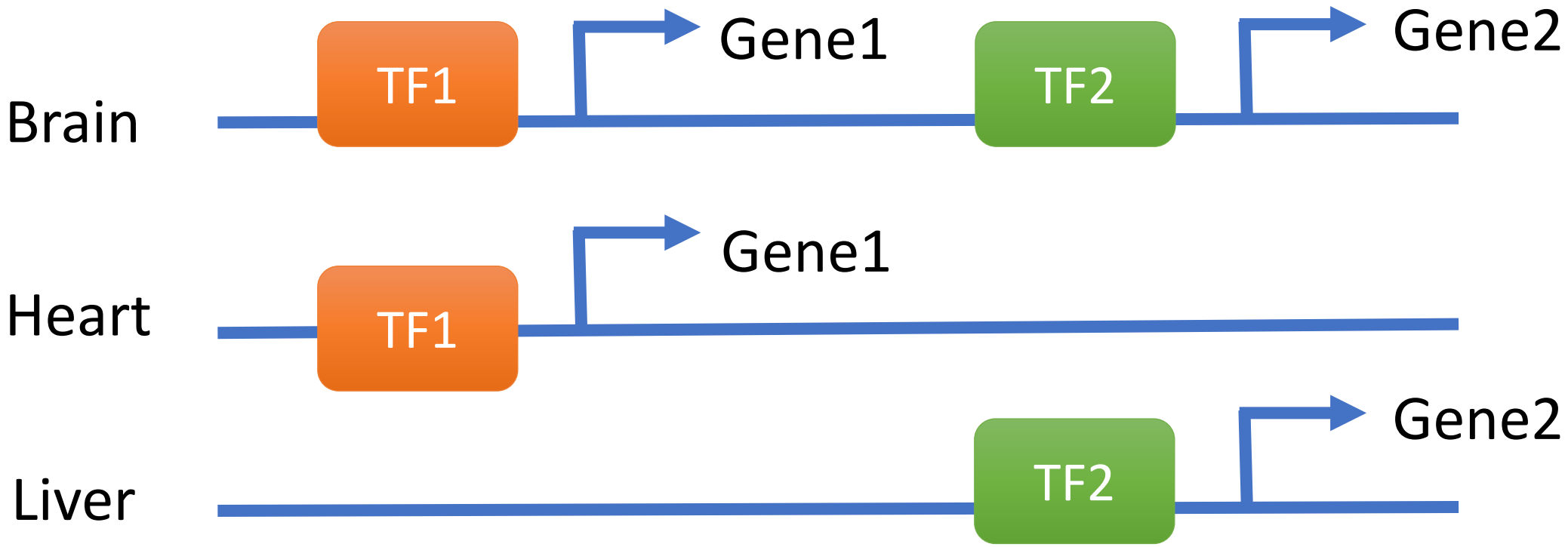
Genome-wide prediction of chromatin accessibility based on gene expression

Junyue Cao, Jiayuan Guo, Yu Liu

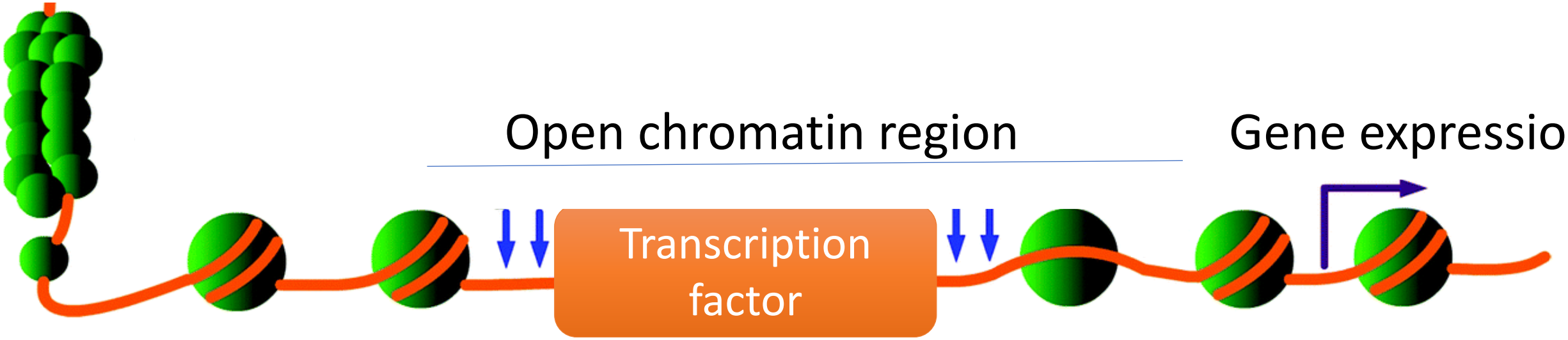
Outline

- Aim and Background
- Data pre-processing and dimension reduction
- Model selection for classification (and regression)

