smILIes
reSilient coMputer archItectures and LIfE Sciences

Politecnico di Torino
Department of Control and Computer Engineering

DAUIN

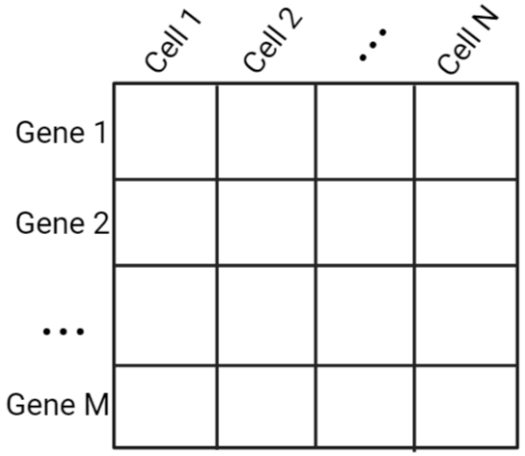
GAGAM: GENOMIC-ANNOTATED GENE ACTIVITY MATRIX FROM scATAC-SEQ DATA

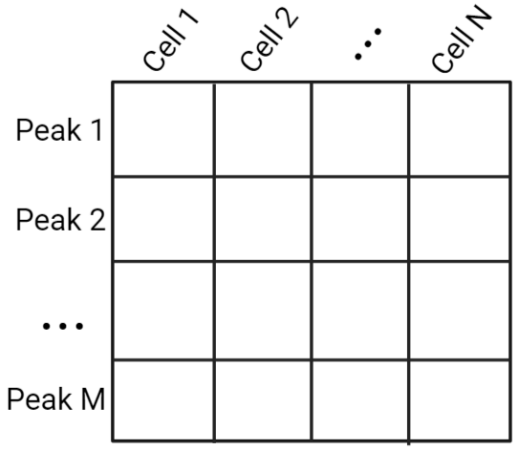
MARTINI LORENZO, BARDINI ROBERTA, SAVINO ALESSANDRO, DI CARLO STEFANO

INTRODUCTION

- Next Generation Sequencing Multi-Modal technologies are becoming particularly relevant

