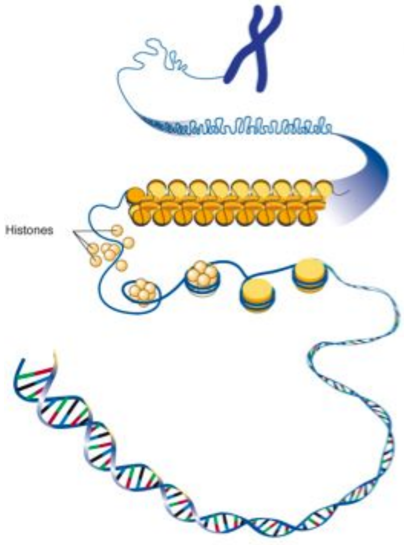
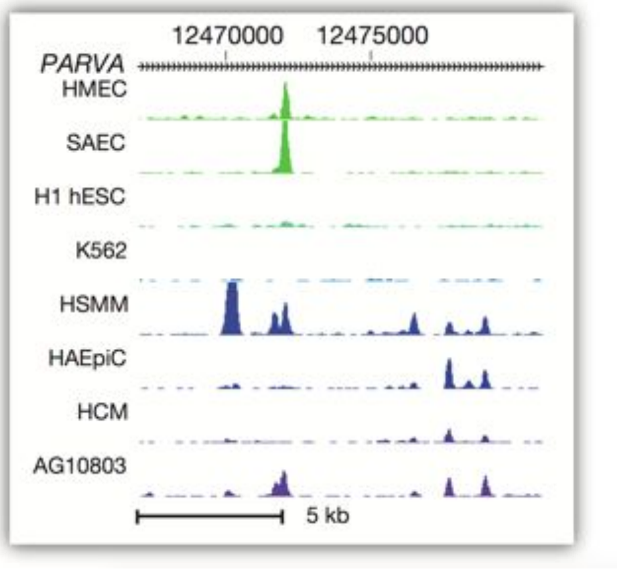
Predicting Pioneer Transcription Factors from Chromatin Accessibility and Gene Expression

Nae-Chyun Chen, Arun Das, Xinyu Feng

Motivation



Density of DNaseI cleavages



Thurman *et al. Nature*, 2012.

- The eukaryotic genome is packaged into chromatin
  - Different cell types vary in their accessible chromatin landscape
  - Open chromatin regions play an important role in cell identity, marking the gene regulatory regions which are accessible to the cellular environment (thus regulating expression)
    - Misregulation of genes can lead to diseases such as cancer
- Open chromatin is the hallmark of active expression, and understanding how it is established and regulated is important
  - However, the principles behind how open chromatin landscape is established and regulated are poorly understood

Introduction

- Recent studies have proposed that a subset of transcription factors, termed **pioneer factors**, might play a role in regulating chromatin accessibility by targeting specific genomic regions.
- Goal: Identify pioneer transcription factors that regulate chromatin accessibility
- Although there are a few well-studied pioneer factors, there has been no systematic experimental investigation to characterize all the pioneer factors.
  - Traditional experimental approaches are low-throughput and limited in scope

Introduction

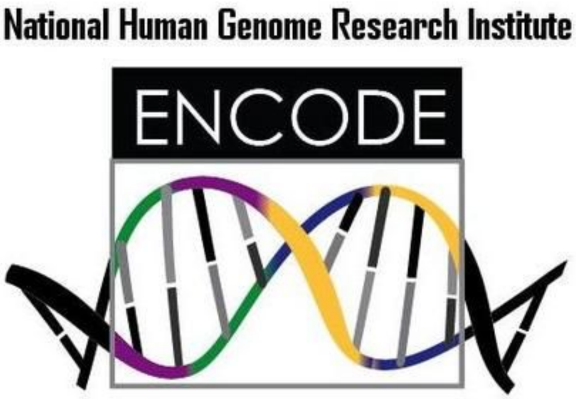
- Our aim is to uncover a list of candidate human pioneer TFs to complement and guide experimental research
- Our preliminary results show that for each cell type, there are a unique set of TFs that predict chromatin accessibility.
    - The cell type variability may be due to differential expression of pioneer factors, which regulate different downstream targets