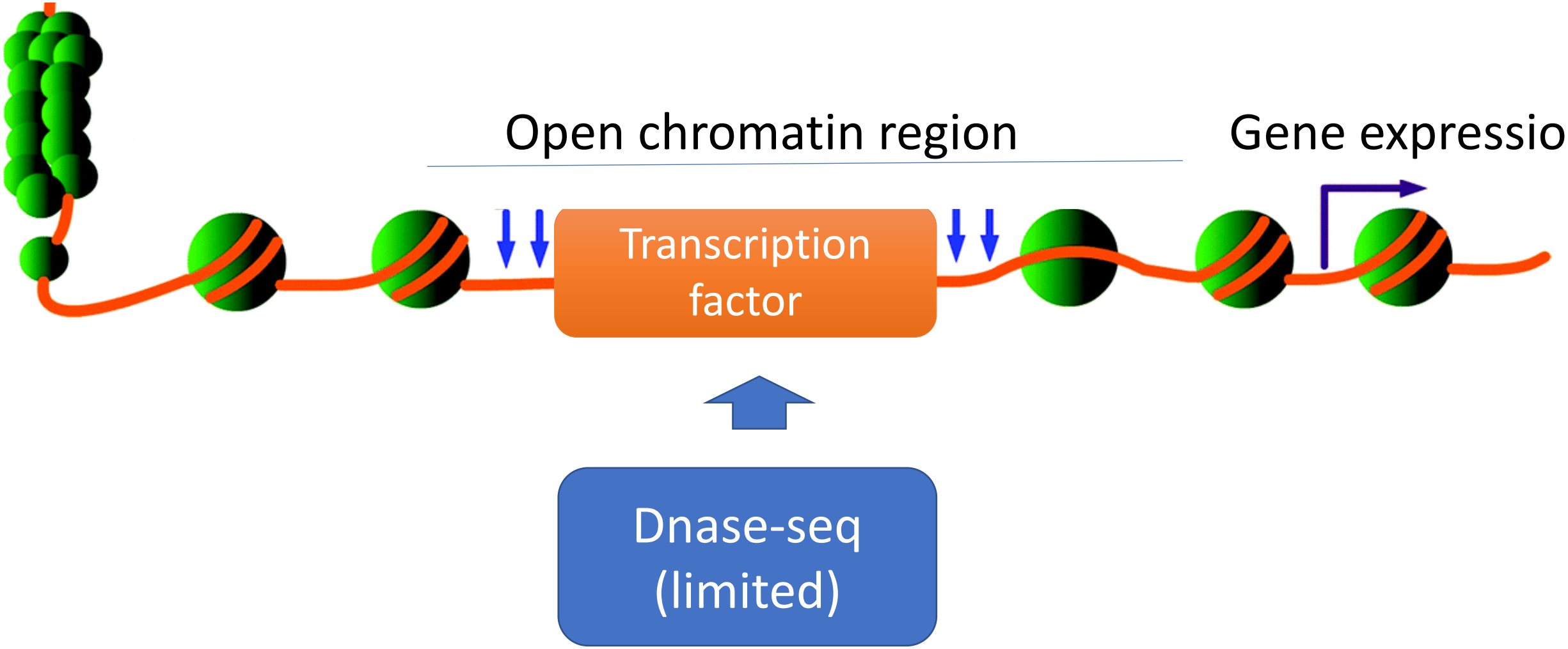
Tissues are regulated by transcription factors



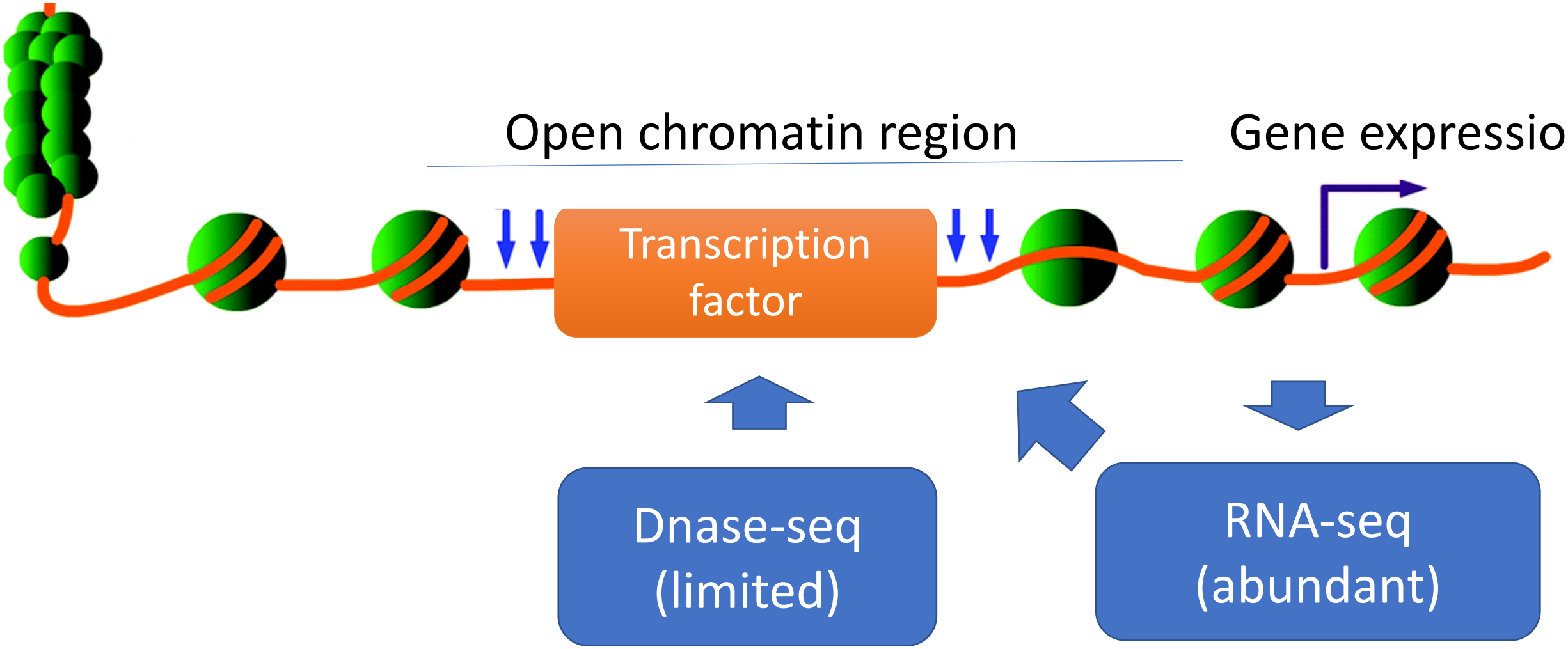
Aim: predicting chromatin accessibility based on RNA expression