- ScRNA-seq Transcriptomic + scATAC-seq Epigenomics

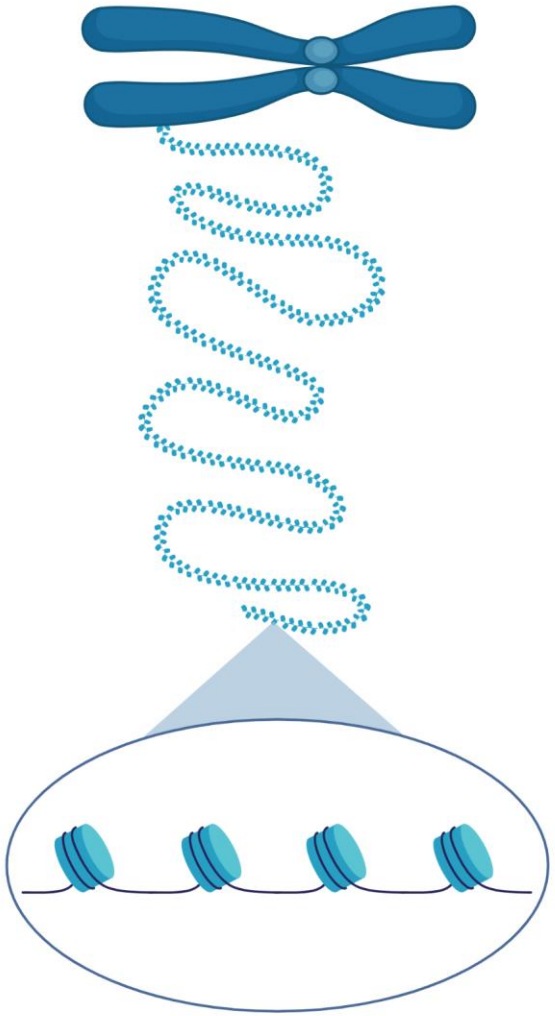




- Necessity to link accessible regions (peaks) and expressed genes

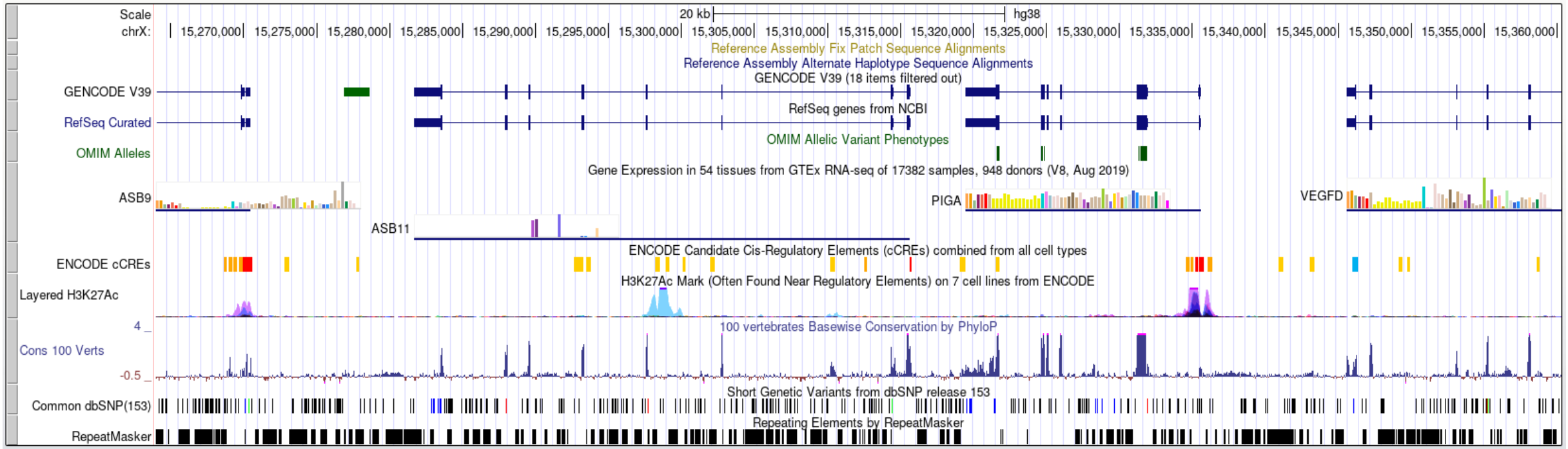
GENE ACTIVITY MATRIX

- Creating a Gene x Cells matrix from scATAC-seq data
- Recapitulates the “activity” of the genes
  - Activity is the overall accessibility of the genomic region of a gene
- There are multiple ways to define it (i.e., Cicero, Gene Scoring)
  - Look at gene body
  - Can be quite simplistic
  - The gene expression regulation is not considered



GAGAM: GENOMIC-ANNOTATED GAM

- Employ the information from the detailed and curated genomic annotations (i.e., UCSC)