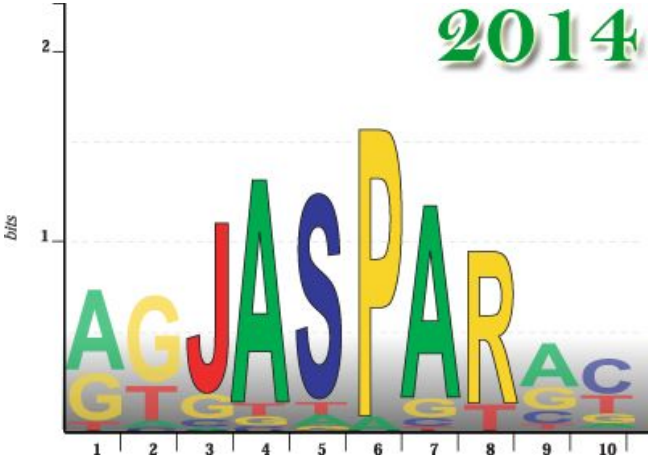
In this project, our goal is to incorporate TF expression level into a model that predicts chromatin accessibility, and to reveal how expression level is related to the pioneer potential of transcription factors.

Data

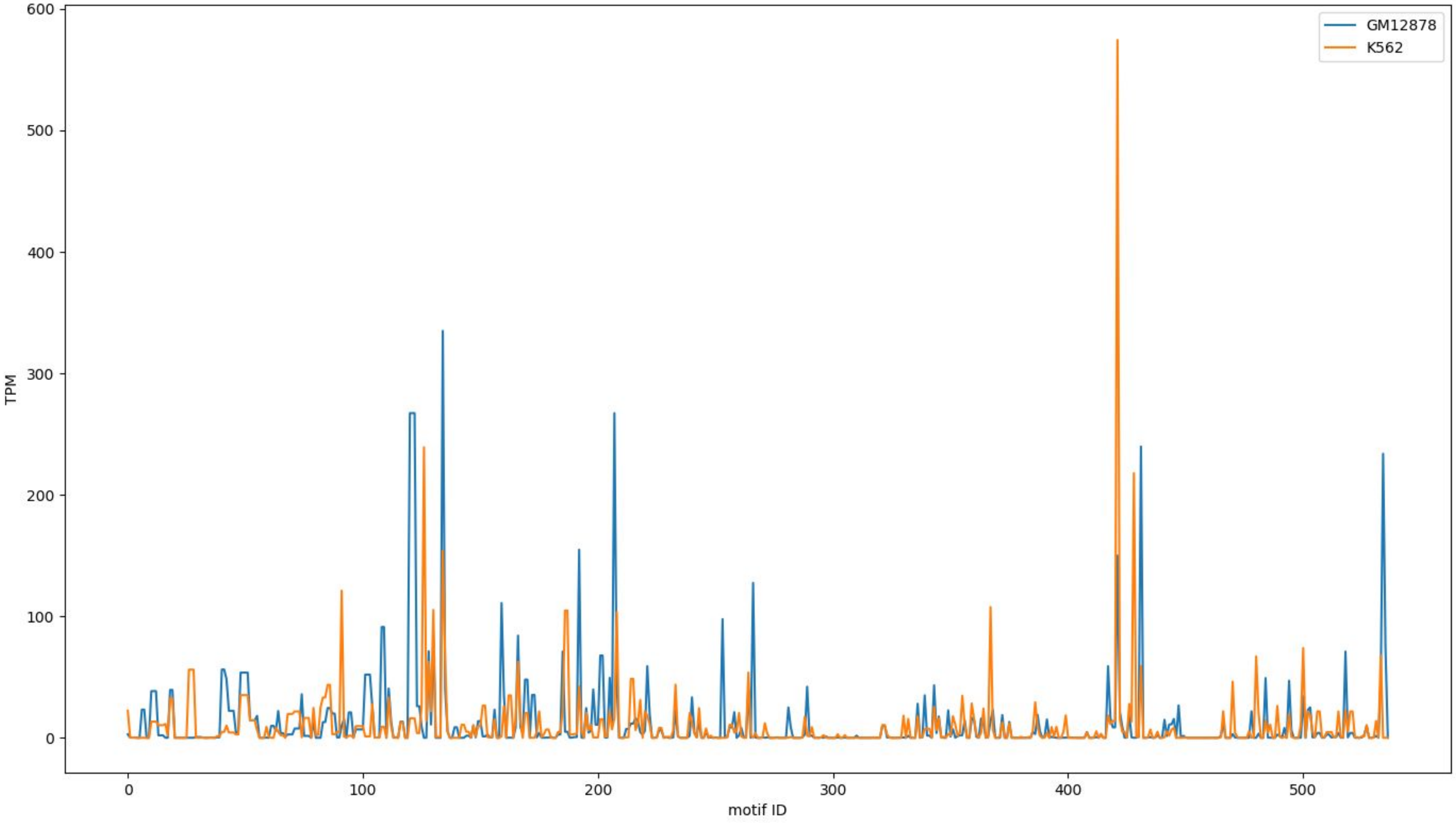
- ENCODE Project: DNase-seq and poly-A RNA-seq
  - Cells lines: human K562 and GM12878
- JASPAR database: 537 TF motifs position weight matrices (PWM)