Aim: predicting chromatin accessibility based on RNA expression



Aim: predicting chromatin accessibility based on RNA expression

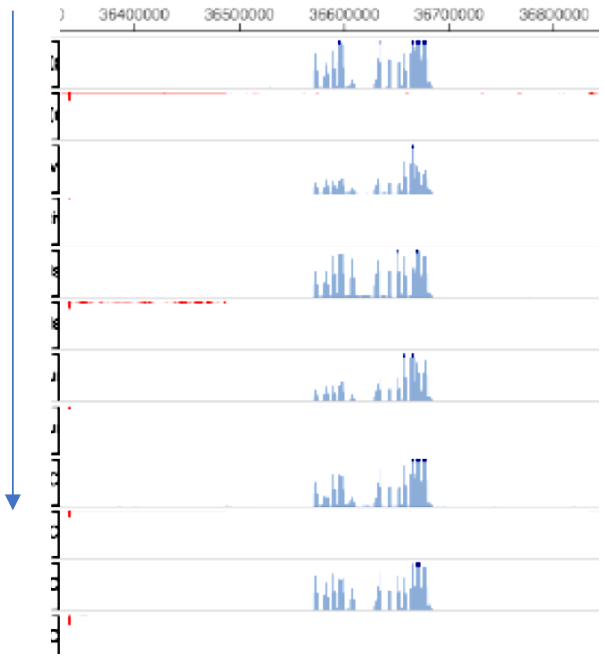