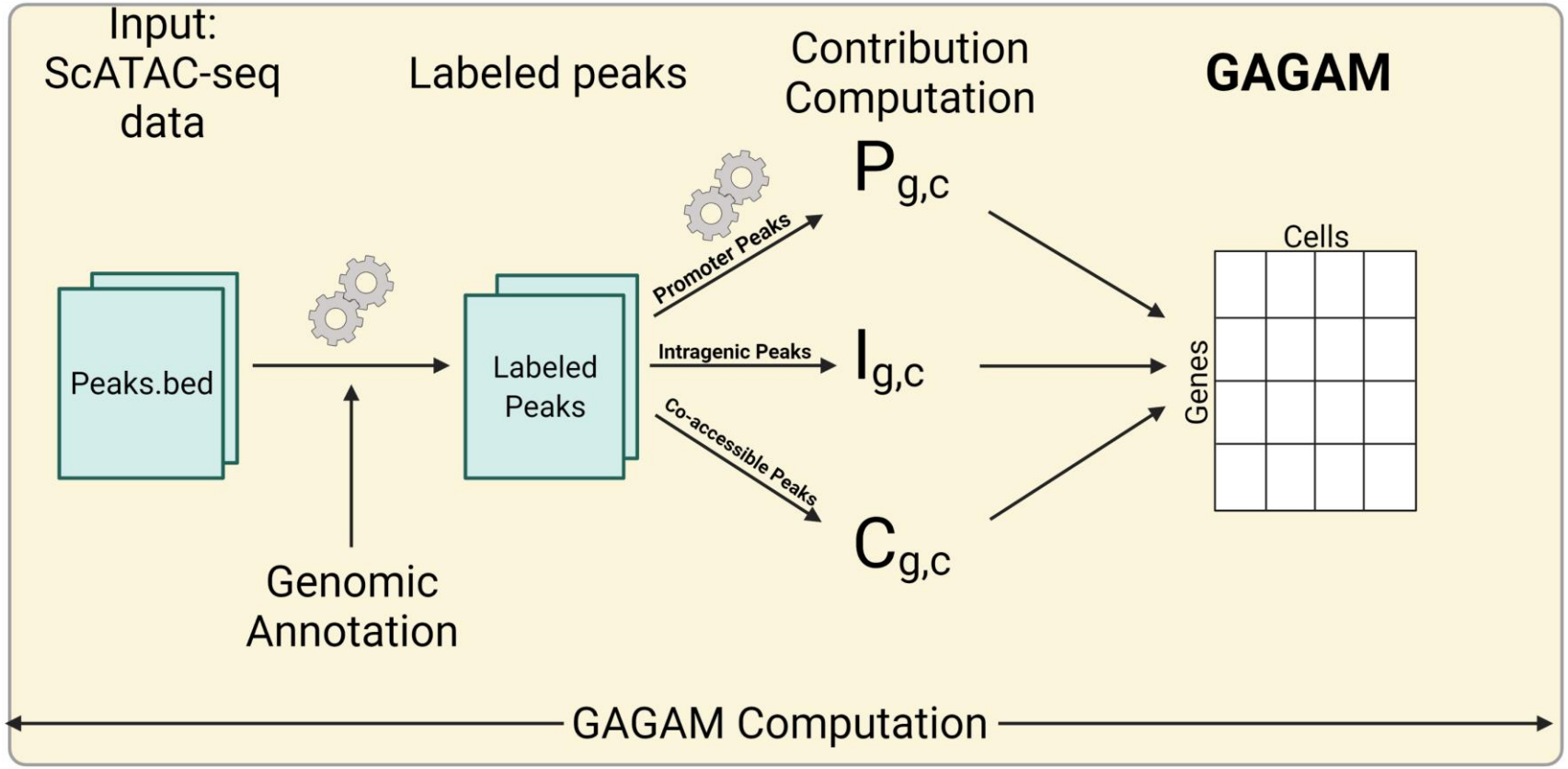


GAGAM: GENOMIC-ANNOTATED GAM

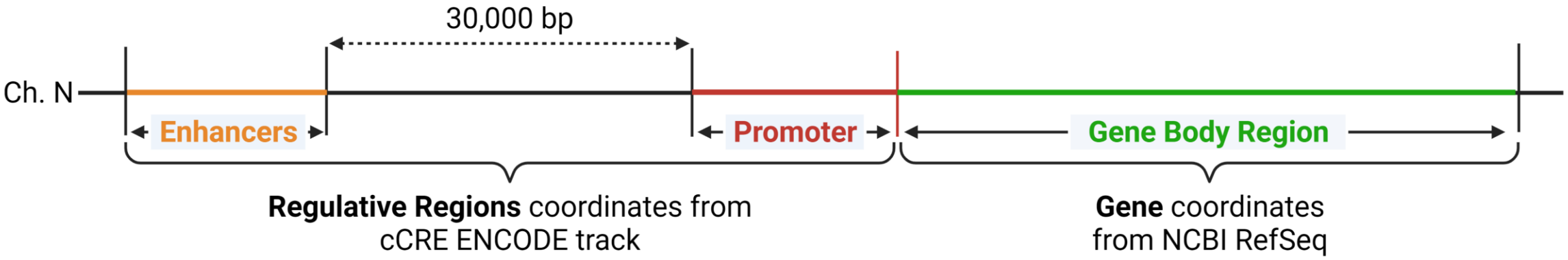
- Employ the information from the detailed and curated genomic annotations (i.e., UCSC)

GAGAM: GENOMIC-ANNOTATED GAM

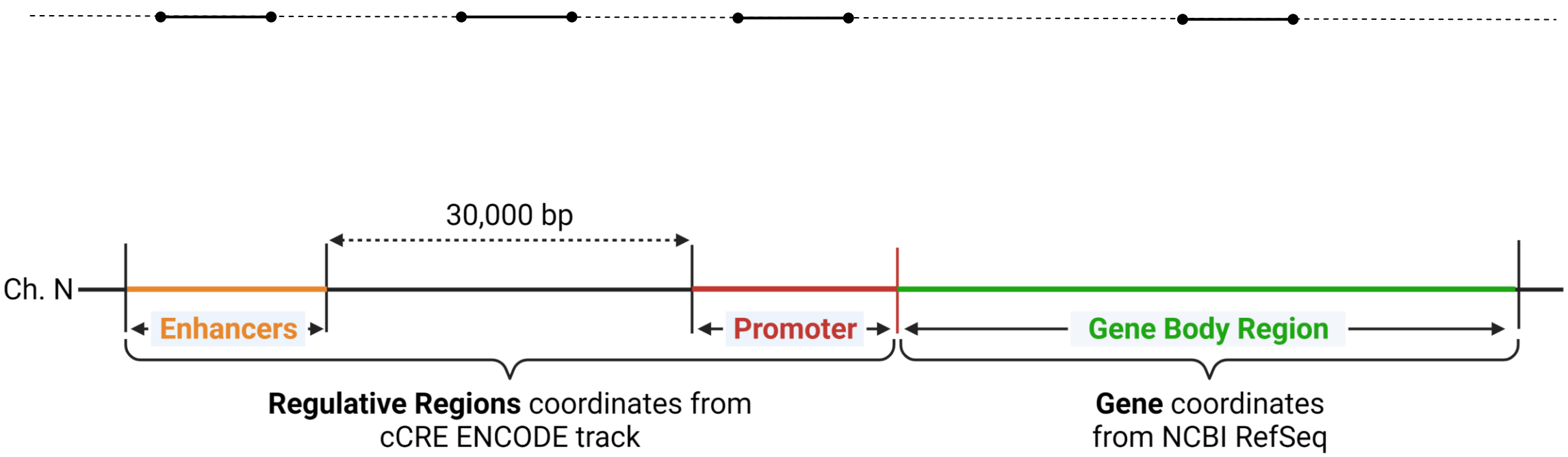
- Employ the information from the detailed and curated genomic annotations (i.e., UCSC)