- Each PWM is scanned across regions of interest to obtain the number of significant motif matches using fimo (Bailey et al. 2009)



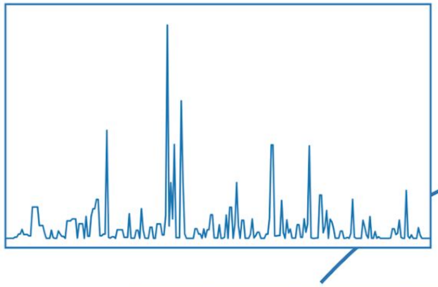


Motivation - TFs are not equally expressed



Data

Labels



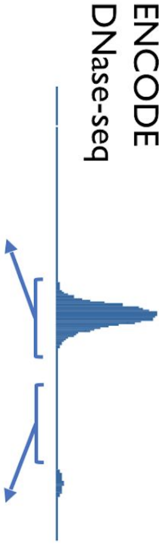
Weight each motif by its TPM

	1	2	…	537	537+1	537+2	…	537+537	openness
site 1	0	3		1	0	4.5		1.5	1
site 2	1	2		0	1	2		0	1
…									
site N	0	0		1	0	0		1.5	1

	1	2	…	537	537+1	537+2	…	537+537	openness
site N+1	1	0		2	1	0		3	0
site N+2	0	2		0	0	3		0	0
…									
site 2N	0	1		1	0	1.5		1.5	0

open chromatin regions (DNase-seq peaks)