Data source

RNA-seq

Chromatin location

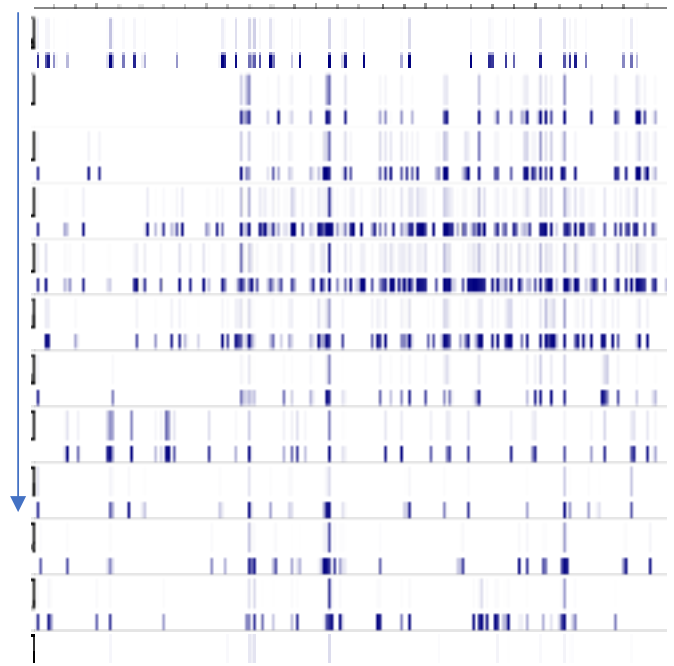
tissues



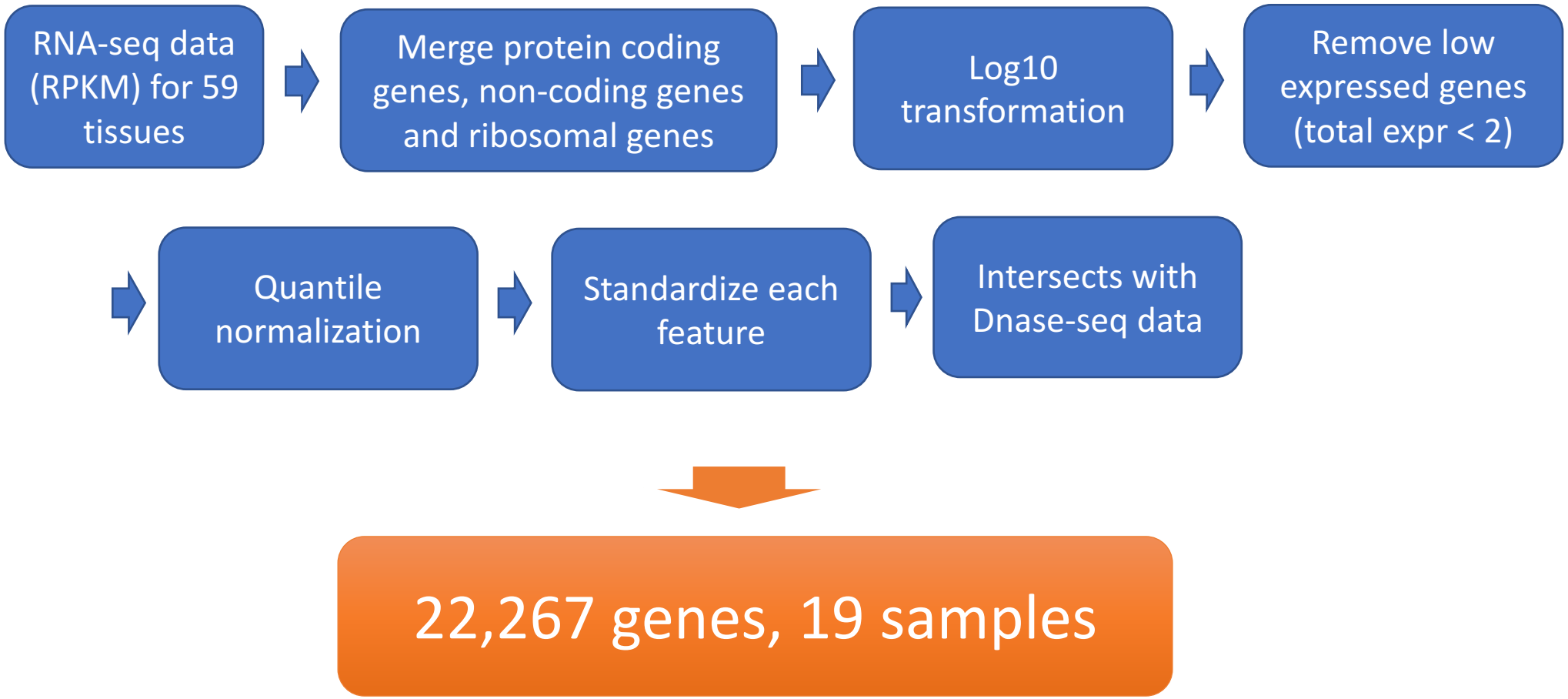
Dnase-seq

Chromatin location