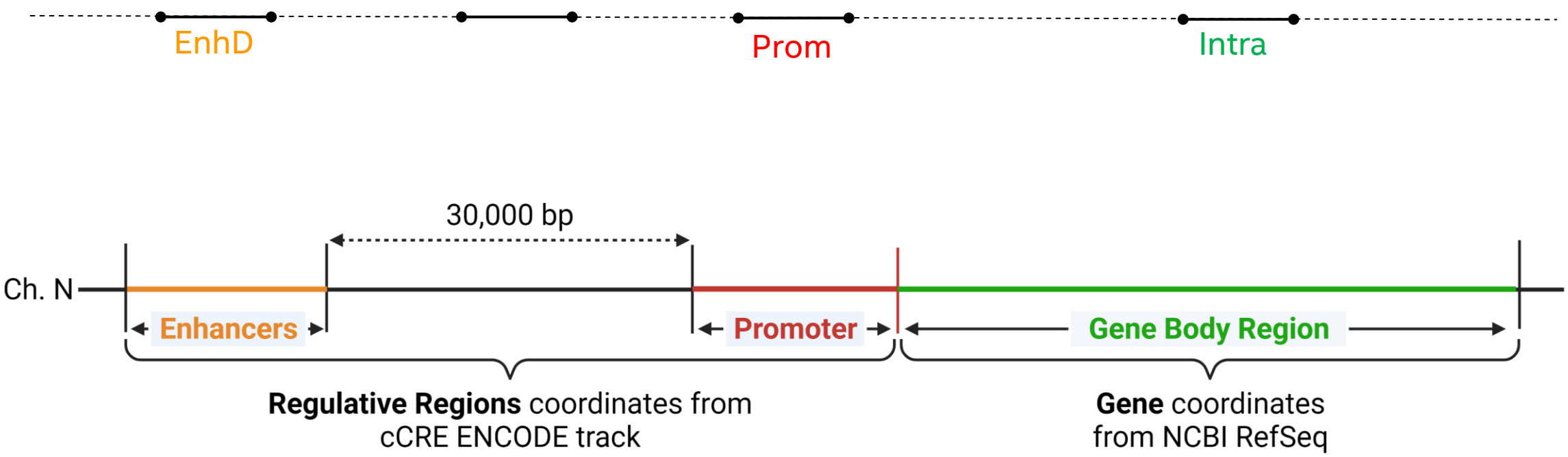
GENOMIC ACCESSIBILITY MODEL



GENOMIC ACCESSIBILITY MODEL



GENOMIC ACCESSIBILITY MODEL



MATRIX CONTRIBUTIONS

$P_{g,c}$ Promoter peaks

- Binary matrix
- Accessibility of the genes' promoters

$I_{g,c}$ Intragenic peaks

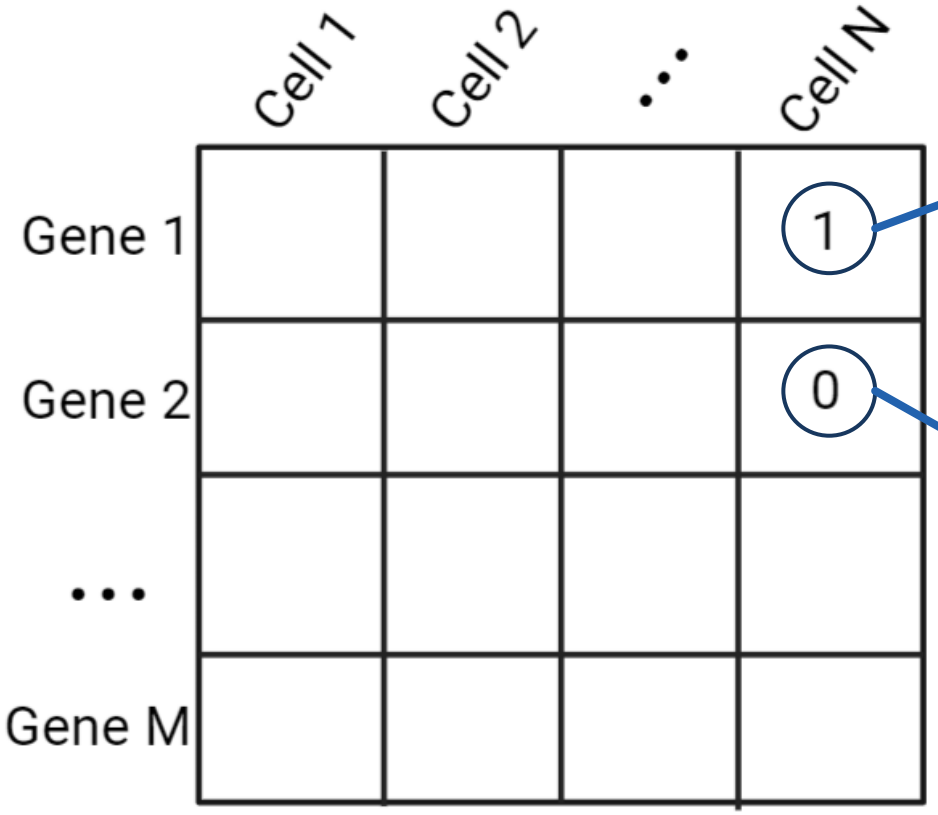
- Overall accessibility of genes body
- Least model driven contribution

