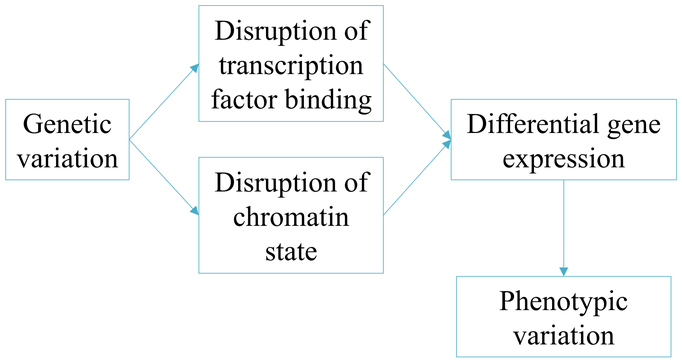
Computational methods for studying the variation in expression and chromatin state across individuals and populations

Sofia Kyriazopoulou-Panagiotopoulou[1], Maya Kasowski[2], Fabian Grubert[2], Judith Zaugg[2], Anshul Kundaje[1,3], Yuling Liu[4], Michael Snyder[2], Serafim Batzoglou[1]

1 Department of Computer Science, Stanford University 2 Department of Genetics, Stanford University School of Medicine 3 Department of Computer Science, Massachusetts Institute of Technology 4 Department of Chemistry, Stanford University

Motivation and background



Questions:

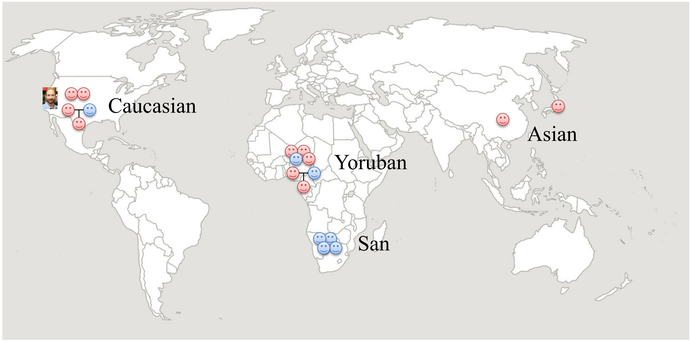
- Is chromatin state heritable? Is it divergent across populations? What are the implications of such differences?
- Can we systematically link variation in transcription factor binding to differences in gene expression?

Why do we care?

- Better understanding of regulatory (genetic and epigenetic) mechanisms and their involvement in disease.

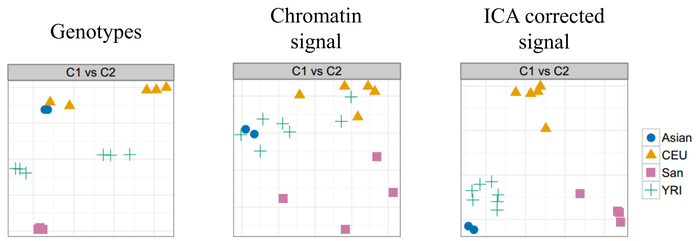
Datasets

Gene expression (RNA-seq) and chromatin state (ChIP-seq) from 19 individuals of diverse ancestry.



1. Correlating chromatin state differences with genetic divergence

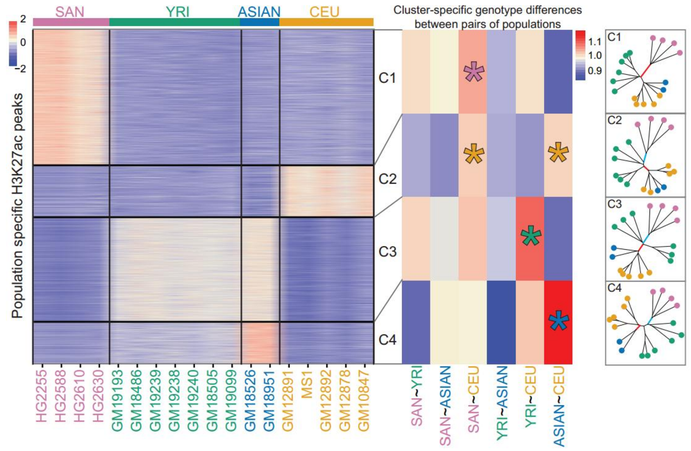
Goal: Understand whether chromatin variability relates to genetic differences.
Problem: Confounding factors, such as experimental batch effects, lead to unexpected structure in the chromatin signal.



Solution: Independent Component Analysis (ICA) to remove confounders.

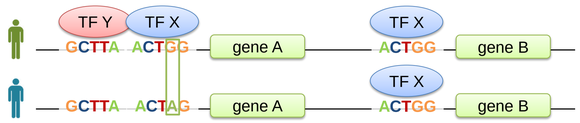
- Input signal X_{ij}: signal at peak i of the mark in individual j.
- Regression on population to get residual signal matrix R.
- ICA: $R \approx SA$. The rows of A are the confounding factors.
- Regression on population and the confounders.
- Corrected signal: $\tilde{X}_{ij} = X_{ij} - \sum_k a_{kj}\theta_{ki}$

The corrected signal correlates with genetic divergence!



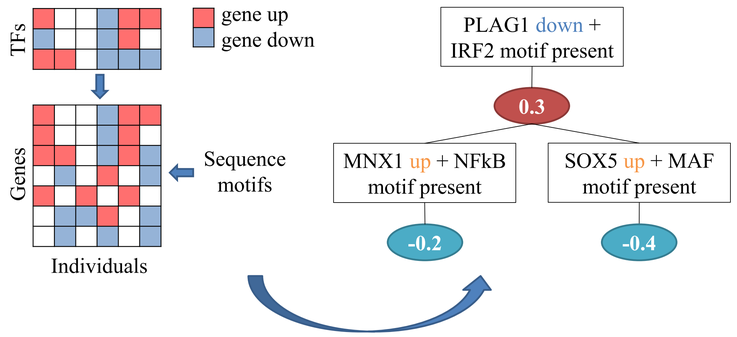
2. Linking variation in sequence motifs to context-specific gene expression

Goal: Use patterns of gene behavior both across the genome and across individuals to find rules that determine gene up- and down-regulation.
Example:



Rule: genes with the motif ACTGG are turned on by TF X.

Method:



LogitBoost (boosting with log-loss) to build an Alternating Decision Tree

- Model AUC: 0.88 on just 3252 genes.
- Learned models can **predict TF associations and sequence affinities in different contexts**!

Conclusion

- We show how we can **systematically remove confounding factors** to identify population structure in chromatin data that correlates with genetic diversity.
- We present a novel method for **modeling how genetic diversity leads to context-specific gene expression**.