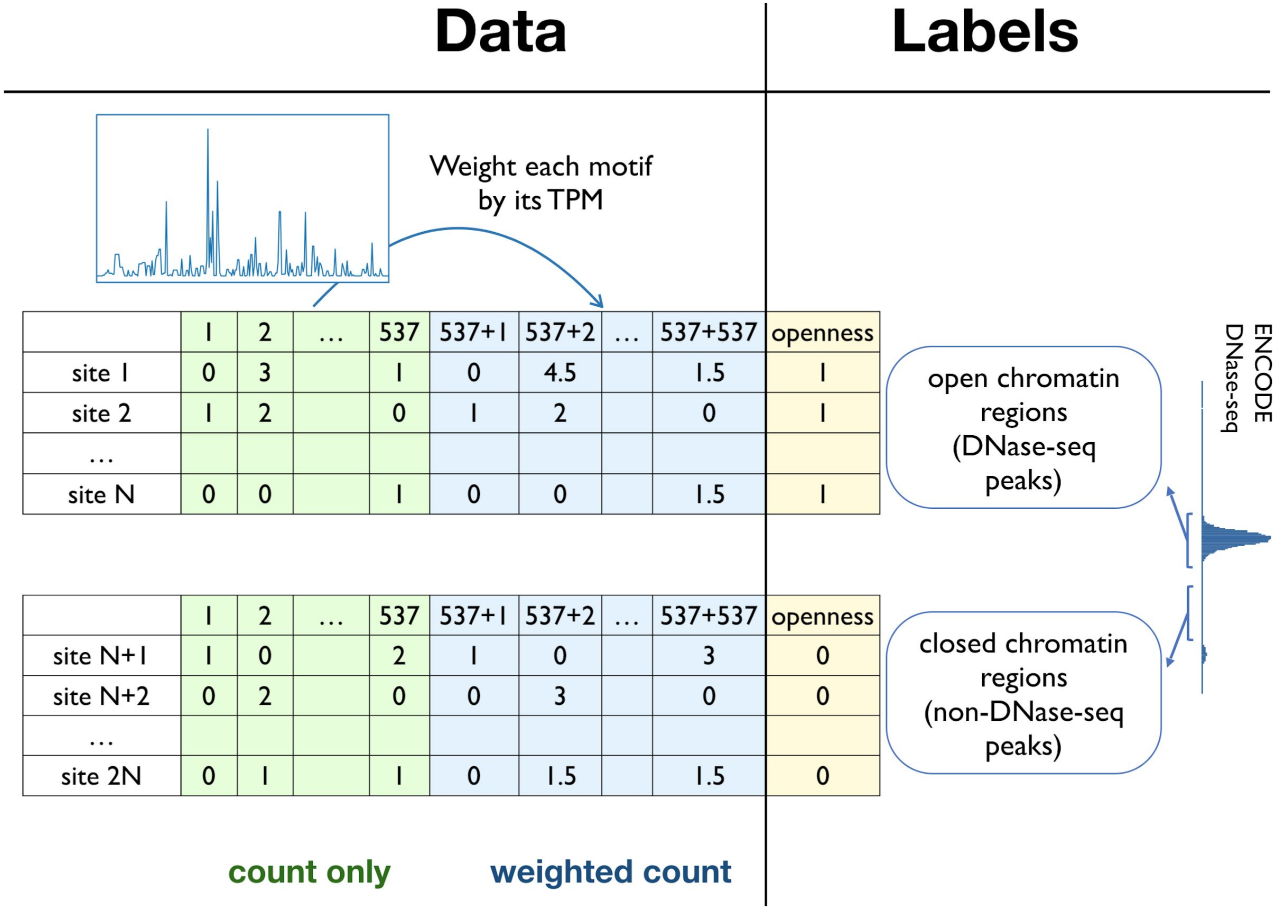
closed chromatin regions (non-DNase-seq peaks)



count only

weighted count

Training a logistic regression model



Data | **Labels**

Weight each motif by its TPM

	1	2	…	537	537+1	537+2	…	537+537	openness
site 1	0	3		1	0	4.5		1.5	1
site 2	1	2		0	1	2		0	1
…									
site N	0	0		1	0	0		1.5	1

	1	2	…	537	537+1	537+2	…	537+537	openness
site N+1	1	0		2	1	0		3	0
site N+2	0	2		0	0	3		0	0
…									
site 2N	0	1		1	0	1.5		1.5	0

