Gene Regulatory Network Inference Using Single-cell Multiome ATAC-seq and RNA-seq Data

Yuge Wang, Zhongyu Cai, Keran Chen, Hongyu Zhao

A mid-term summary of Zhongyu Cai's internship in Zhao Lab

Zhongyu Cai

10/10/2022

Outline

- Famework of **GRN Inference** using single-cell **multiome ATAC-seq and RNA-seq data**
  - Models for GRN inference:
    - Model 1 **non-filter**, Model 2 **filter**, Model 3 **multiply**, Model 4 **score**
    - most straightforward → use **open chromatin regions** for collecting candidate TFs → incorporate chromatin accessibility information → incorporate TF-peak binding score
  - Model evaluation: AUPRC & Comparison line chart
- Results to Date
  - Main Results
  - Other details
- Recent Work
  - Use new data: BMMC
  - Results were inconsistent with previous conclusion
  - New conjecture and verification

Famework of Gene Regulatory Network Inference

- To build a **Gene Regulatory Network**:
  - Select candidate transcription factors **(TFs)** for a certain gene **(target gene)**
  - Determine the **weight** of each **TF-target gene edge**

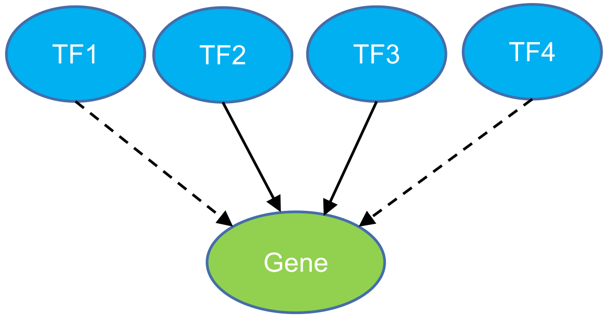


Fig 1. Select candidate TFs for a target gene

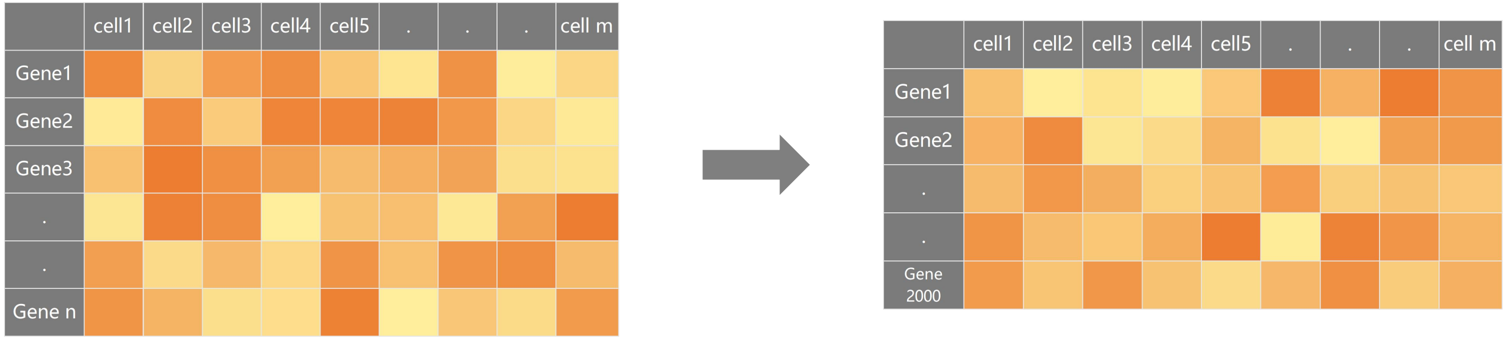


Fig 2. Select 2000 highly variable genes from PBMC gene expression matrix

- **Data**: PBMC （Peripheral blood mononuclear cell）
- **Target gene set**: select **2000 highly variable genes** from PBMC gene expression matrix
- **TF set**: cis-BP

MODEL 1: The most straightforward model (previous one)

- **Model 1 (non-filter):** Use **gene expression level for all TFs** and the **target gene** to build a regression tree using GENIE3.