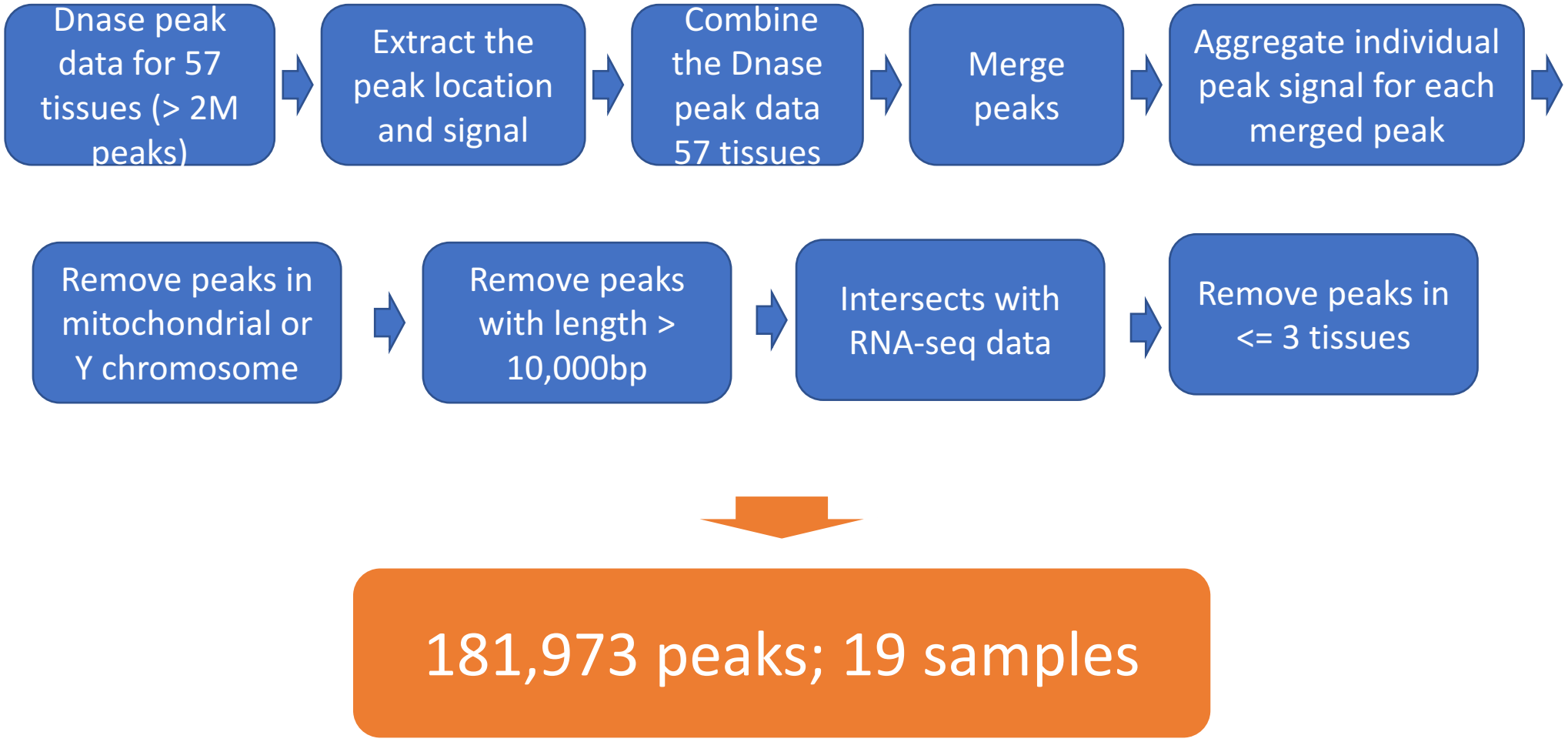
tissues



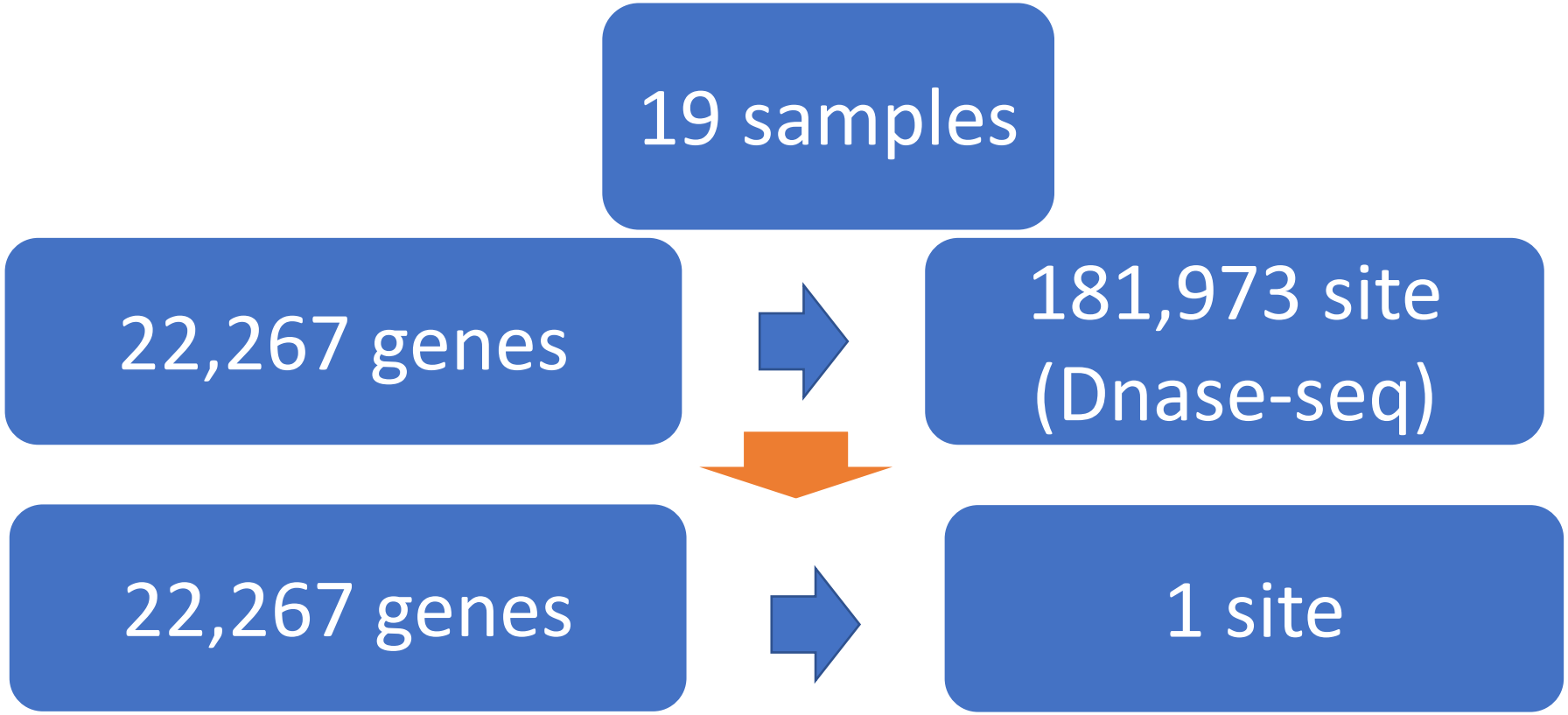
RNA-seq data pre-processing



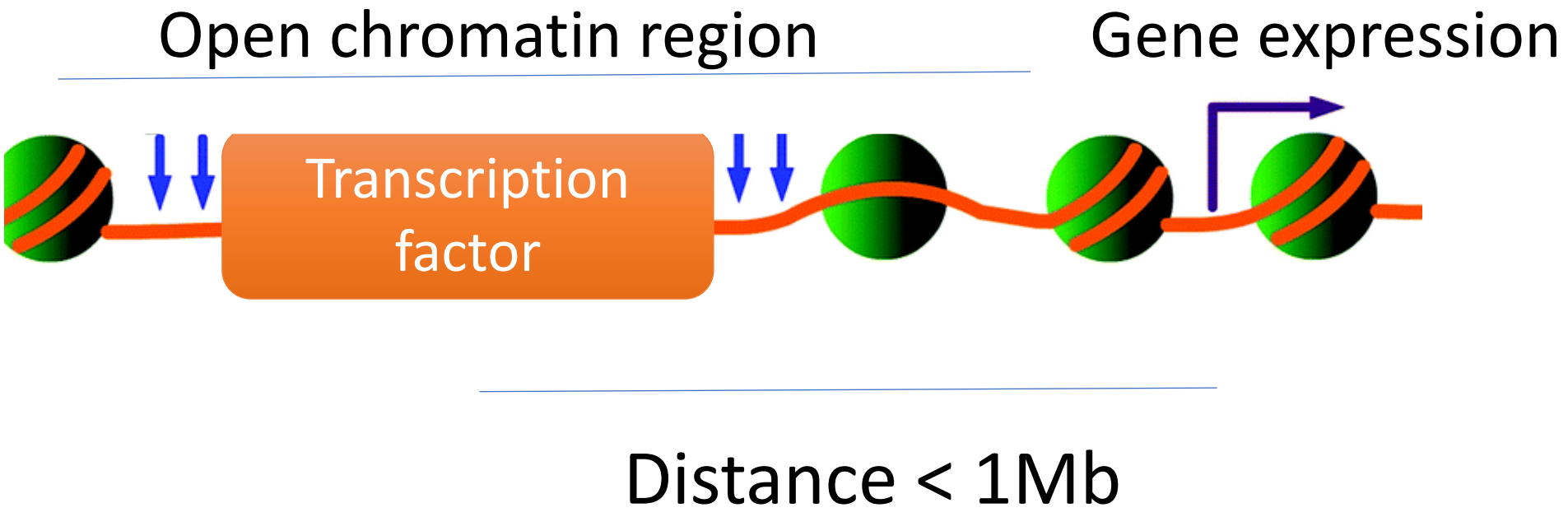