count only

weighted count

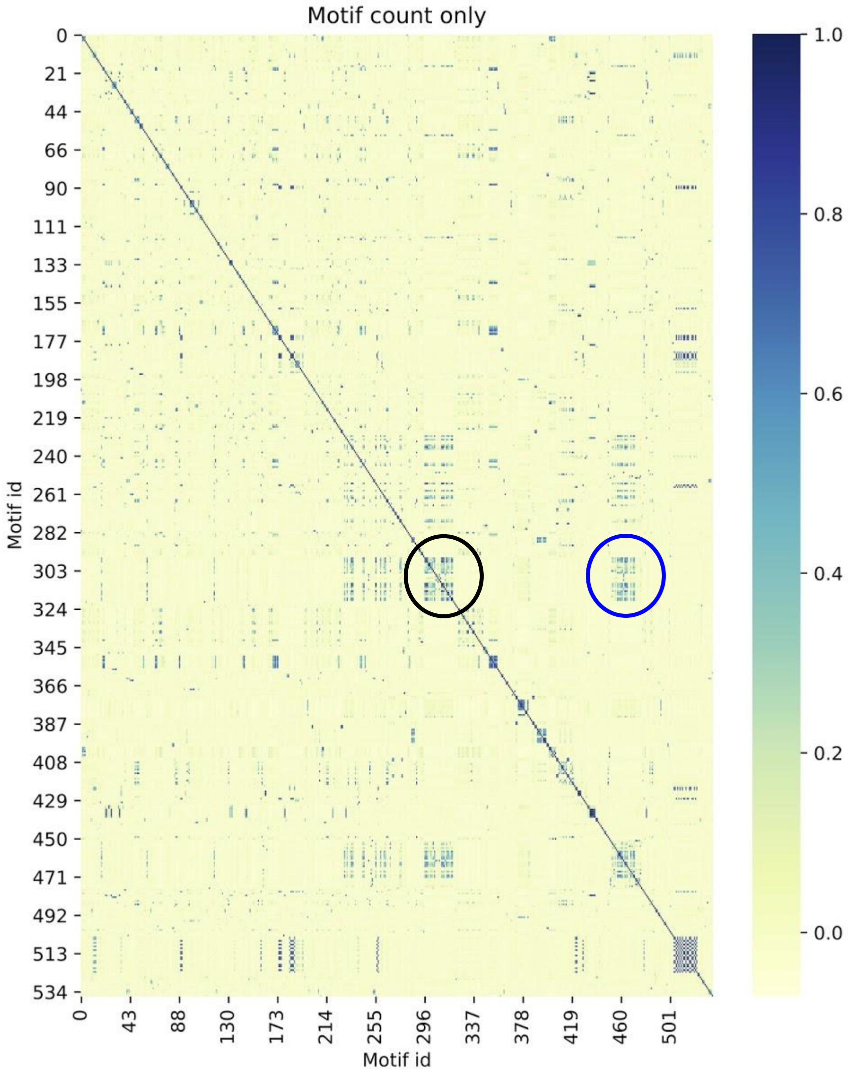
open chromatin regions (DNase-seq peaks)

closed chromatin regions (non-DNase-seq peaks)

ENCODE DNase-seq

- Data matrix is shuffled and split, ⅔ for training and ⅓ for testing
- Train a logistic regression classifier using sklearn
- Model parameters are optimized using 5-fold cross validation
- Accuracy and area under ROC curve (AUC) are reported to evaluate model performance

