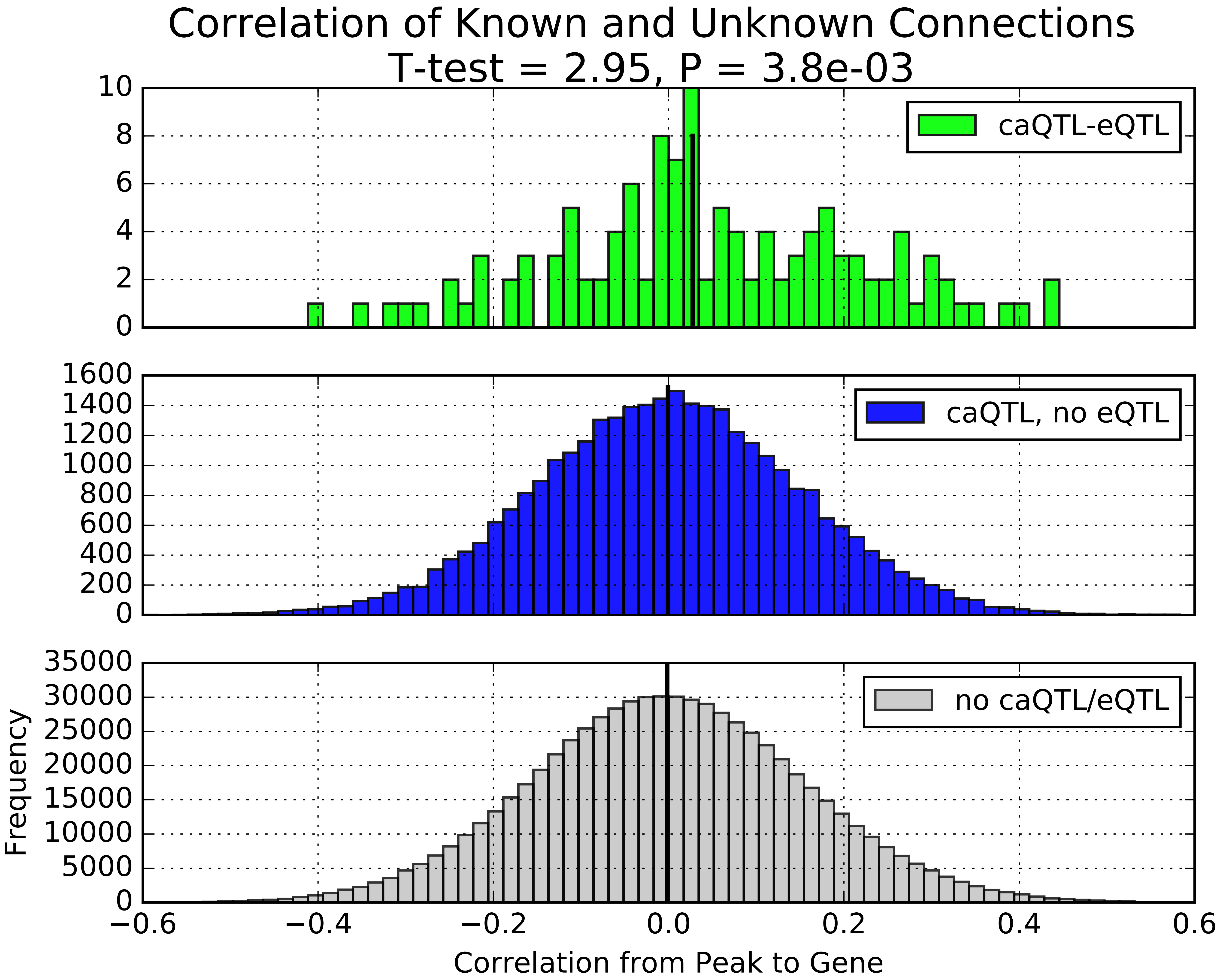
Limitations of the Validation Dataset

- No *known negatives*
- Cannot detect connections with peaks or genes unaffected by SNPs
- caQTL-eQTL may be independent
- We don’t have access to their entire dataset
- They likely cannot detect connections with small expression / peak intensity / low connectivity