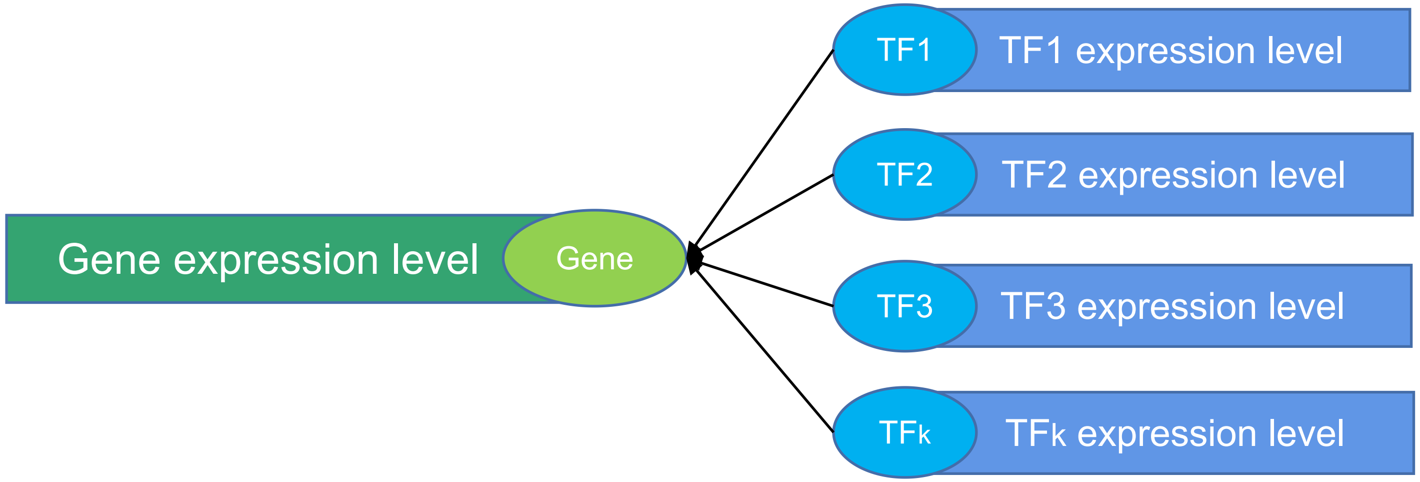
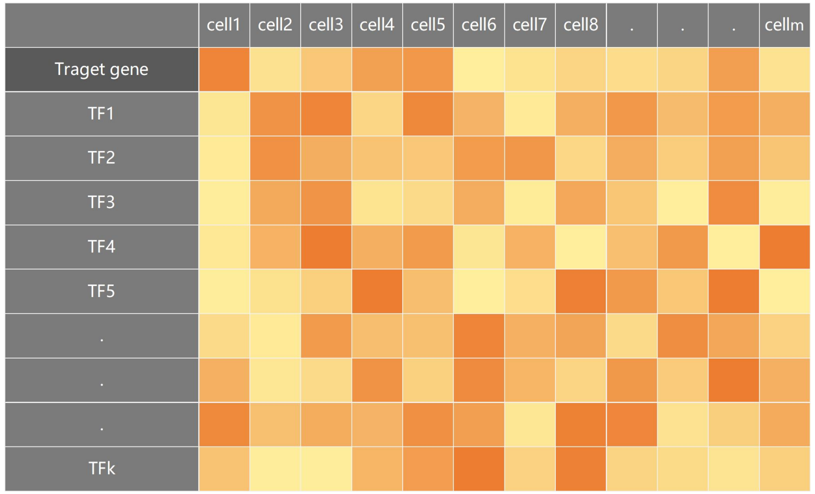


Fig 3. Use gene expression level for all TFs and the target gene to build a regression tree



	cell1	cell2	cell3	cell4	cell5	cell6	cell7	cell8	.	.	.	cellm
Traget gene												
TF1												
TF2												
TF3												
TF4												
TF5												
.												
.												
.												
TFk												

Fig 4. Input matrix of model 1 for GENIE3

- **Limitations:**
  - Regulation does not only rely on co-expression but also on the binding of TFs on nearby regulatory regions
- **Improving direction**:
  - Incorporate information of nearby regulatory regions

How to Incorporate Information of Nearby Regulatory Regions?

- Use **important open chromatin regions** to select some **candidate TFs** for a certain **target gene** before building a tree based model

- **TF** regulates target genes by **binding on** transcription factor binding sites **(TFBS)** around the transcription starting site **(TSS)** of the **target gene**

- Find candidate TFs for every target gene:
    - Select **important chromatin regions** around the TSS of a gene as promoters
    - **Pair TFs with candidate ATAC peaks** by applying TF binding site analysis tools

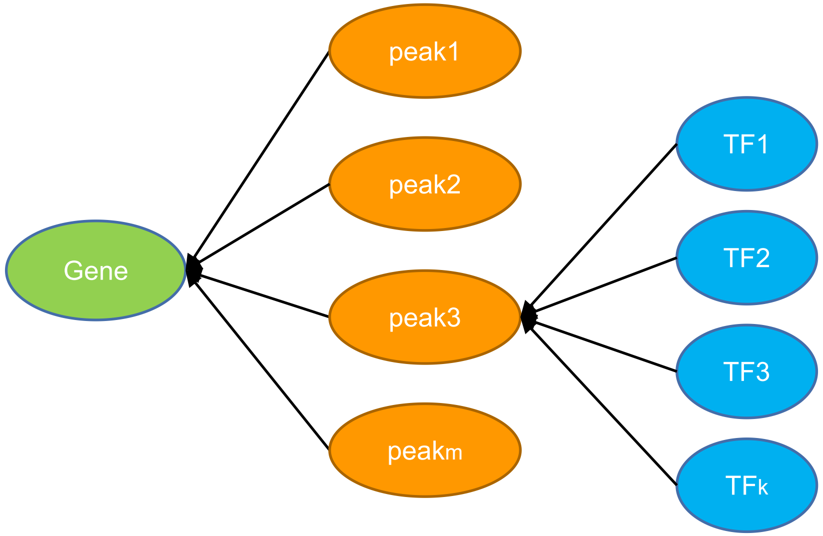


Fig 5.Use ATAC peaks as a bridge to find candidate TFs for each target gene

Collect candidate TFs for target gene: Collect peaks for a target gene & pair TFs with peaks

- **For each target gene**, there are two choices to collect **ATAC peaks**:
  - Collect **ATAC peaks** within **500 kb around the TSS** of a target gene **(500kb)**
  - Collect **ATAC peaks** within **500 kb around the TSS** of a target gene **&** only retain gene-peak links with a **correlation coefficient** above **a certain threshold (LinkPeaks)**