$C_{g,c}$ Promoter-enhancer co-accessibility

- Accessibility of enhancers linked to promoters

PROMOTER PEAKS $P_{G,C}$

- A gene is active if its promoter is accessible

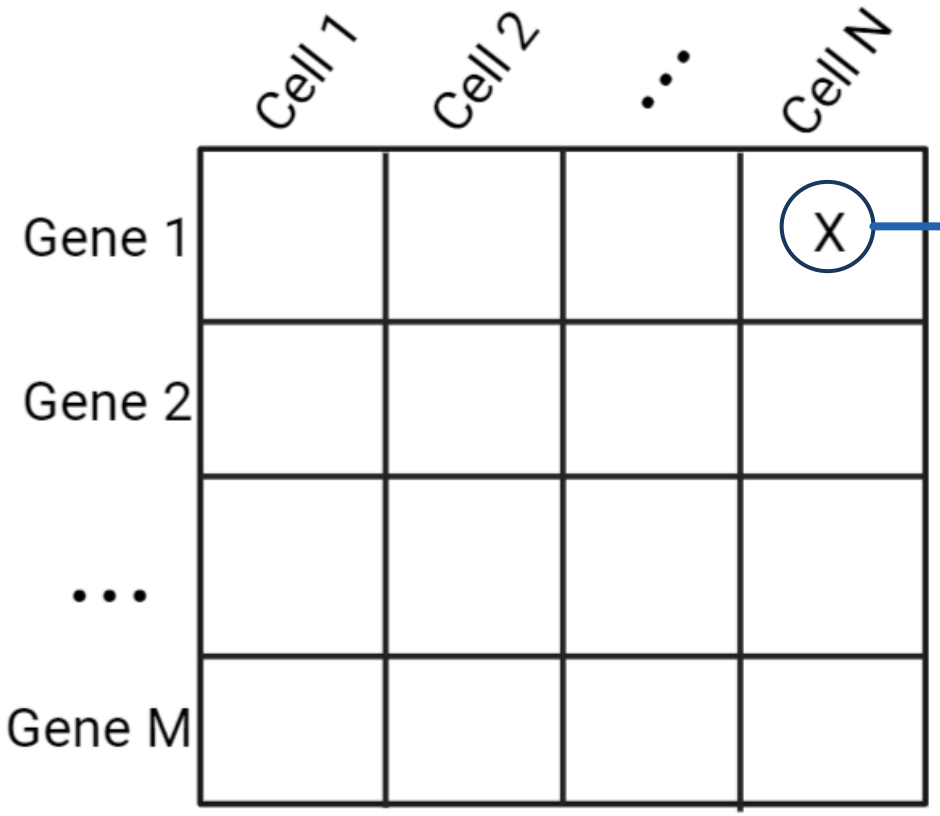


Gene 1 has a peak on its promoter region and it is accessible in Cell N

Gene 2 has a peak on its promoter region but it is not accessible in Cell N

INTRAGENIC PEAKS $I_{G,C}$

- Contribution from the peaks inside the gene body



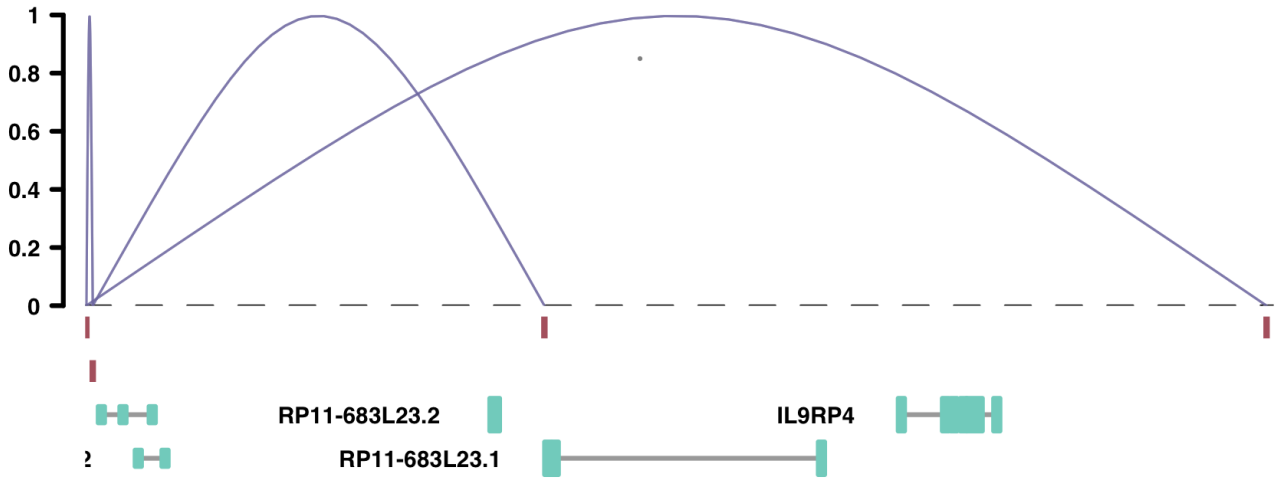
Gene 1 has a contribution X from the intragenic peaks (i.e., the ones overlapping the gene body)

X is the sum of all peaks in the gene body weigthed by the distance from the Transcription Starting Site (TSS)