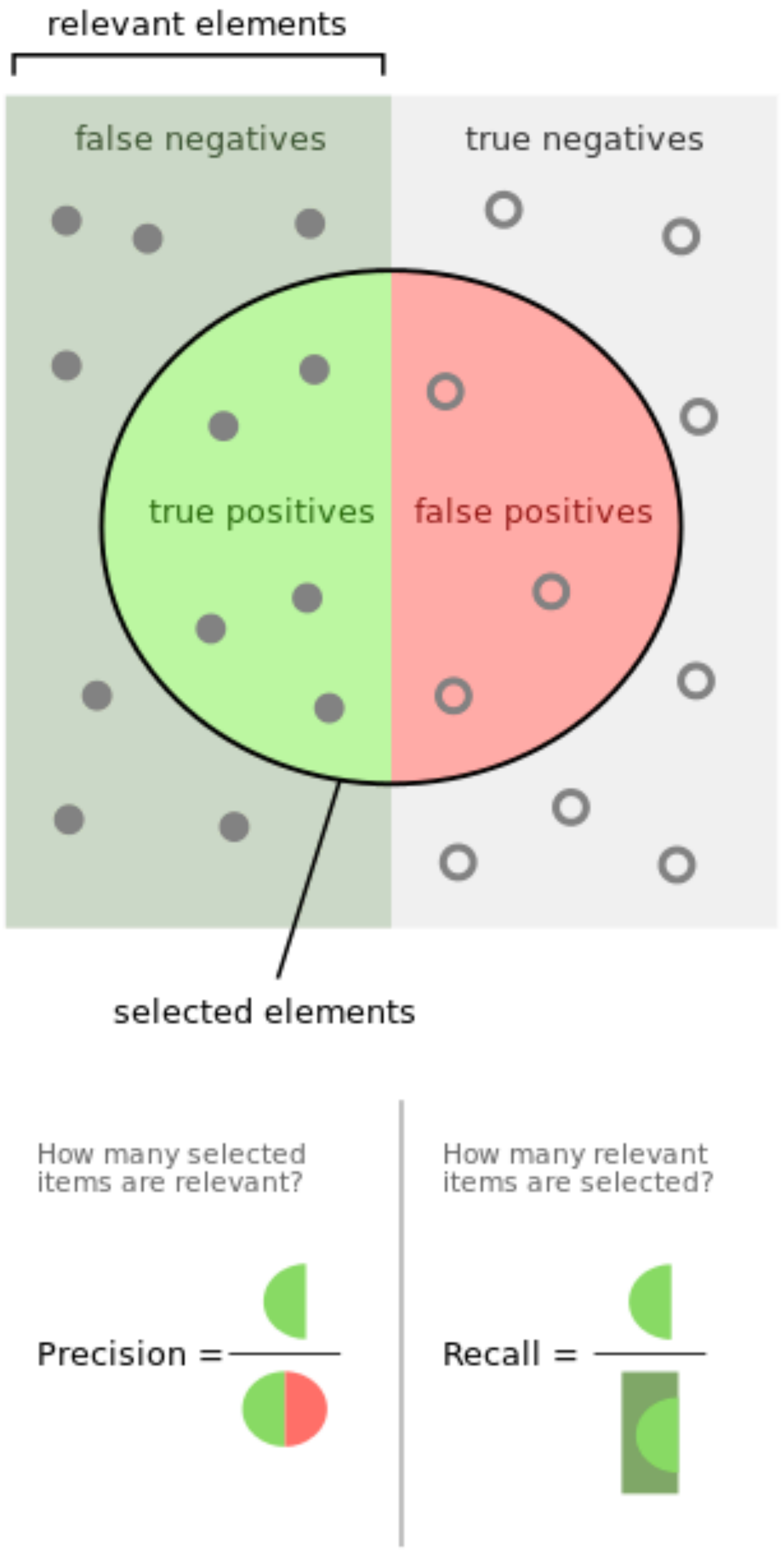
ABC and Distance Log-Log Plot

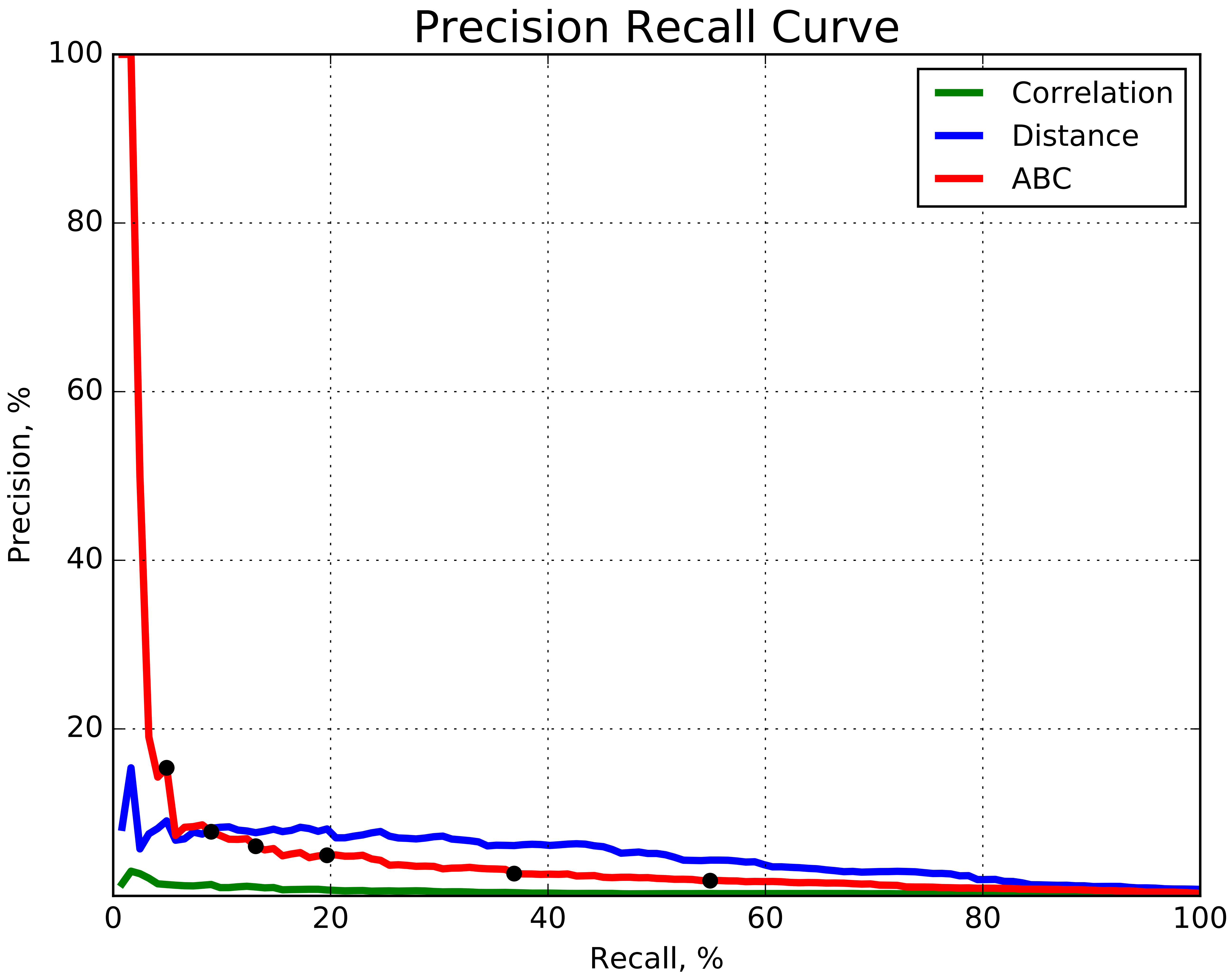


Distributions by Correlation

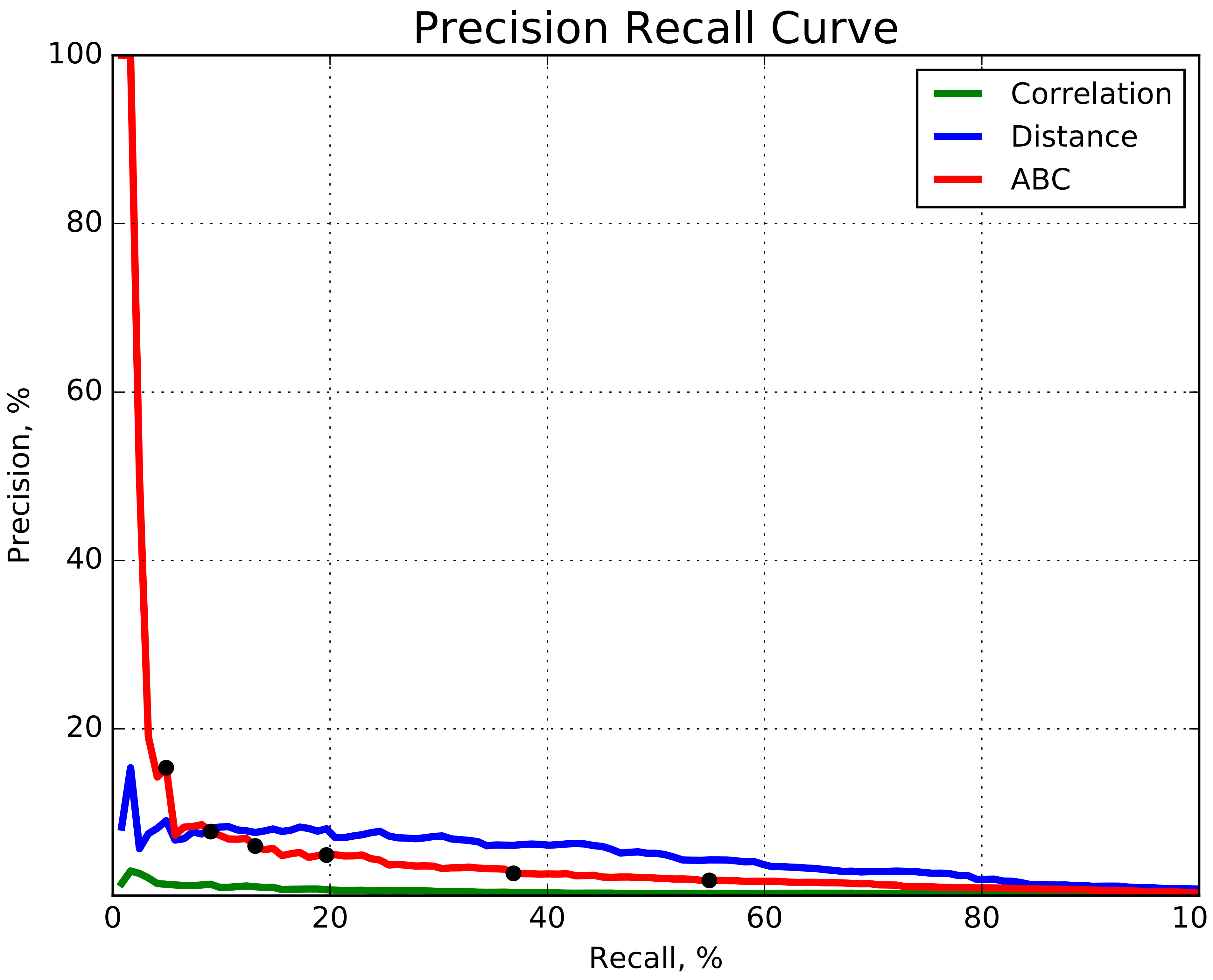


Precision Recall Curve