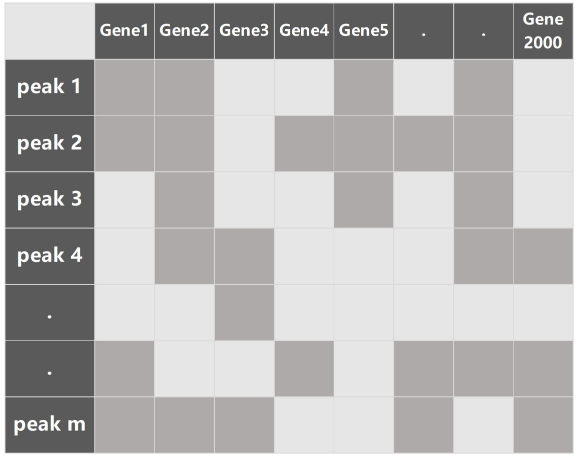


Fig 6. Generate a peak by Gene binary matrix by collecting ATAC peaks for genes

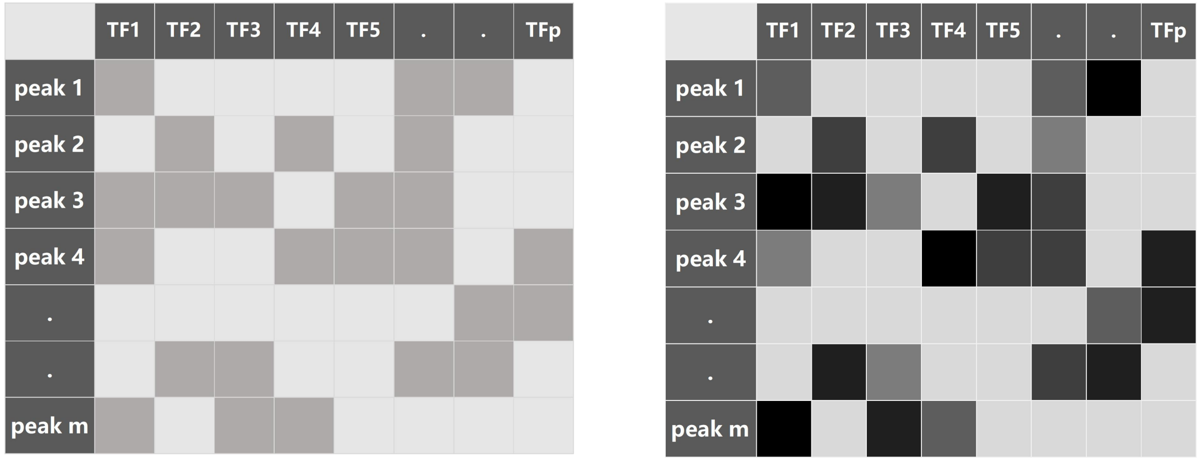


Fig 7. Generate a peak by TF binary matrix and a score matrix by pairing ATAC peaks and TFs

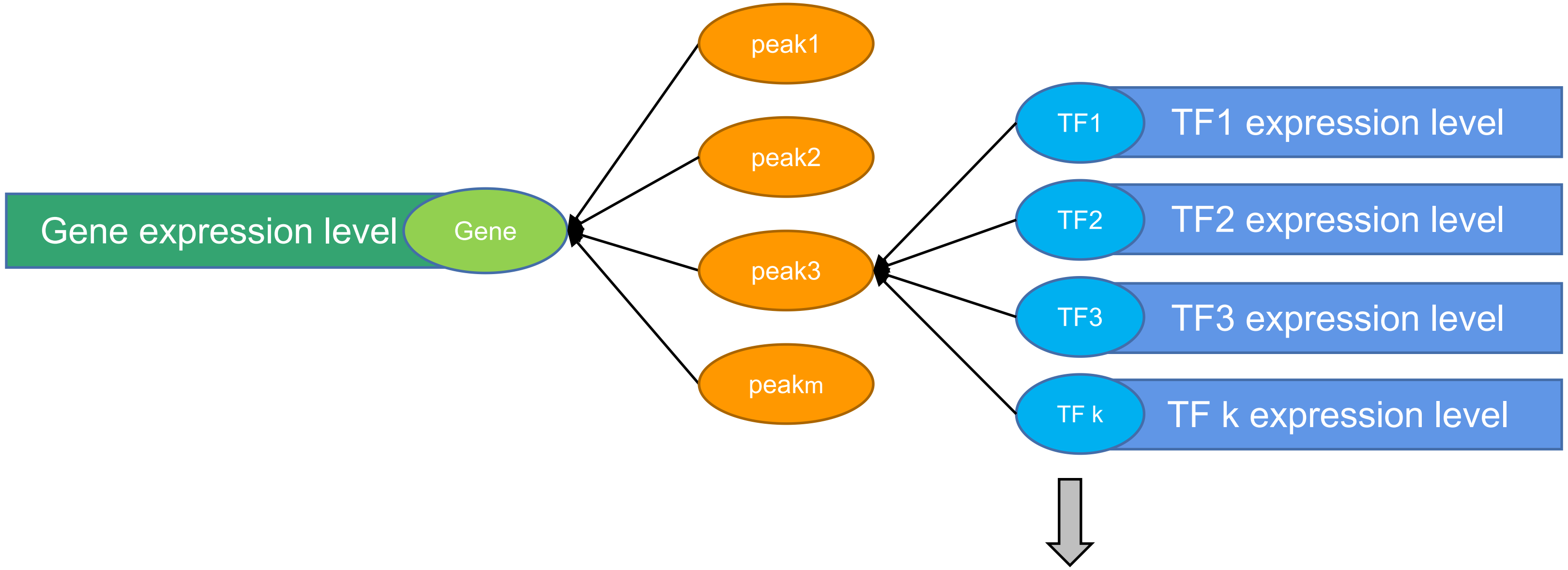
- **For each peak**, a base sequence is provided. (chr1-102938401-102938550: ACTGAGTGATC...ATAGCATGC)
- **For each TF**, a position frequency matrix (**PFM**) is provided.

$$\begin{array}{c|cccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline \mathrm{A} & 0.24 & 0.10 & 0.45 & 0.27 & 0.49 & 0.15 & 0.45 & 0.31 \\ \mathrm{T} & 0.03 & 0.28 & 0.41 & 0.19 & 0.42 & 0.41 & 0.39 & 0.22 \\ \mathrm{C} & 0.26 & 0.40 & 0.05 & 0.23 & 0.01 & 0.35 & 0.05 & 0.19 \\ \mathrm{G} & 0.47 & 0.22 & 0.09 & 0.31 & 0.08 & 0.09 & 0.11 & 0.28 \end{array}$$

- **For each TF-ATAC peak pair**, a score can be calculated to **indicate the binding intensity of this pair**
- There are two choices to collect **TFs** for each **peak**: FIMO & motifmatchr

MODEL 2: use open chromatin regions for collecting candidate TFs

- **Model 2 (filter):** Use **gene expression level for candidate TFs** and the **target gene** to build a regression tree using GENIE3.