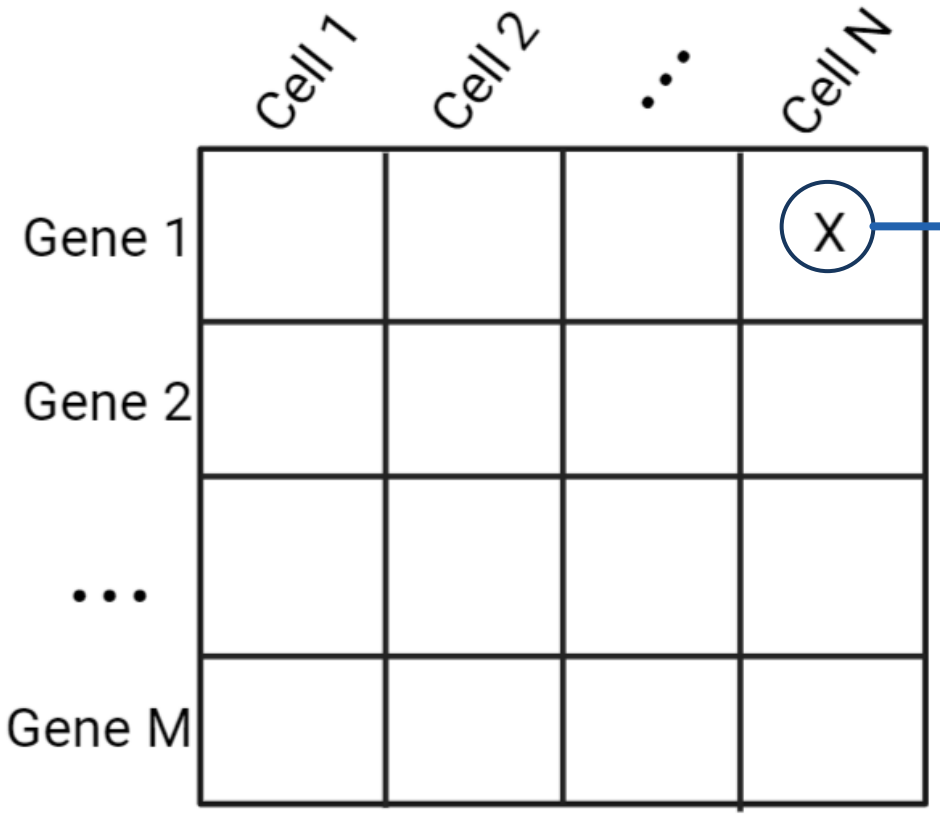
PROMOTER-ENHANCER CO-ACCESSIBILITY $C_{G,C}$

- Cicero allows calculating the co-accessibility between couples of peaks
  - How much couples of peaks tend to be simoultaneously accessible
  - It is a way to model the gene regulation

- If a peak labeled as enhancer is co-accessible with a promoter, it can be viewed as a promoter-enhancer relationship



PROMOTER-ENHANCER CO-ACCESSIBILITY $C_{G,C}$

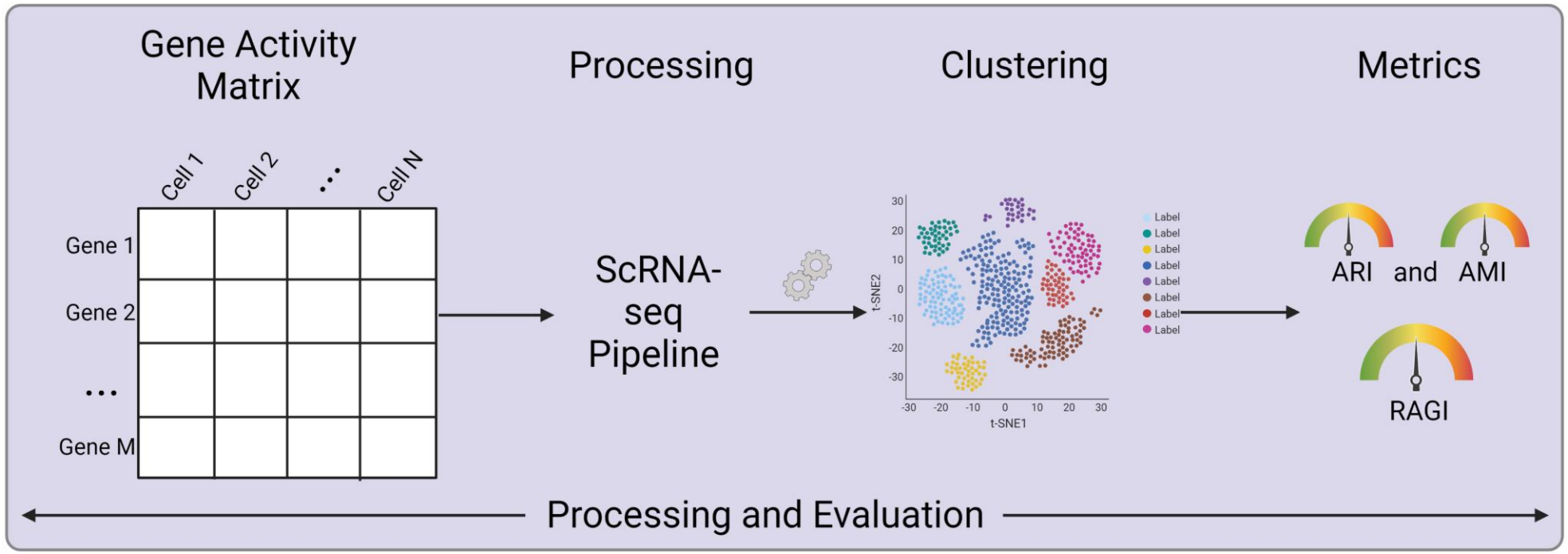


Gene 1 has a contribution X from the enhancer peaks related to its promoter that are accessible in Cell N.