Applying Validation To Leukemia Data

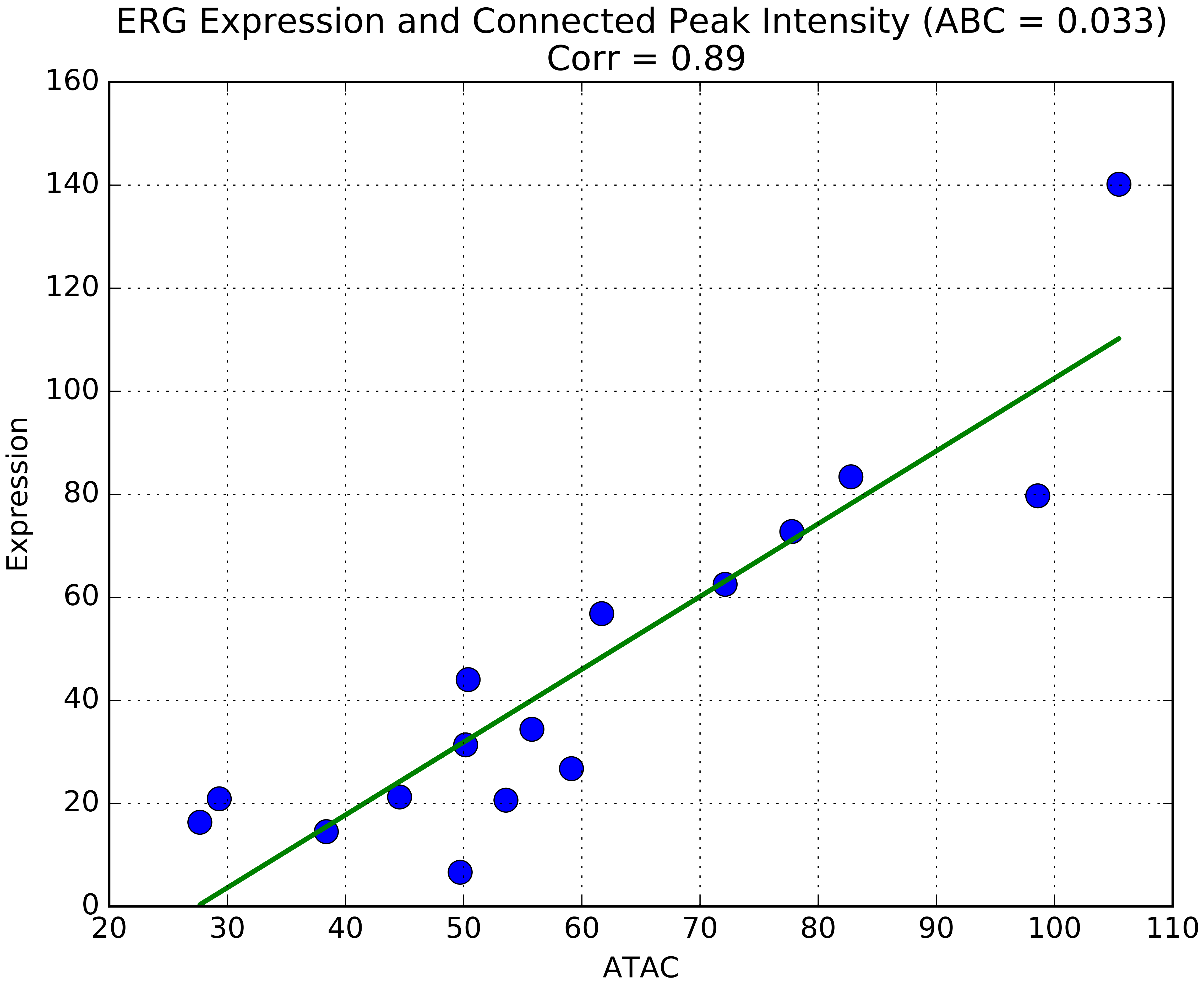


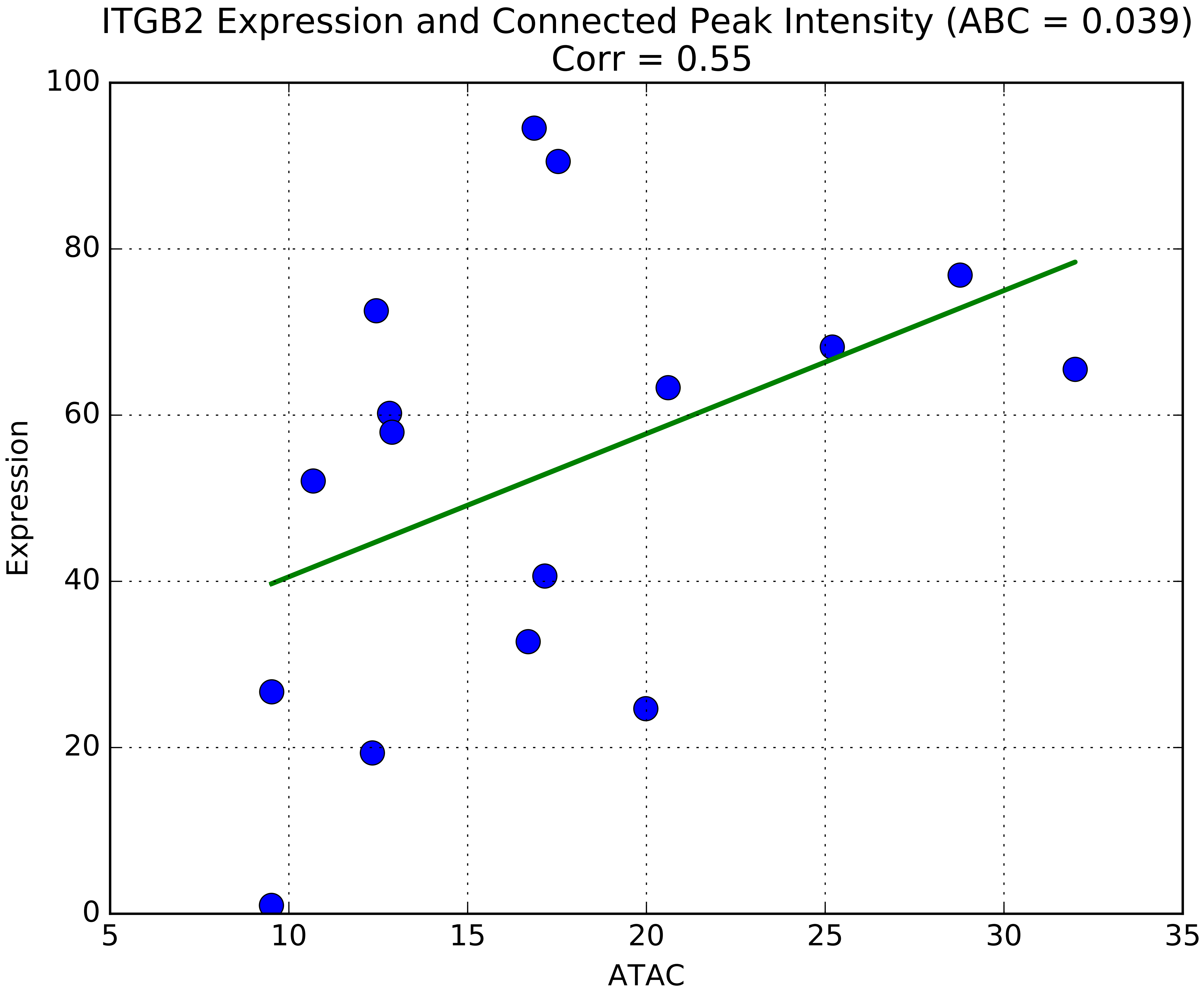
ABC	0.45	0.25	0.2	0.14	0.06	0.03
Precision	15%	8%	6%	5%	3%	2%
Recall	5%	9%	13%	20%	37%	55%