TFs in at least one gene-peak-TF relationship

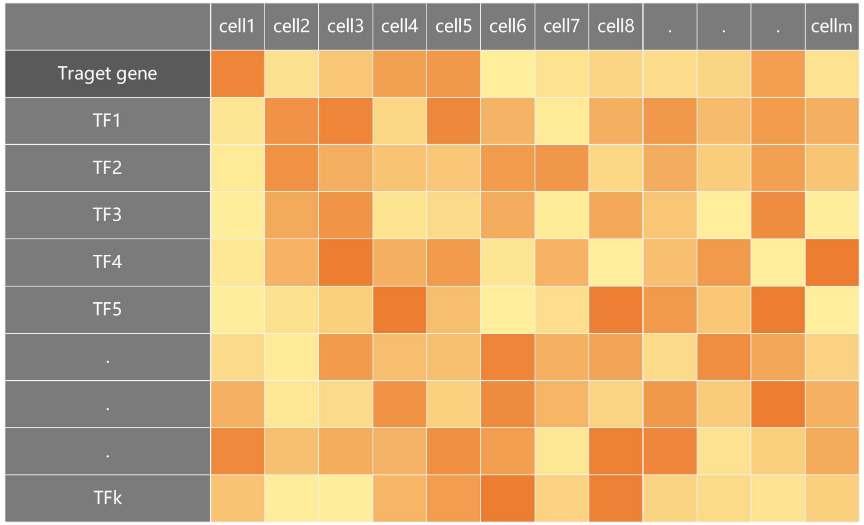


Fig 8. Use gene expression level for candidate TFs and the target gene to build a regression tree

Fig 9. Input matrix of model 2 for GENIE3

- **Limitations: Gene regulation** also involve **the accessibility of important chromatin regions** around TSS.
- Improving direction: Incorporate information from **chromatin accessibility** into tree-based model

MODEL 3: incorporate chromatin accessibility information

- **Model 3 (multiply):** Use **gene exprssion level** **multiplied by** **peak accessiblity** for each **TF-peak pair** and **the expression level** of **target gene** to build a regression model using GENIE3.

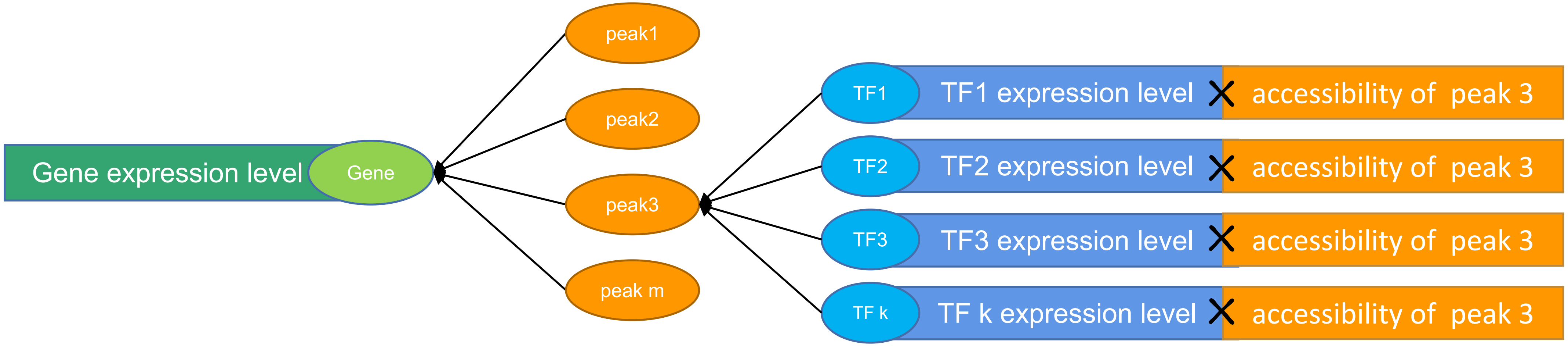


Fig 10. Use gene exprssion level multiplied by peak accessiblity for each TF-peak pair and the expression level of target gene to build a regression model

- For each target gene, the **weight for each TF** is calculated as the summation of the importance scores of all **TF-ATAC pairs that involve the TF**.

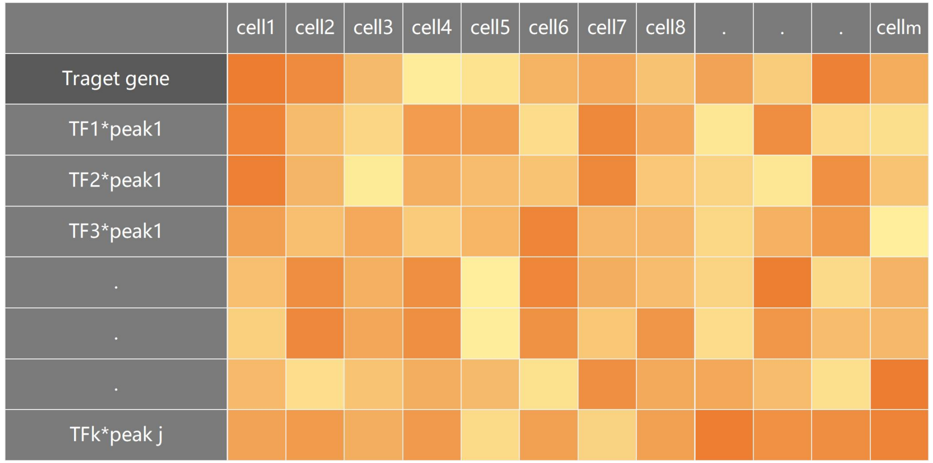


Fig 11. Input matrix of model 3 for GENIE3

MODEL 4: incorporate TF-peak binding score

- **Model 4 (score):** Use **gene exprssion level times peak accessiblity times binding score** for **each TF-peak pair** and the target gene to build a regression model using GENIE3