Dnase-seq data pre-processing



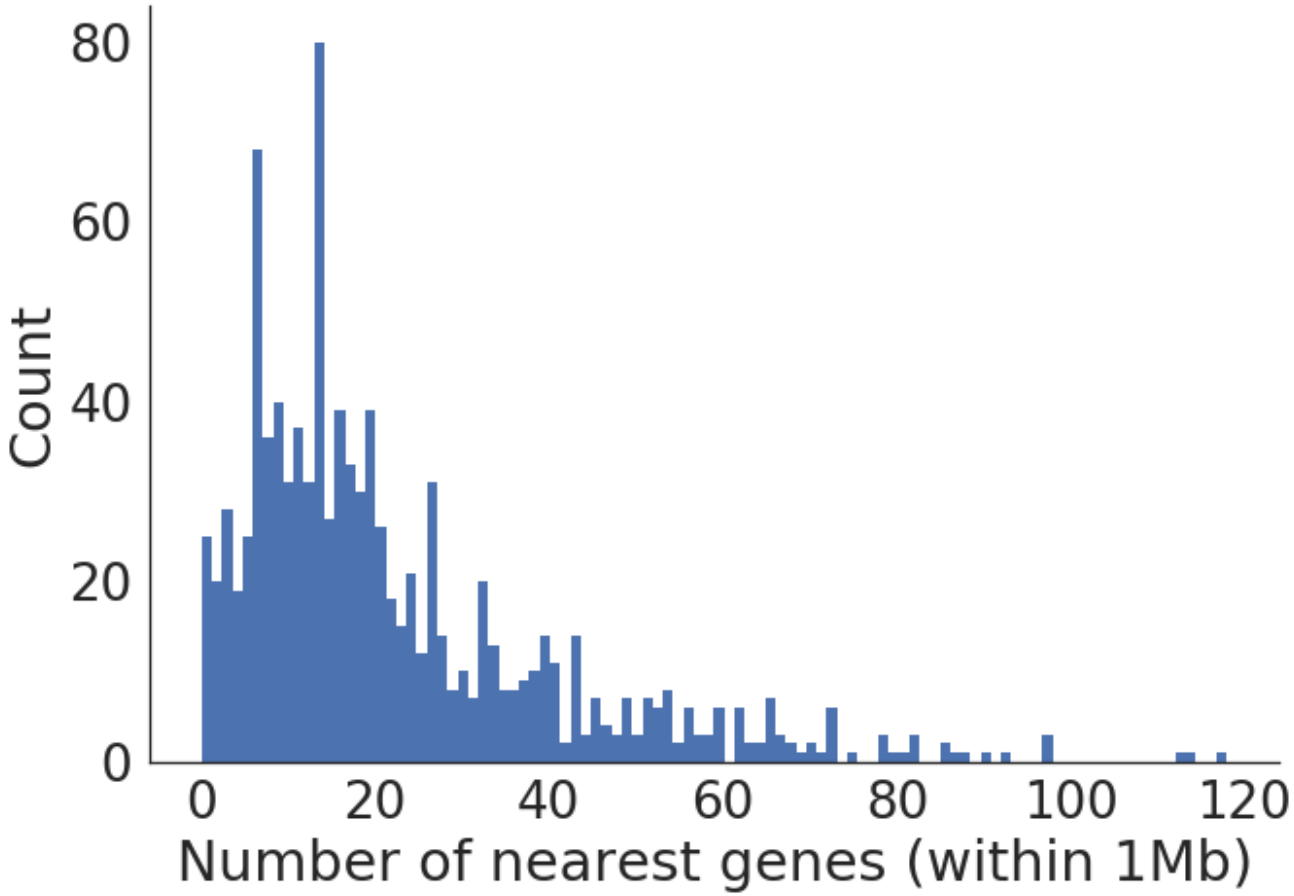
Challenge: Predicting high dimensional data from high dimensional data with very limited samples