Leukemia Data: Process

- Calculated ABC Score for every possible peak-gene connection (within 1.5Mb)
- Also computed computed:
    - Distance
    - Correlation
    - Slope
- Identified peak-gene connections by ABC cutoffs identified in validation (6 groups)

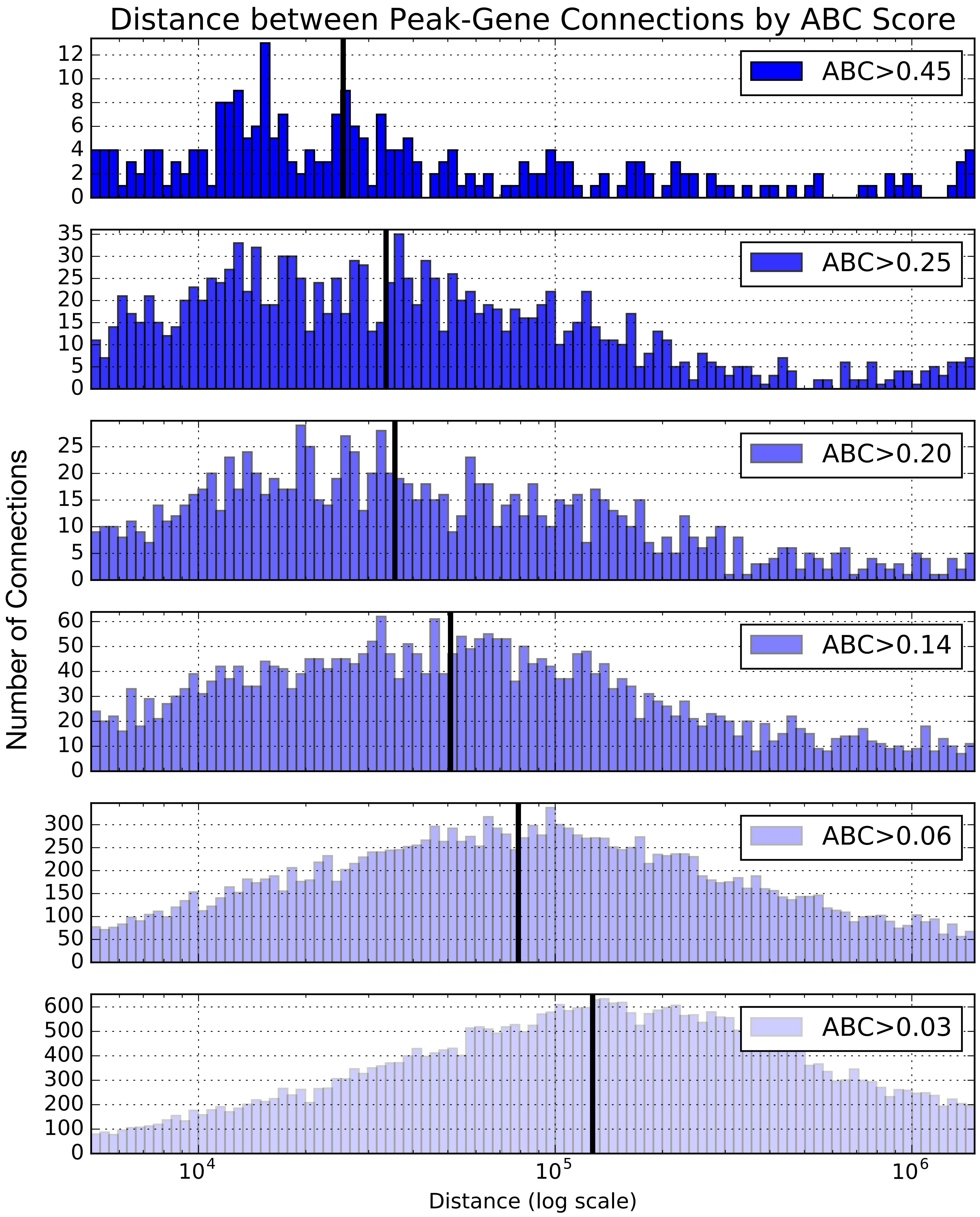
Genes in Leukemia Dataset

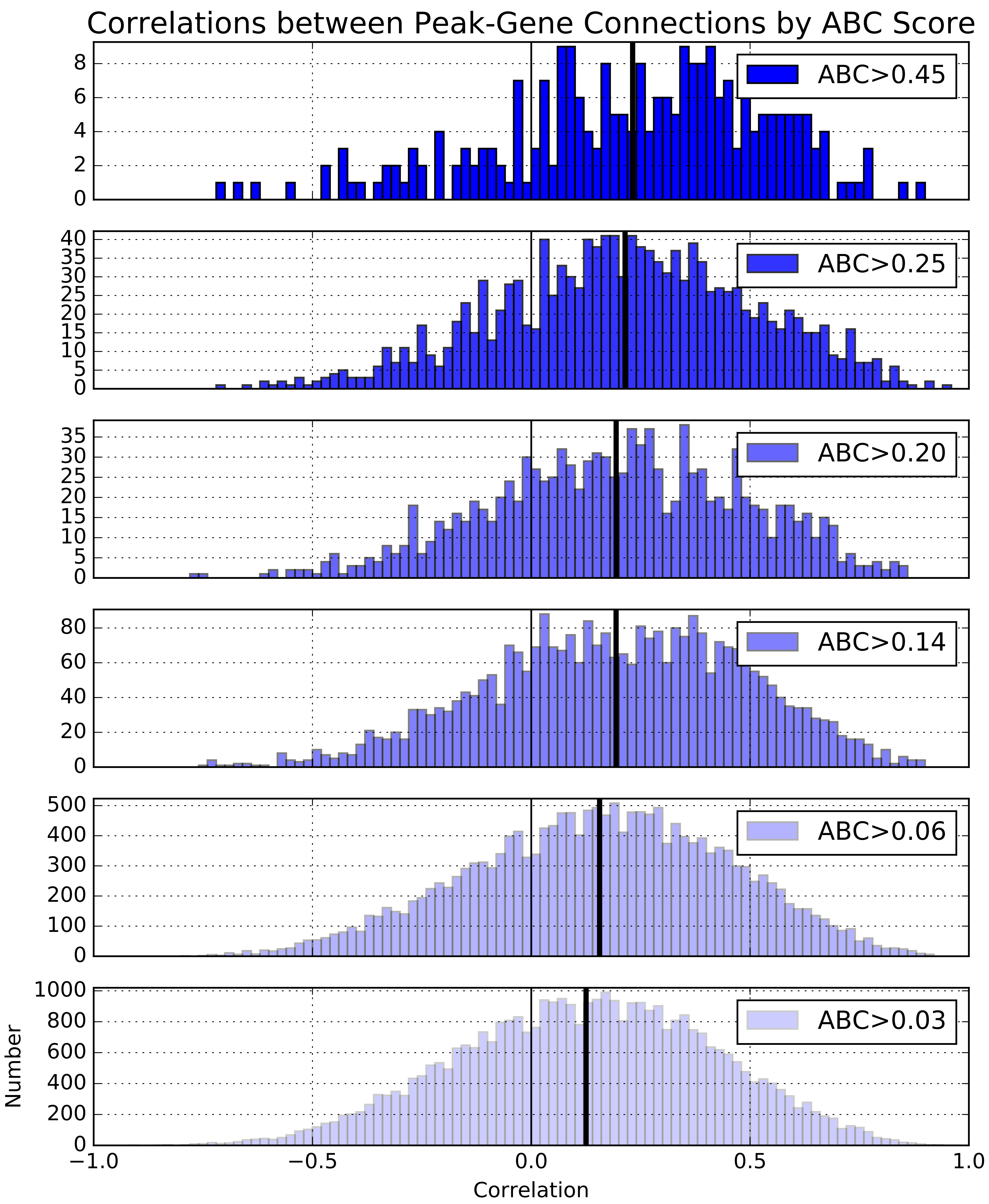