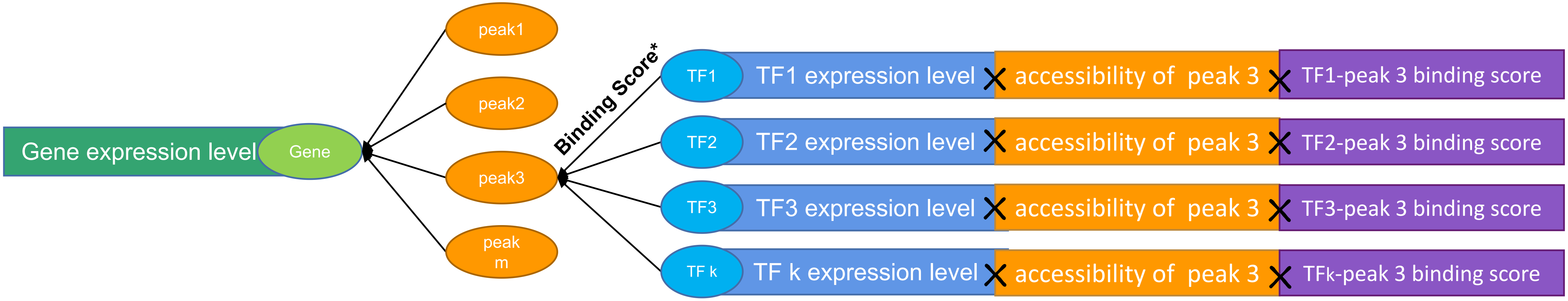


Fig 12. Use gene exprssion level multiplied by peak accessiblity for each TF-peak pair and the expression level of target gene to build a regression model

***Binding Score =** $\varphi^{-1}(1-\frac{p.value}{2})$

- For each target gene, the **weight for each TF** is calculated as the summation of the importance scores of all TF-ATAC pairs that involve the TF.

- **Results**: a list of TF-gene pairs with weight (importance)

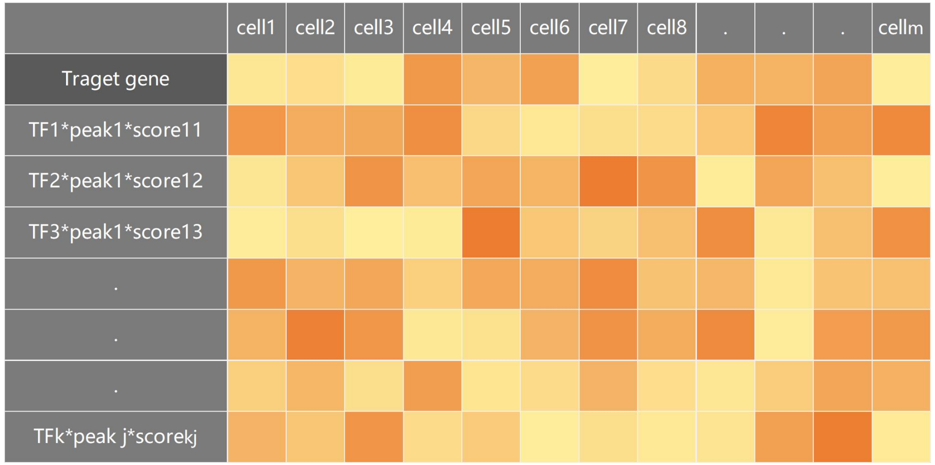


Fig 13. Input matrix of model 4 for GENIE3

Model Evaluation: 2 methods

- **True network**: GRNs from existing databases
- Two methods are used for **model evaluation**:
  - Area Under Precision-Recall Curve(**AUPRC**)
  - Comparison line chart: compare the **true positives** in **top k edges** of two predicted network

    (True positives in the top k edges of predicted network 1) - (True postives in the top k of predicted network 2)