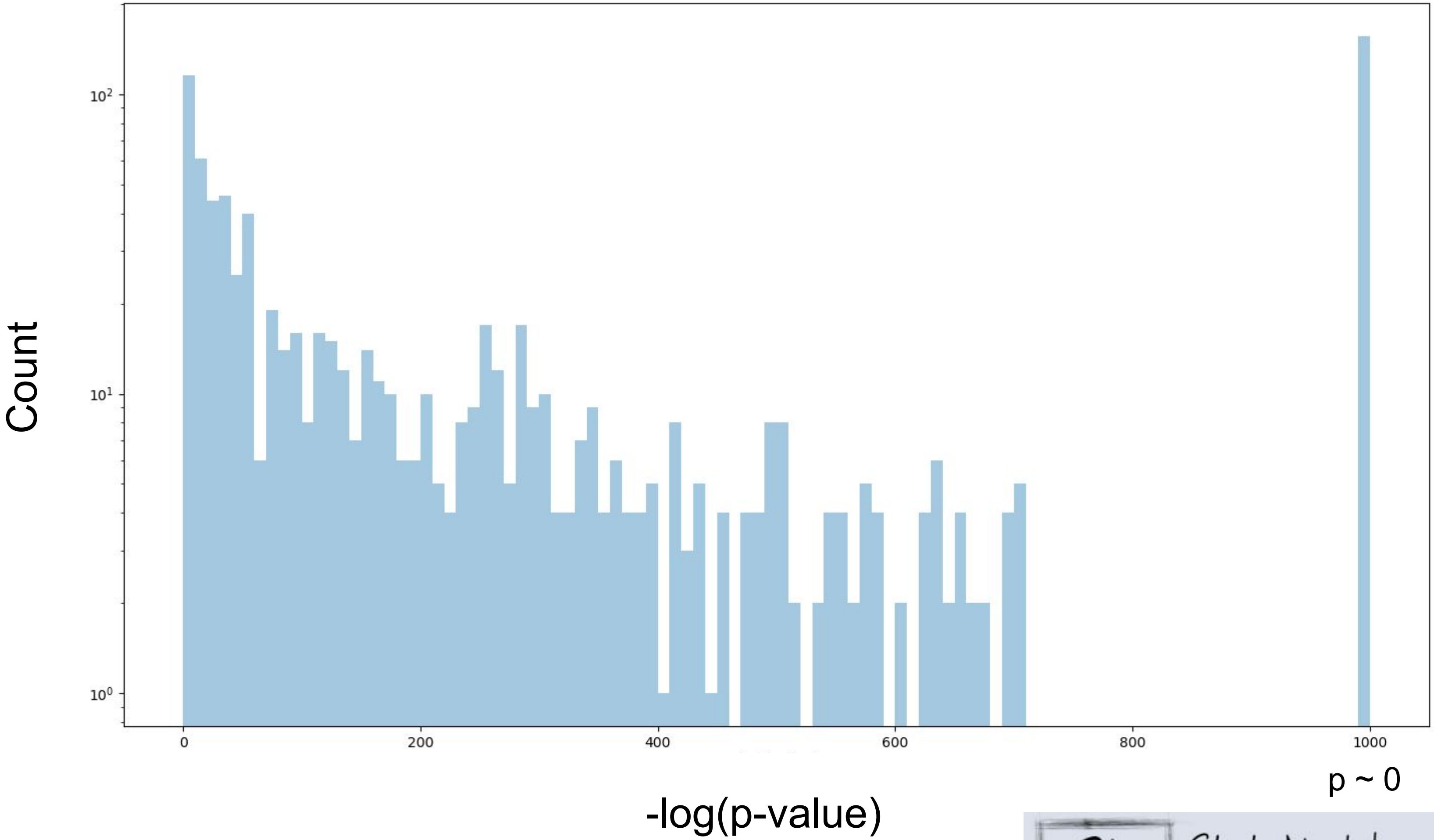
Correlation in feature space



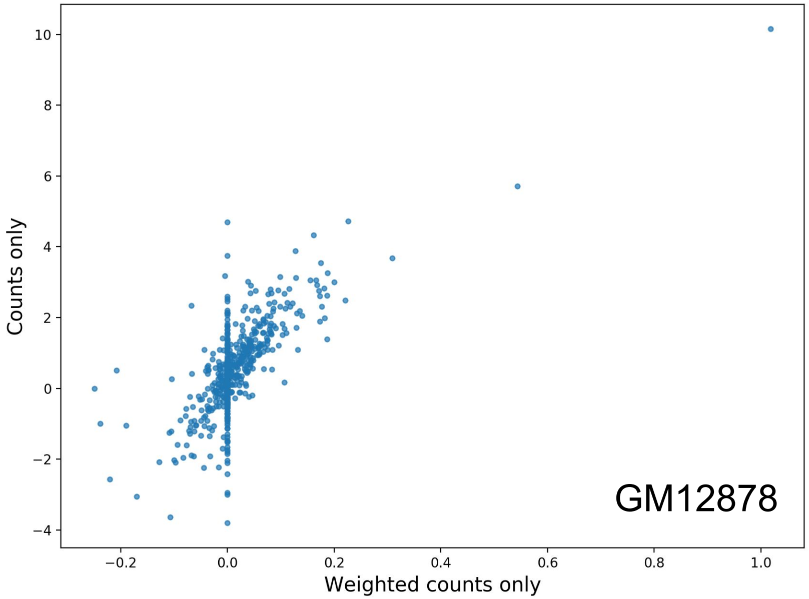
- Diagonal correlations are due to similar binding motifs of the same TF

- Off-diagonal correlations could be two different TFs that bind to the same region cooperatively