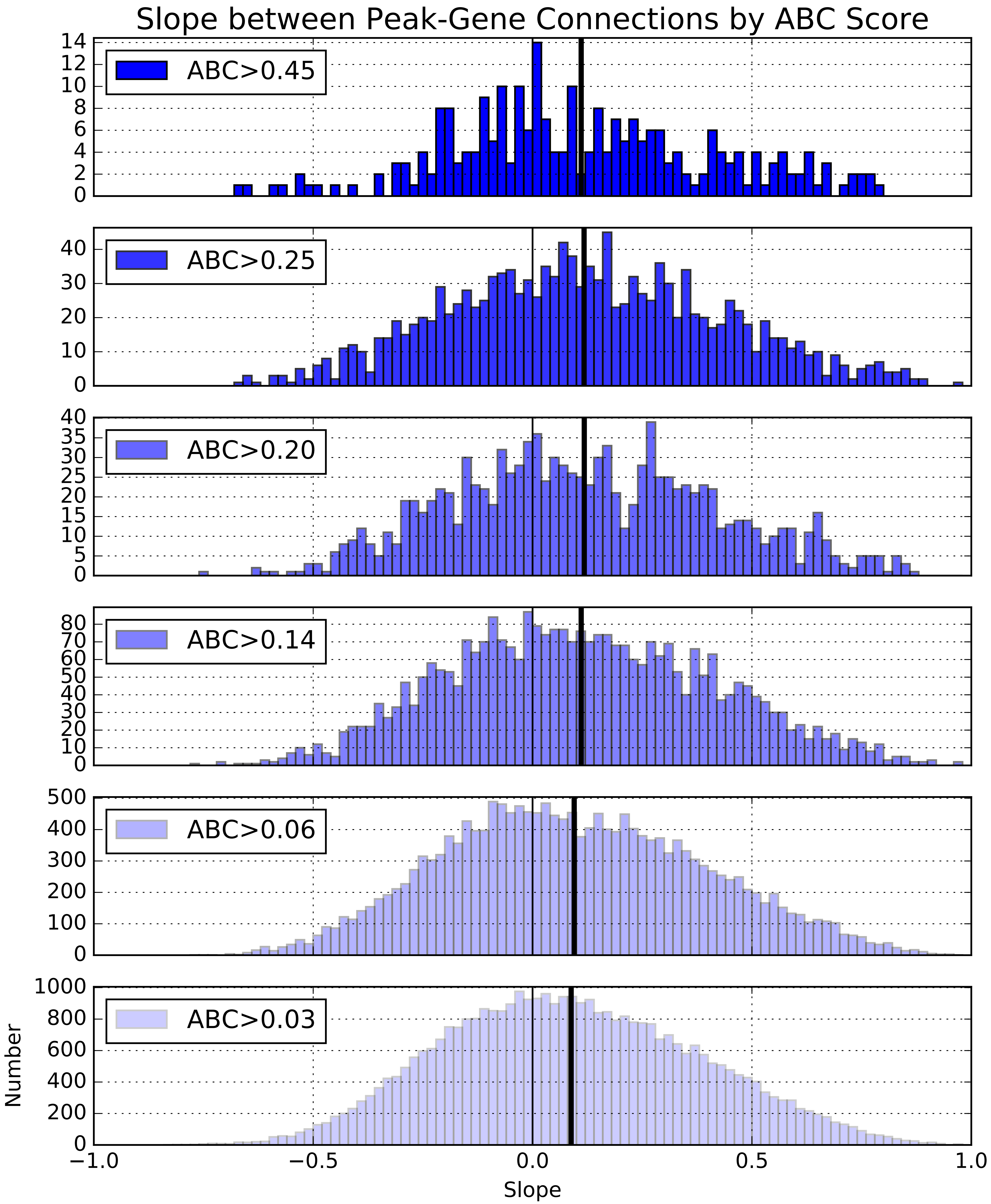


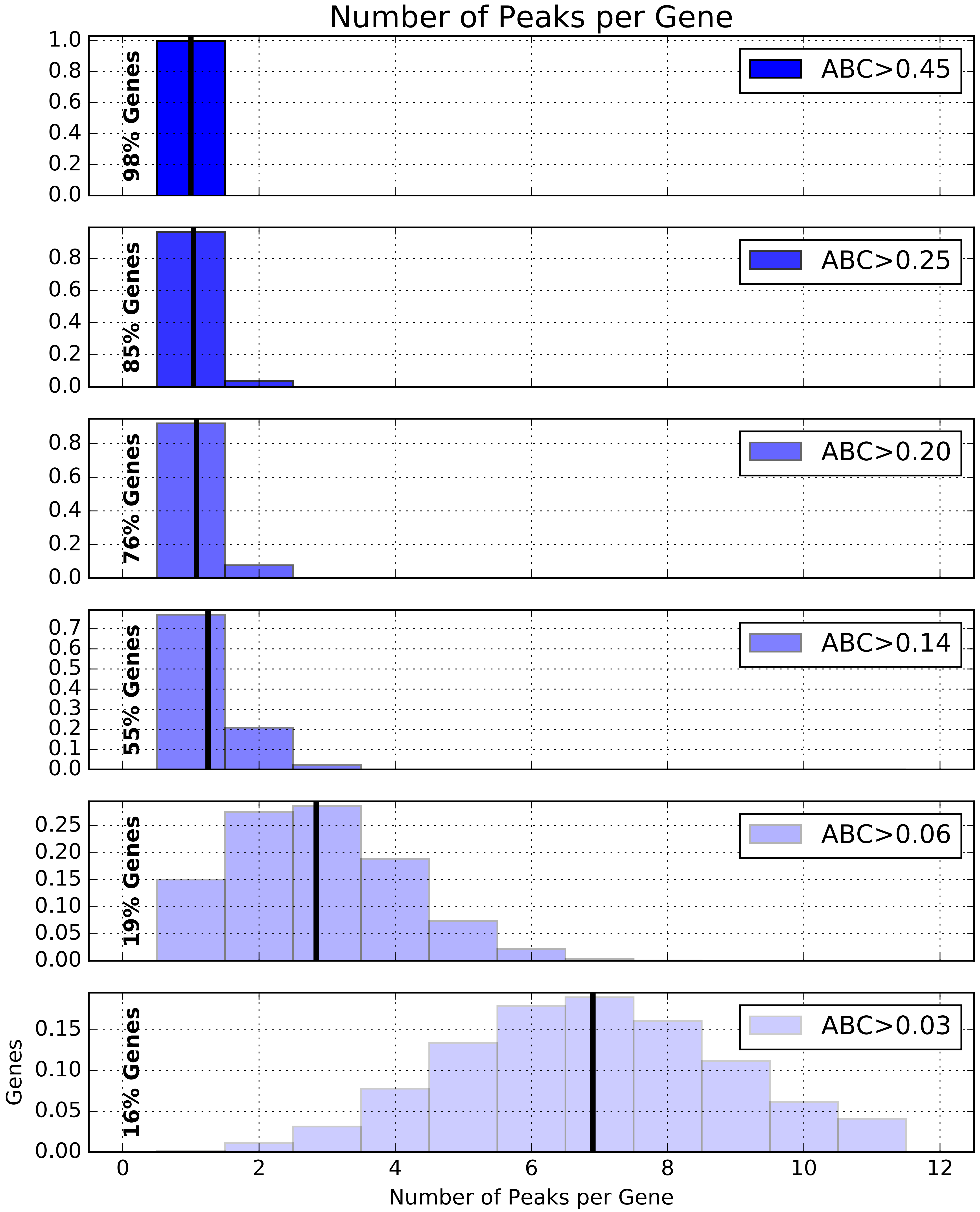
Leukemia Data

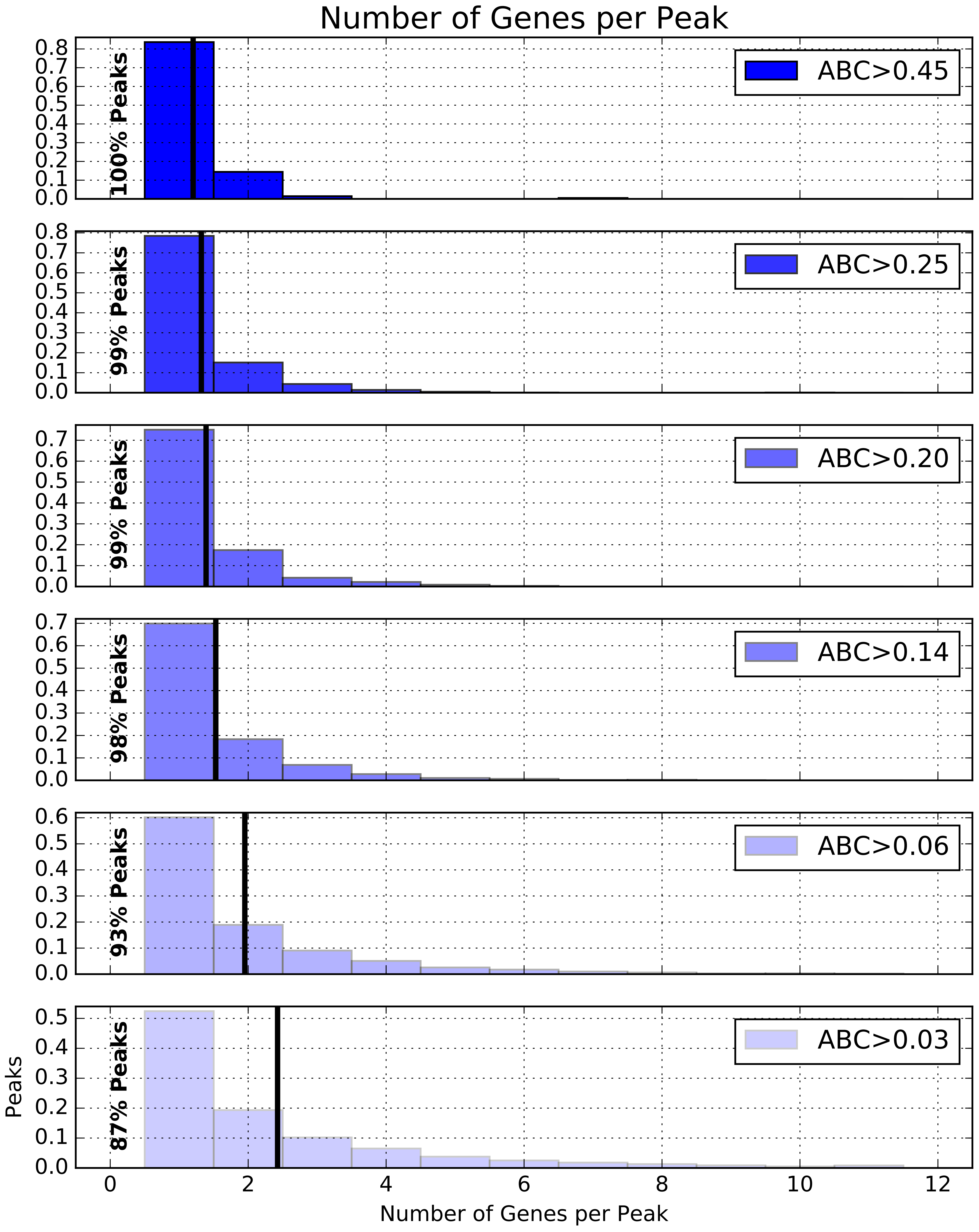
- Trends by ABC score:
  - Distance
  - Correlation
  - Slope
  - Peaks per gene
  - Genes per peak




Correlations between Peak-Gene Connections by ABC Score
ABC>0.45
ABC>0.25
ABC>0.20
ABC>0.14
ABC>0.06
ABC>0.03
0
2
4
6
8
0
5
10
15
20
25
30
35
40
0
5
10
15
20
25
30
35
0
20
40
60
80
0
100
200
300
400
500
0
200
400
600
800
1000
Number
−1.0
−0.5
0.0
0.5
1.0
Correlation


Slope between Peak-Gene Connections by ABC Score
ABC>0.45
ABC>0.25
ABC>0.20
ABC>0.14
ABC>0.06
ABC>0.03
0
2
4
6
8
10
12
14
0
10
20
30
40
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60
70
80
0
100
200
300
400
500
0
200
400
600
800
1000
Number
−1.0
−0.5
0.0
0.5
1.0
Slope