X is the sum of all co-accessibility scores for the enhancer peaks

EVALUATION

- Compare GAGAM to Cicero and Gene Scoring GAMs performances on different datasets
- GAGAM1 (all three contributions), GAGAM2 (without intragenic contribution) RIF

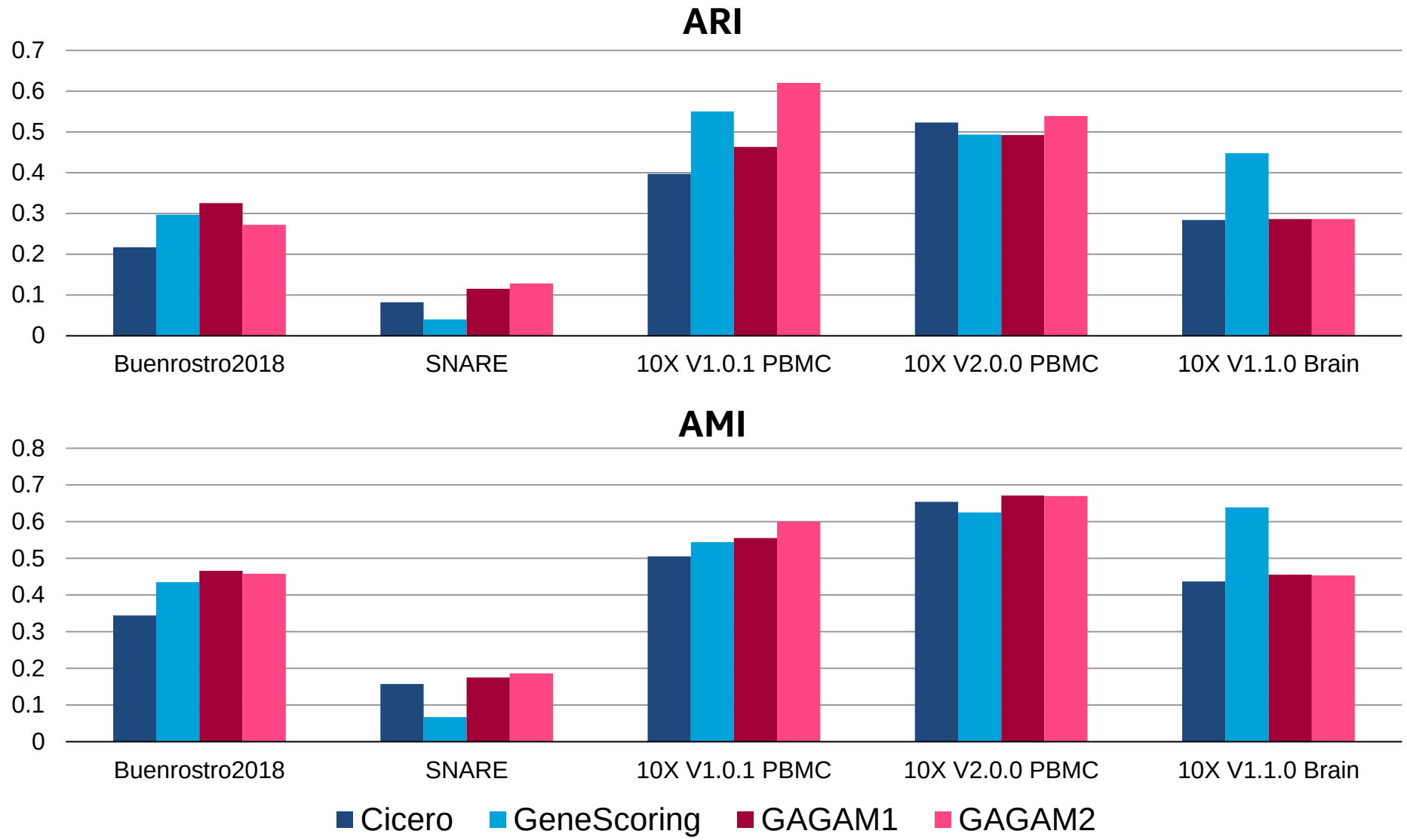


EVALUATION

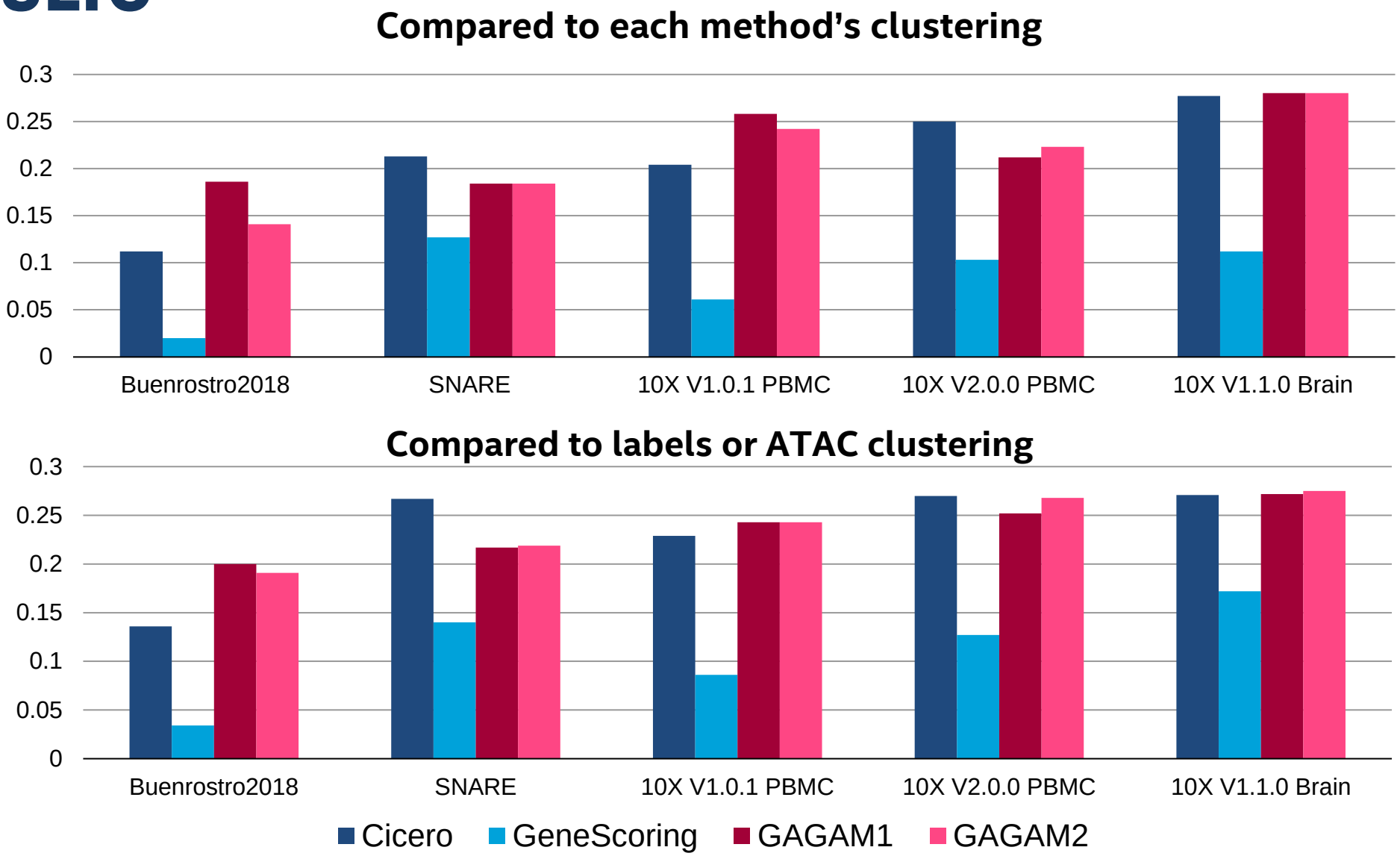
- Adjust Rand Index (**ARI**) and Adjust Mutual Information (**AMI**)
  - Used to **evaluate clustering** with respect to some **ground-truth classification**
  - Cell-type labels or clustering performed on raw scATAC-seq data

- Residual Average Gini Index (**RAGI**)
  - **Difference of Gini Index** between **markers and housekeeping** genes
  - It is also an evaluation on the quality of the GAM itself
  - Performed on the clusters obtained and the ground-truth classification

RESULTS


ARI
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Buenrostro2018
SNARE
10X V1.0.1 PBMC
10X V2.0.0 PBMC
10X V1.1.0 Brain
AMI
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Buenrostro2018
SNARE
10X V1.0.1 PBMC
10X V2.0.0 PBMC
10X V1.1.0 Brain
Cicero
GeneScoring
GAGAM1
GAGAM2

RESULTS


Compared to each method's clustering
0.3
0.25
0.2
0.15
0.1
0.05
0
Buenrostro2018
SNARE
10X V1.0.1 PBMC
10X V2.0.0 PBMC
10X V1.1.0 Brain
Compared to labels or ATAC clustering
0.3
0.25
0.2
0.15
0.1
0.05
0
Buenrostro2018
SNARE
10X V1.0.1 PBMC
10X V2.0.0 PBMC
10X V1.1.0 Brain
Cicero
GeneScoring
GAGAM1
GAGAM2

CONCLUSIONS

GAGAM is a new and model-driven way to compute Gene Activity Matrix

It employs genomic annotation to label the peaks

Its performances are consistent throughout all the metrics

Advancements can be done by better tuning the already existing contributions and by adding other genomic information in the computation

Promising point of view

THANK YOU VERY MUCH FOR THE ATTENTION

- Contacts:

  lorenzo.martini@polito.it

  SMILIES Polito on LinkedIn