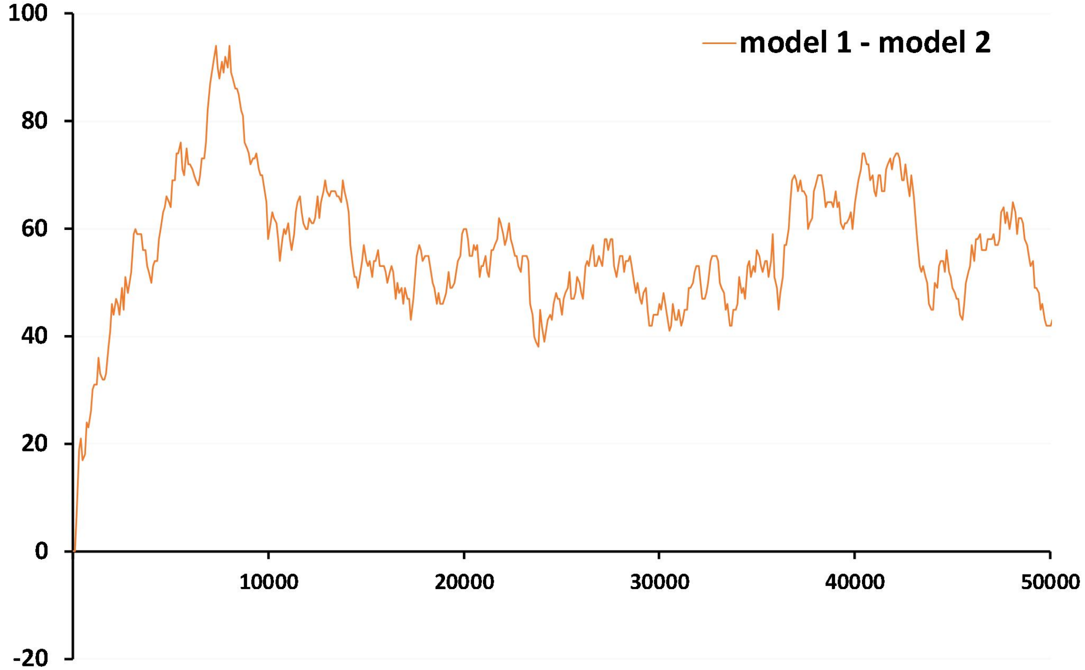
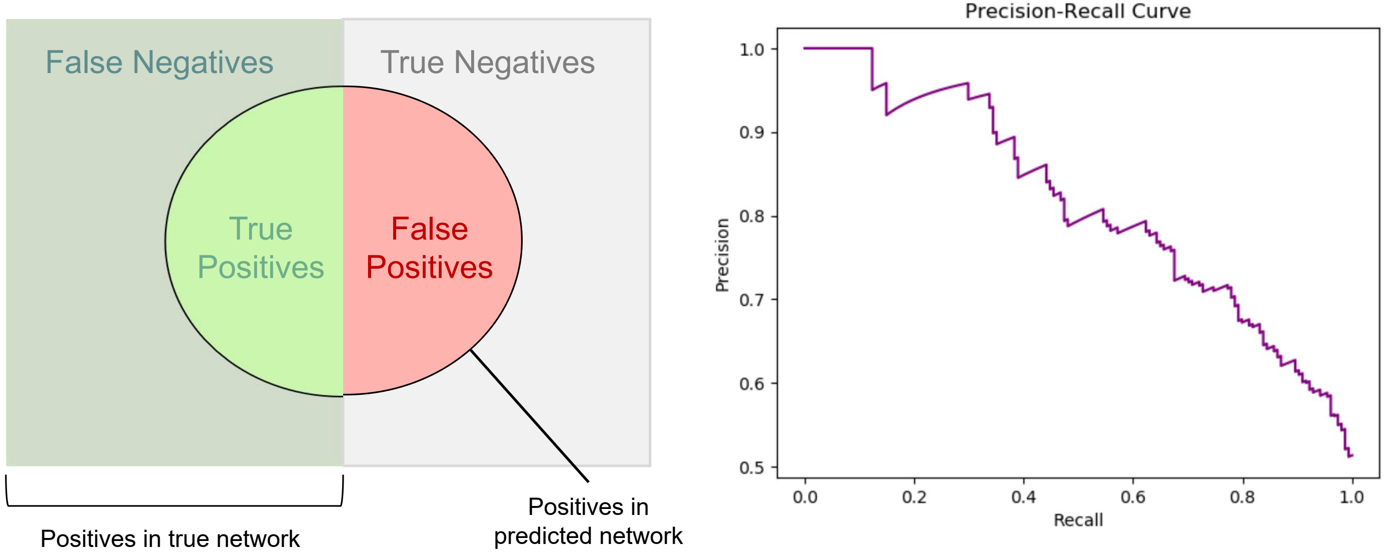


Fig 14. Model performance comparison between model 1 and model 2 which indicate model 1 outperforms model 2



$$precision: \frac{True\ Positive}{True\ Positive+False\ Positive}$$

$$recall: \frac{True\ Positive}{True\ Positive+False\ Negative}$$

Fig 15. The calculation principle of AUPRC, precision and recall

Model Evaluation: background for precision and recall

- Choices **for the background**:
  - Target Genes in **true network** × TFs in **true newtork**: edges involving

    TFs and genes only in true network will all be marked as **negative** in **predicted network**
  - Target Genes in **predicted network** × TFs in **predicted newtwork**:

    edges involving **TFs and genes only in predicted network** will all be marked as **negative in true network**
  - Target Genes **in both network** × TFs in **both network**