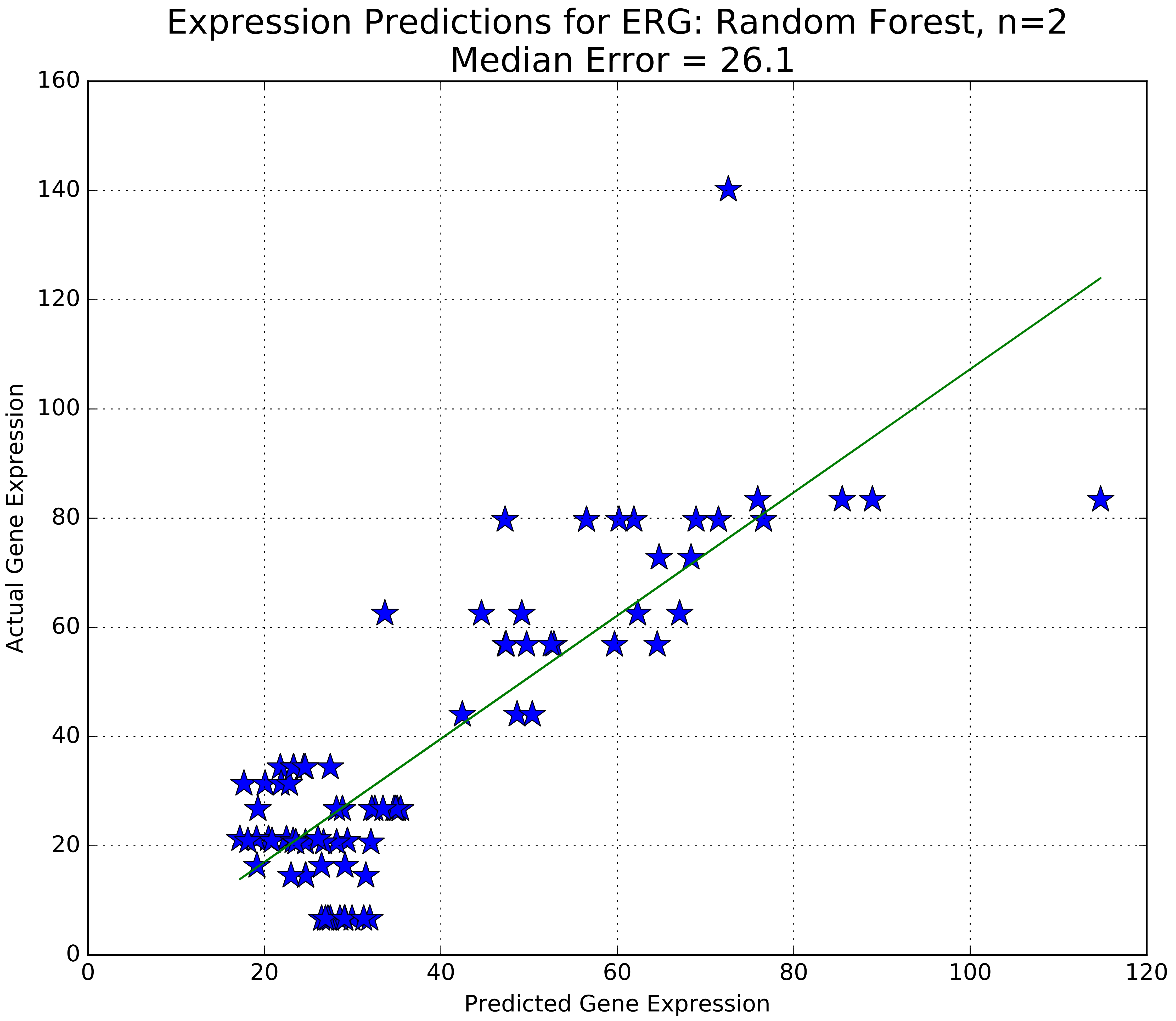

Number of Peaks per Gene
98% Genes
ABC>0.45
85% Genes
ABC>0.25
76% Genes
ABC>0.20
55% Genes
ABC>0.14
19% Genes
ABC>0.06
16% Genes
ABC>0.03
Genes
0
2
4
6
8
10
12
Number of Peaks per Gene

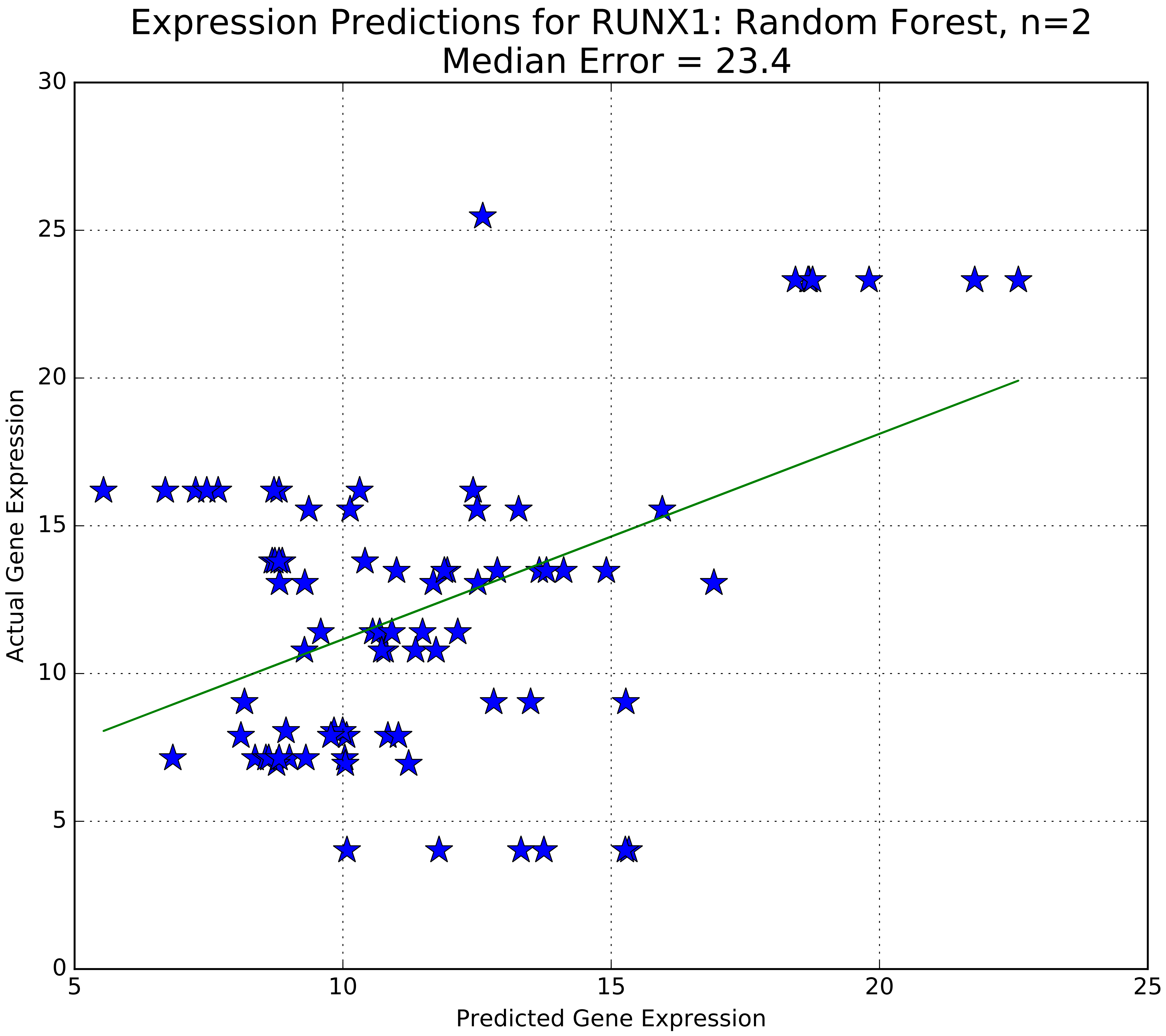

Number of Genes per Peak
100% Peaks
ABC>0.45
99% Peaks
ABC>0.25
99% Peaks
ABC>0.20
98% Peaks
ABC>0.14
93% Peaks
ABC>0.06
87% Peaks
ABC>0.03
Peaks
0
2
4
6
8
10
12
Number of Genes per Peak

Predicting Gene Expression: Process

1. For every gene, identified peaks with ABC > 0.03
    - Features: Peak intensities (16 samples X n peaks)
    - Targets: Gene expression (16 samples)
2. Split targets and features into a training (80%) and testing (20%) set
3. Trained a Multivariate Regression and Random Forest Regression on the training set
4. Predicted the testing set
    - Computed Error and R2
5. Repeated steps 2 — 4 20 times for each gene