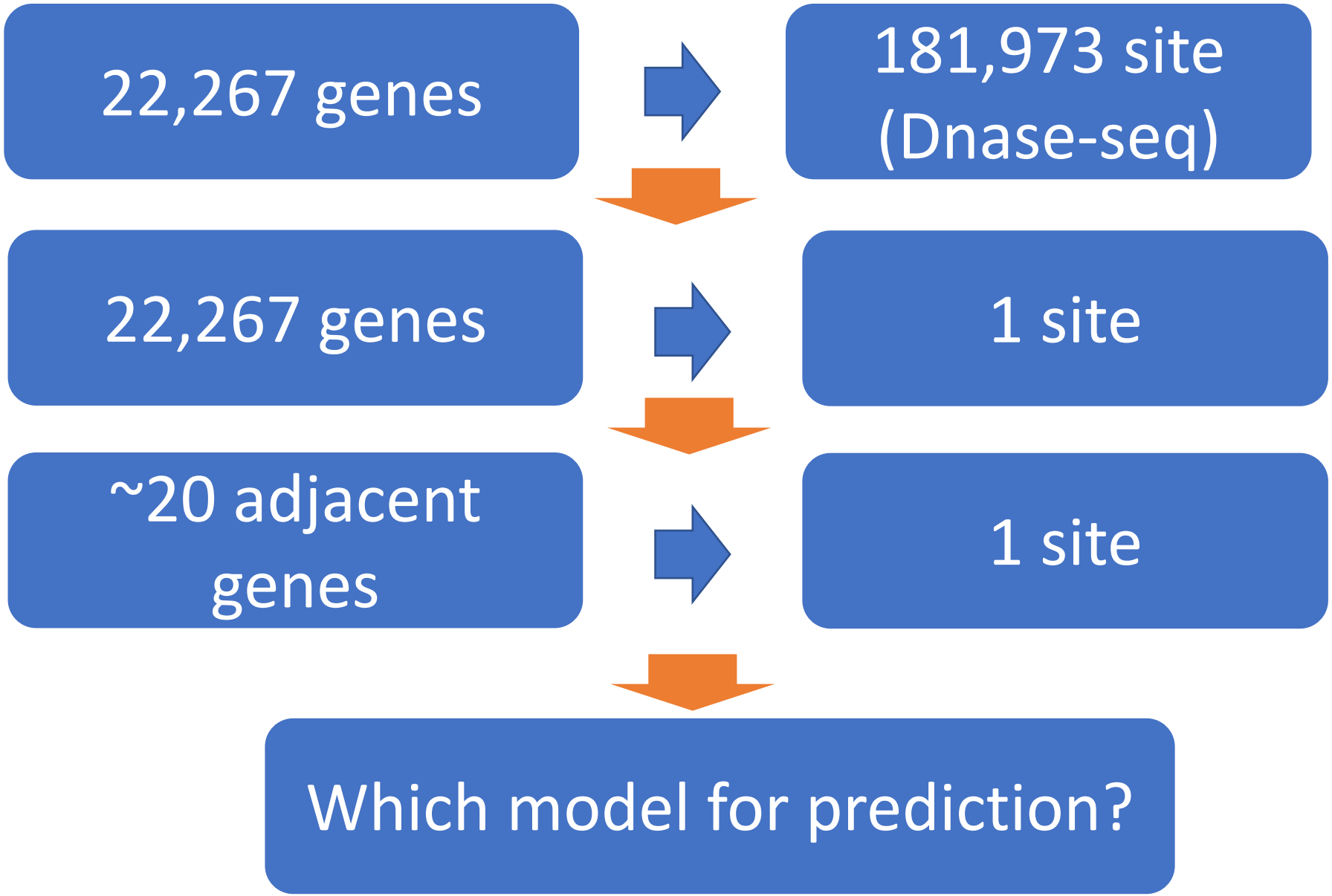


Chromatin open sites are close to regulated genes

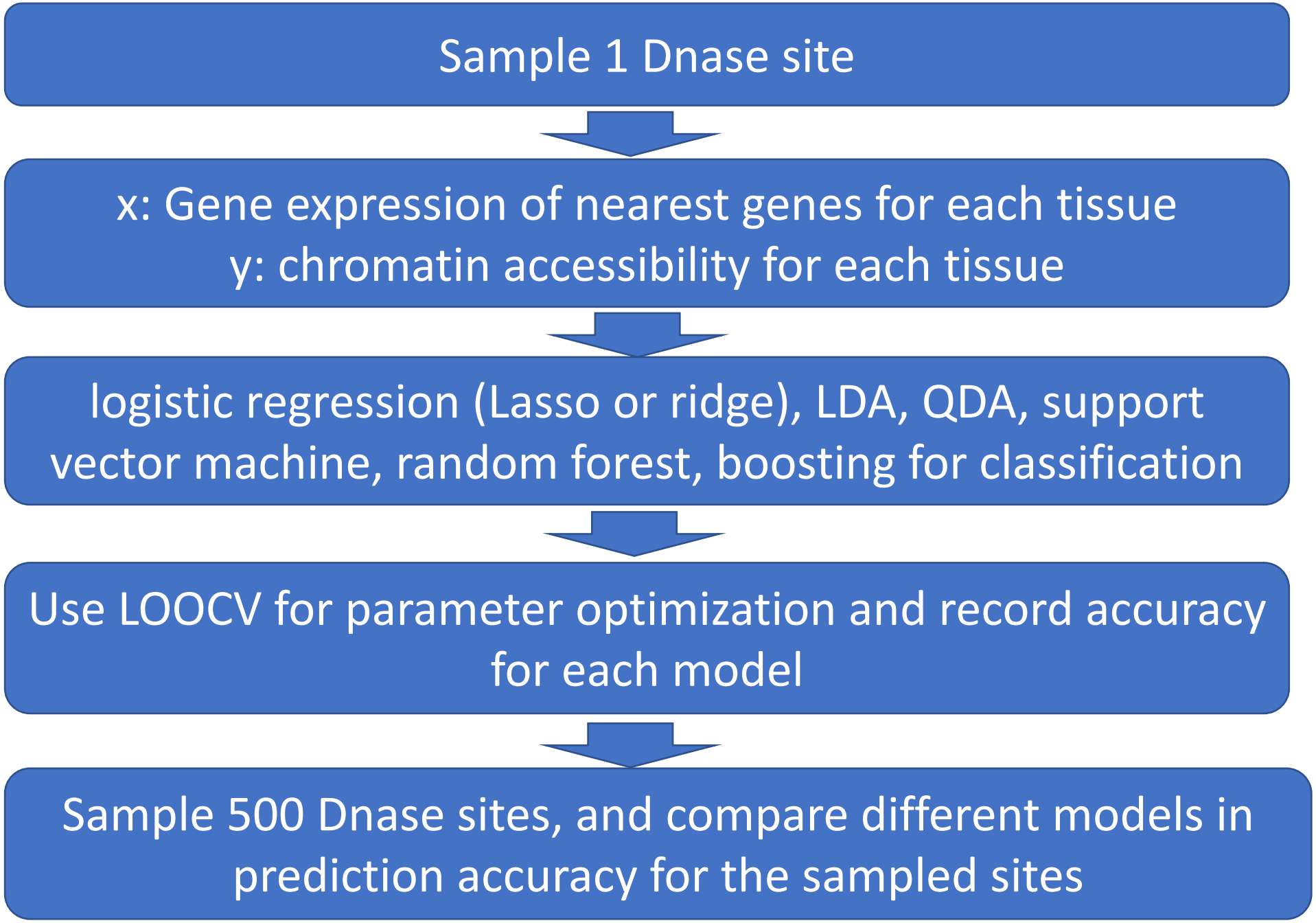


Chromatin open sites are close to regulated genes




22,267 genes
181,973 site (Dnase-seq)
22,267 genes
1 site
~20 adjacent genes
1 site
Which model for prediction?