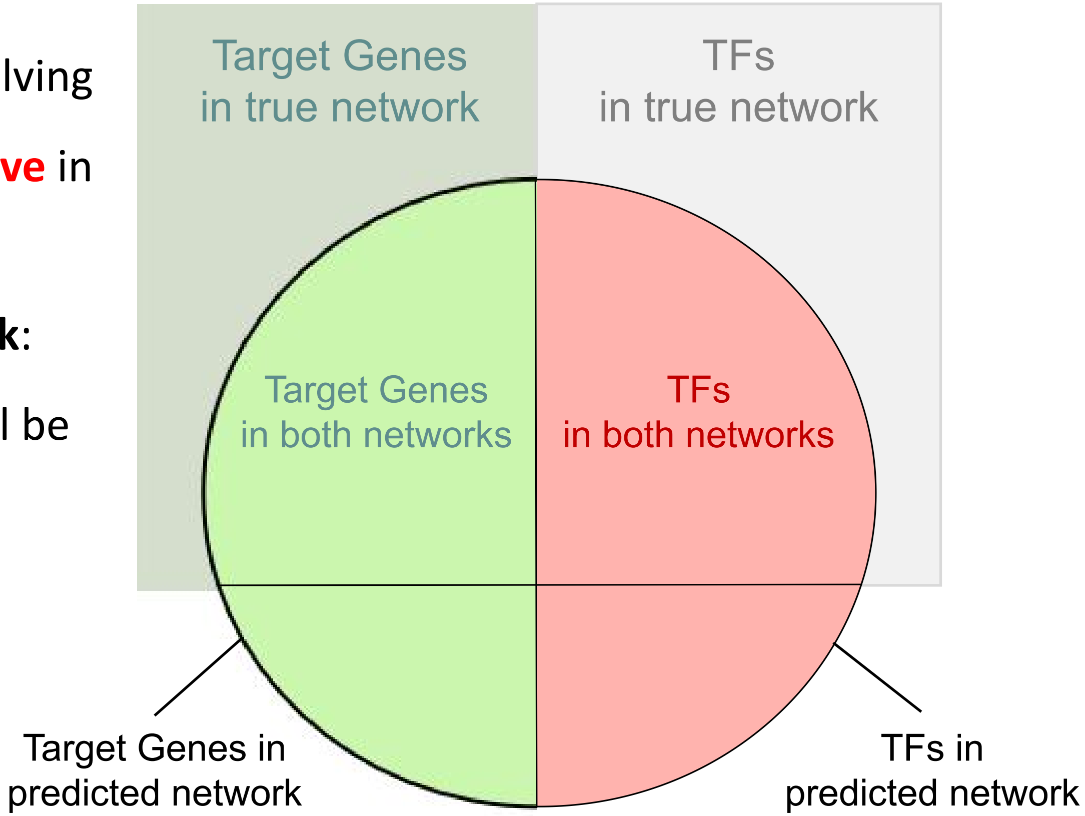


Fig 16. Relationships between target gene set and TF set in predicted and true network

Model Evaluation: background for precision and recall

- If setting background as Target Genes **in both network** × TFs **in both network**, it will result in **inconsistent baselines** of different models because **target genes sets** are different between models.
- model 1 non-filter: Target Genes are **all 2000 highly variable genes**
- model 2&3&4: Target Genes are those who can **be paired with at least 2 candidate TFs**
- **Target Gene only in model 1** can be divided into three parts:
    - genes whose **coordinates are missed** because of lack of annotations
    - genes whose candidate TF set is **empty**
    - genes which have **only one candidate TF**

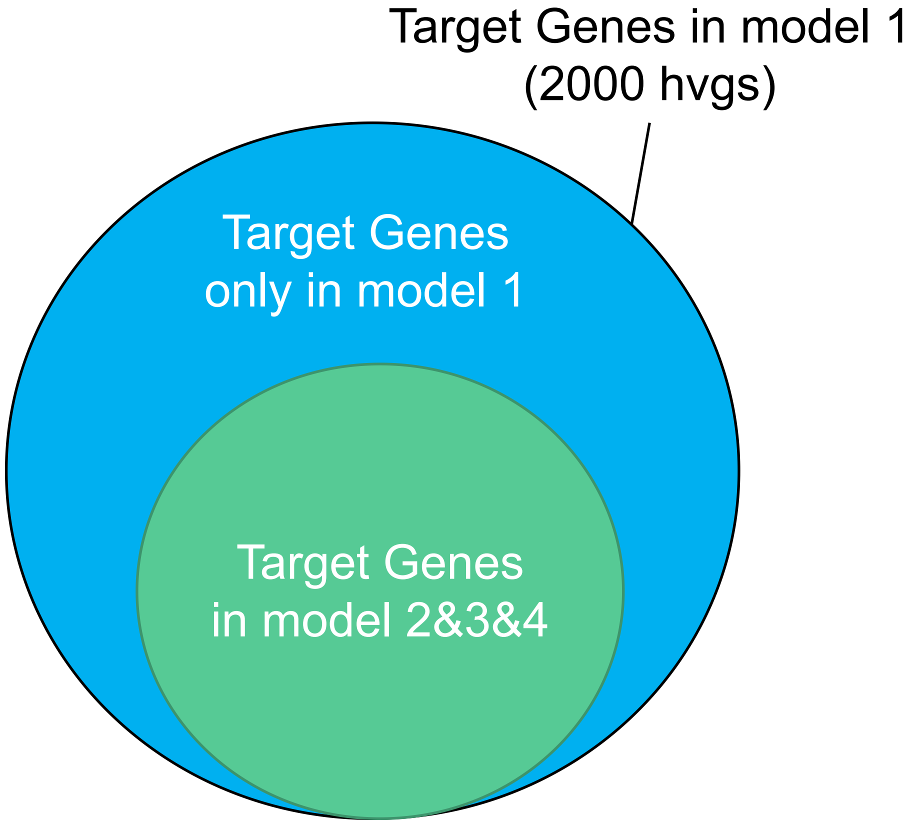


Fig 17. Relationship between target gene set of different models

Model Evaluation: background for precision and recall

- **Previous work**: number of genes in model 2&3&4 is **~1% less than** that in model 1
- **Recent work on BMMC**: **>50% genes** are filtered out because of **pairing with no TF**
- Only **consider target genes with annotations** in all models
- Unify background of different models:
  - **background = highly variable genes with annotations in true network × TFs in both network**