Predicting Gene Expression





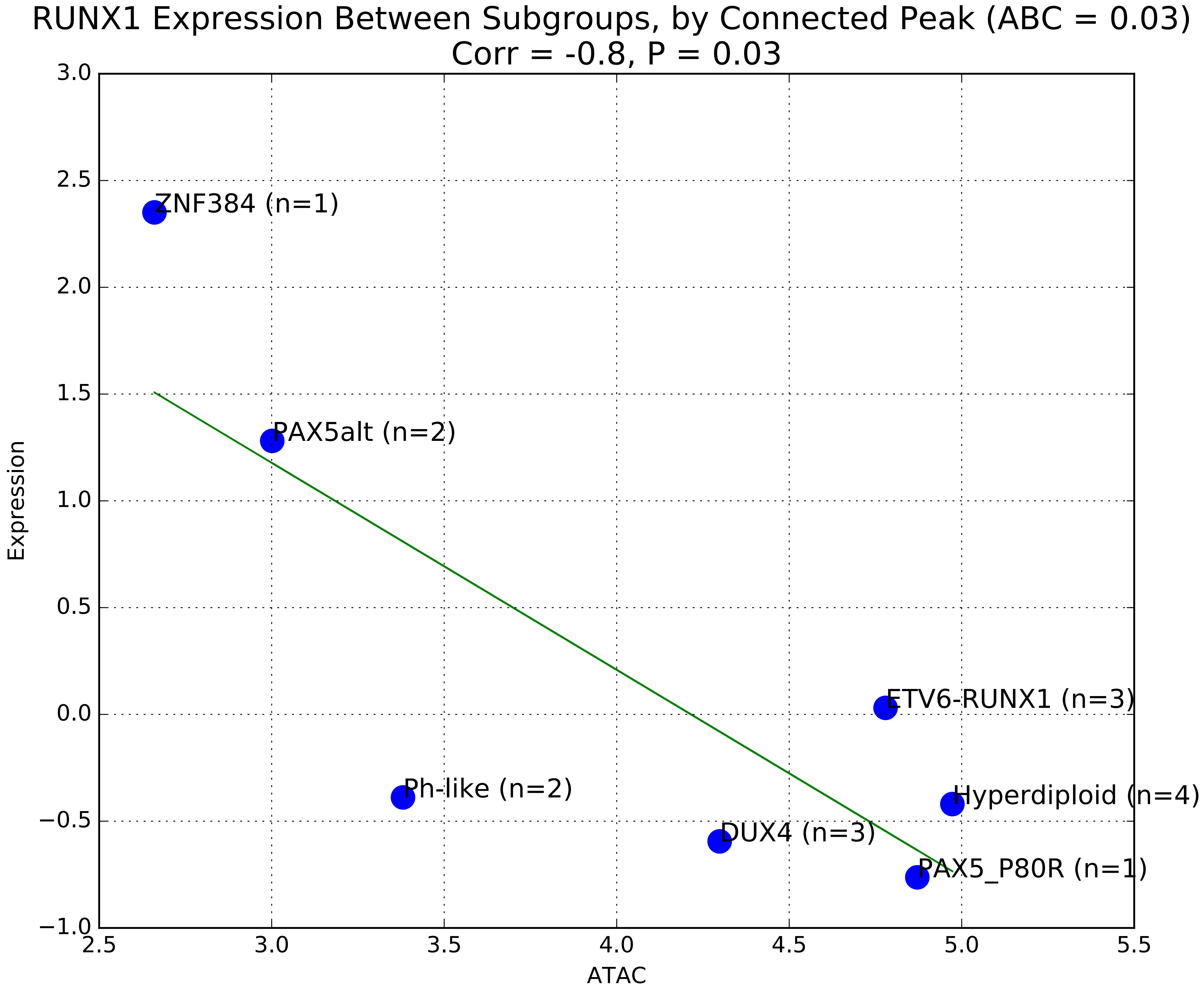
Predicting Gene Expression

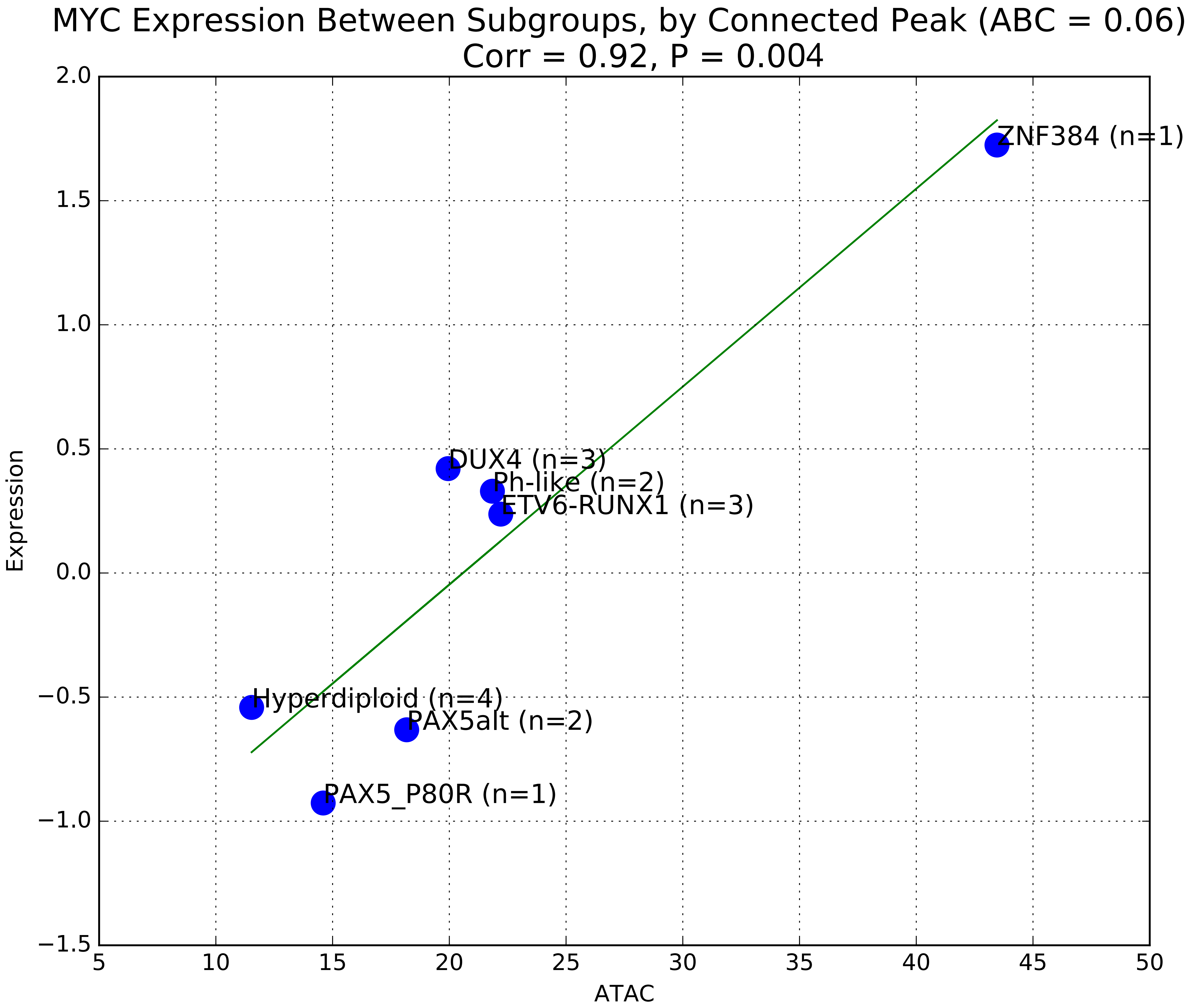
ABC	>0.45	>0.25	>0.2	>0.14	>0.06	>0.03
Median Error: Random Forest	31%	32%	35%	35%	35%	34%
Median Error: Multivariate Regression	50%	34%	32%	32%	32%	31%

Subgroup Variability: Process

- Used the internet to identify a list of 70 leukemia-associated ongogenes
- Averaged gene expression and peak intensity within the 7 subgroups:
    - Hyperdiploid, DUX4, ETV6-RUNX1, PAX5alt, Ph-like, PAX5_P80R, ZNF384
- Computed correlations and slopes between peaks and genes

Subgroup Variability: RUNX1 and MYC





Conclusions

- ABC offers a promising way to identify regulatory elements of genes
  - Performs much better than correlation and is more complex than distance
- In agreement with Fulco et al (2019), we find that enhancers typically regulate multiple genes, and genes are regulated by multiple enhancers
- ABC is highly correlated with distance
    - Not clear whether using contact is better than distance
- ABC might offer a way to examine regulatory elements and interpret the functions of noncoding genetic mutations that influence risk for human diseases, such as Leukemia

Next Steps

- Add components to ABC score
  - DHS or H3K27ac ChiP-Seq
- Motif calling at known connections
- Redo the validation with more types of data
  - CRISPR perturbation data
  - HiChip connections