Most individual features are significantly correlated with chromatin accessibility

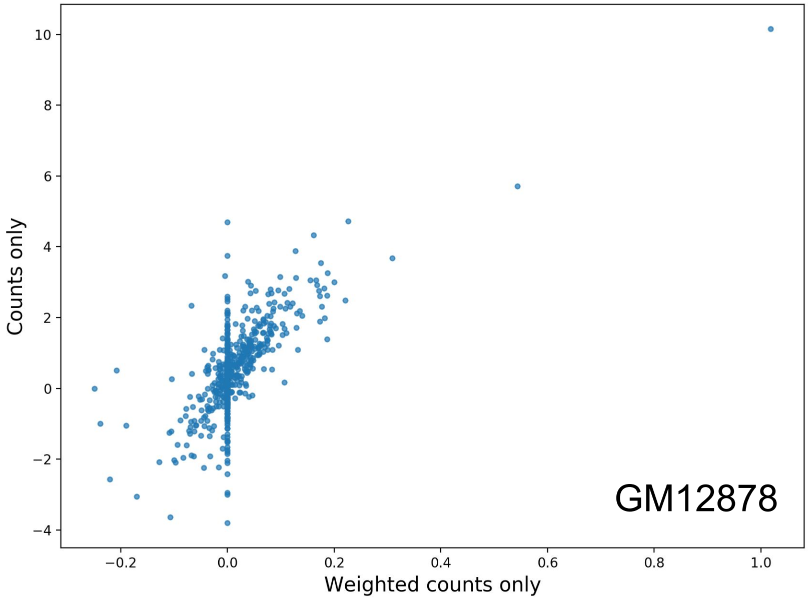


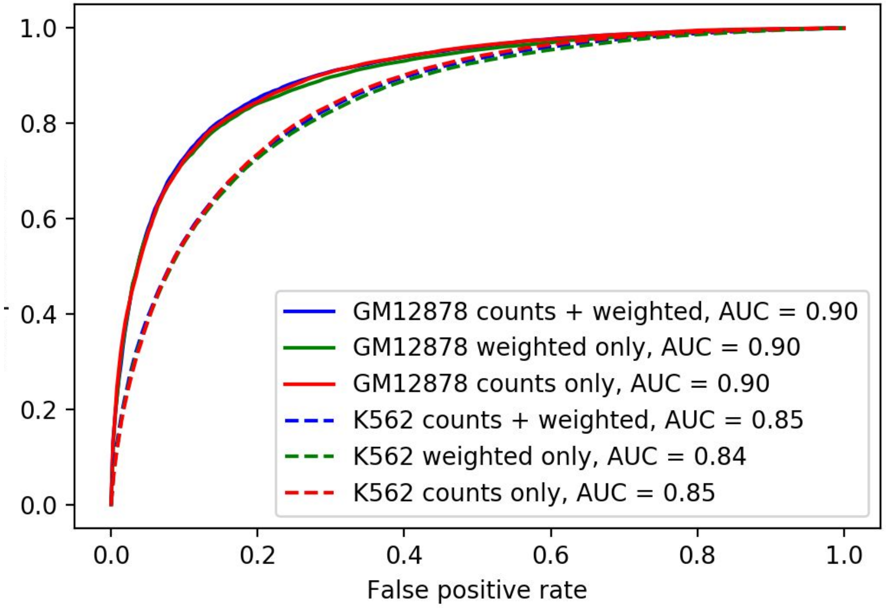
Incorporating expression data filters out 30% features while preserving performance



- 30% of expression-weighted features have zero coefficients
- Coefficients are positively correlated between counts only model and weight counts model