Model selection



Model selection

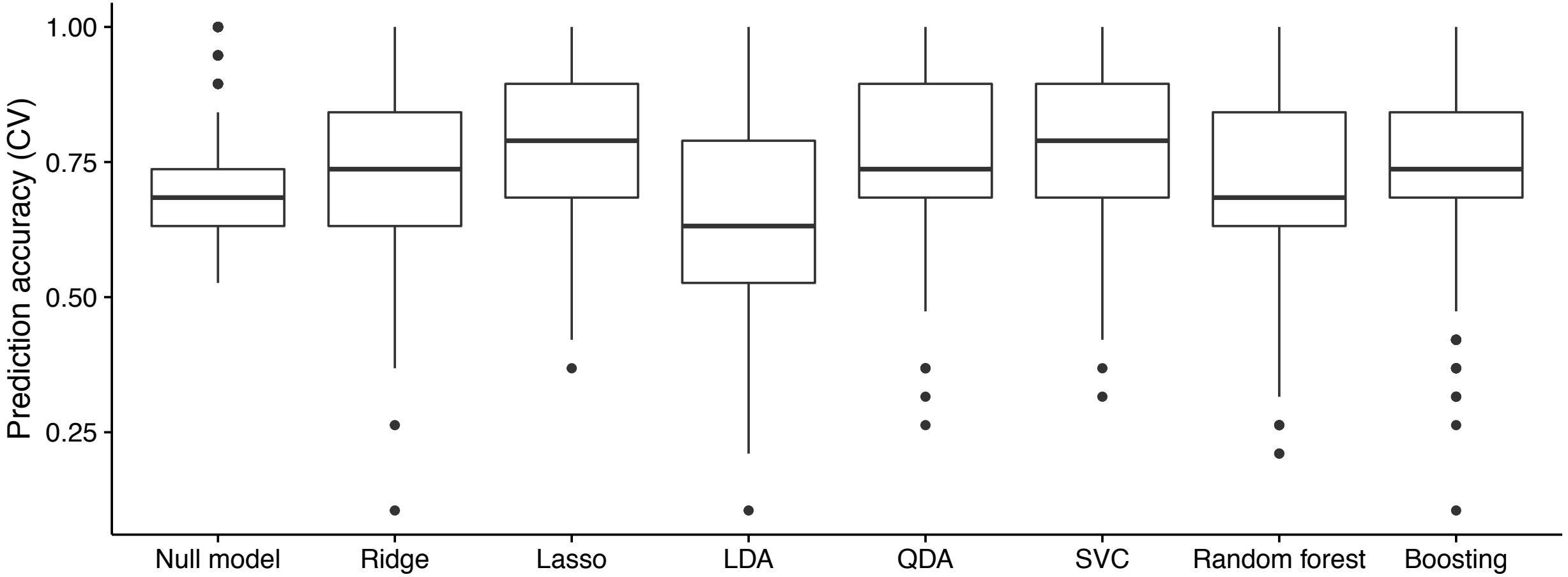
for each Dnase site							
Y vector	DNAseq		X matrix	RNA 1	RNA 2	RNA 3	…
	(open) 1		tissue1	0.22	0.65	0.67	…
	0		tissue2	..	…	…	
	1		tissue3	…	signals	…	
	…		…			…	
	1		tissue19	…	…	…	

logistic regression (Lasso or ridge), LDA, QDA, support vector machine, random forest, boosting for classification

Use LOOCV for parameter optimization and record accuracy for each model

Sample 500 Dnase sites, and compare different models in prediction accuracy for the sampled sites

Logistic regression with Lasso regularization gives the highest average prediction accuracy in sampled 500 sites



Summary and future directions

➢Summary

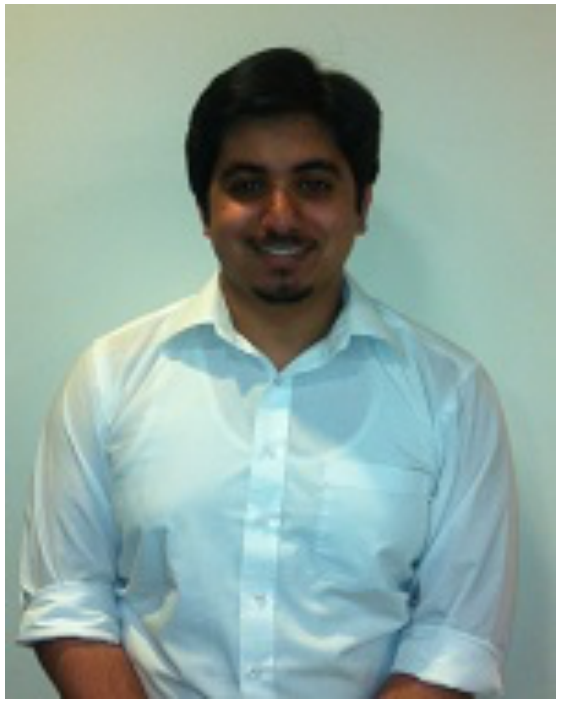
- We preprocessed and combined RNA-seq and Dnase-seq data for 19 tissues.
- We incorporate adjacent genes for chromatin accessibility prediction using logistic regression with lasso regularization, and increase the prediction accuracy from 70% in null model to 80%.

➢Future directions:

- Fit the logistic regression model (Lasso regularization) on all 19 tissues
- Validate the model on the test data set (RNA-seq and Dnase-seq paired data set from ENCODE project)

Acknowledgement

Ali Shojaie



Asad Haris