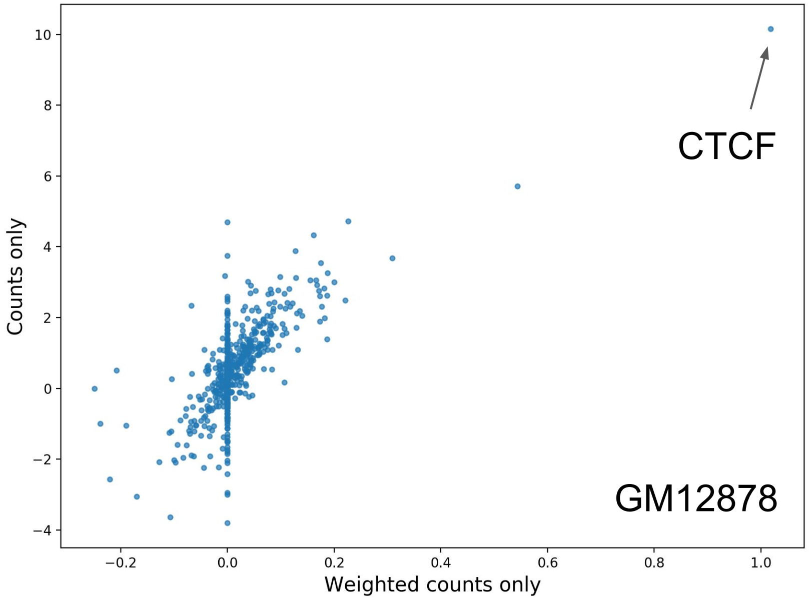
Incorporating expression data filters out 30% features while preserving performance

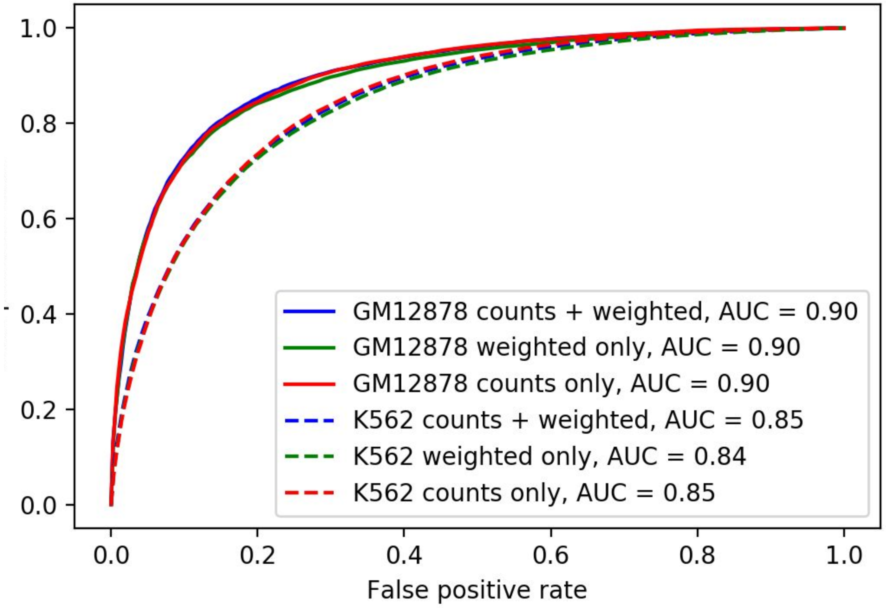




- 30% of expression-weighted features have zero coefficients
- Coefficients are positively correlated between counts only model and weight counts model
- Performance remains the same after incorporating TF expression

Incorporating expression data filters out 30% features while preserving performance





- 30% of expression-weighted features have zero coefficients
- Coefficients are positively correlated between counts only model and weight counts model
- Performance remains the same after incorporating TF expression
- Most predictive features include TF motifs enriched in open chromatin, such as CTCF

Conclusions

- We can predict chromatin accessibility using TF motifs with similar performance as existing methods (Basset, Kelley *et al.* 2016)

- After taking into account TF expression, we can achieve the same performance using a reduced feature space

- Most predictive features include known TFs motifs enriched in open chromatin, thus are promising pioneer TF candidates

Next Steps

- Establish a pipeline to identify TF motifs predictive of chromatin accessibility from DNase-seq and expression data of any cell line

- Once the pipeline is set up, we will compare the most effective features across different cell lines to see if they can be used to identify TFs with pioneer factor properties

- Explore other sources of data such as FAIRE-seq, total RNA-seq and other TF motif databases