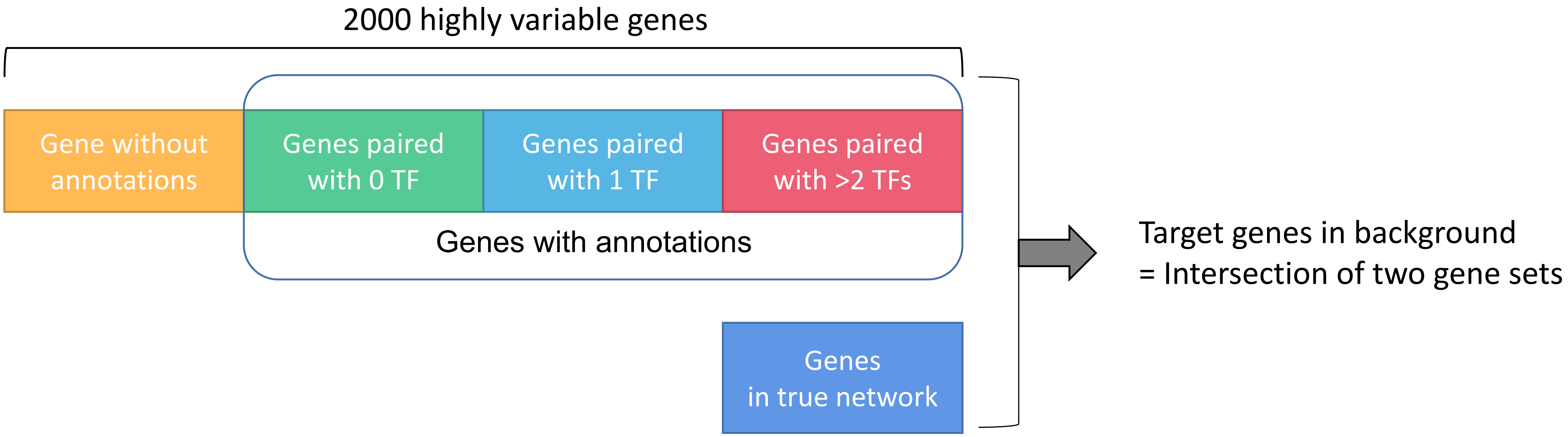


Fig 18. Unify background of different models by using the same target gene set

Incorporating peak accessibility information can improve the model performance

- Model 3 **multiply** ≈ Model 4 **score** > Model 1 **non-filter** > Model 2 **filter**
- AUPRC (500kb + motifmatchr):
  Model 1: 0.02571 Model 2: 0.02531
  Model 3: 0.02609 Model 4: 0.02607
- Incorporating **peak accessibility information** can improve True Positives

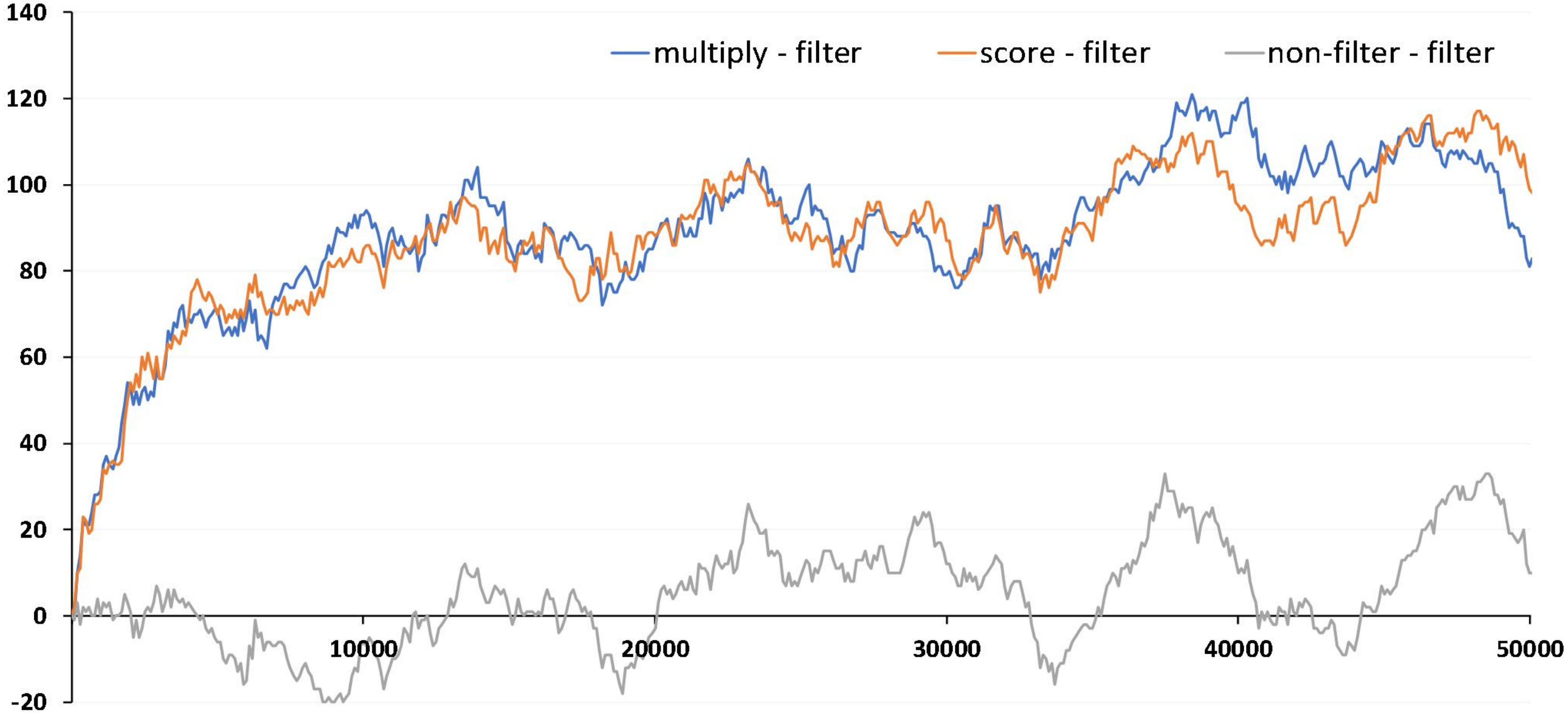


Fig 19. Model performance comparison for peak-gene pairing based on location only (500kb) and peak-TF pairing based on motifmatchr

Incorporating peak-gene correlation can improve the performance

- Collect **ATAC peaks** within **500 kb around the TSS** of a target gene, and only preserve peak-gene pairs that has **a significant correlation**. (LinkPeaks)
- A significant improvement in True Positives
- A significant improvement in **time and memory (over 50%)**

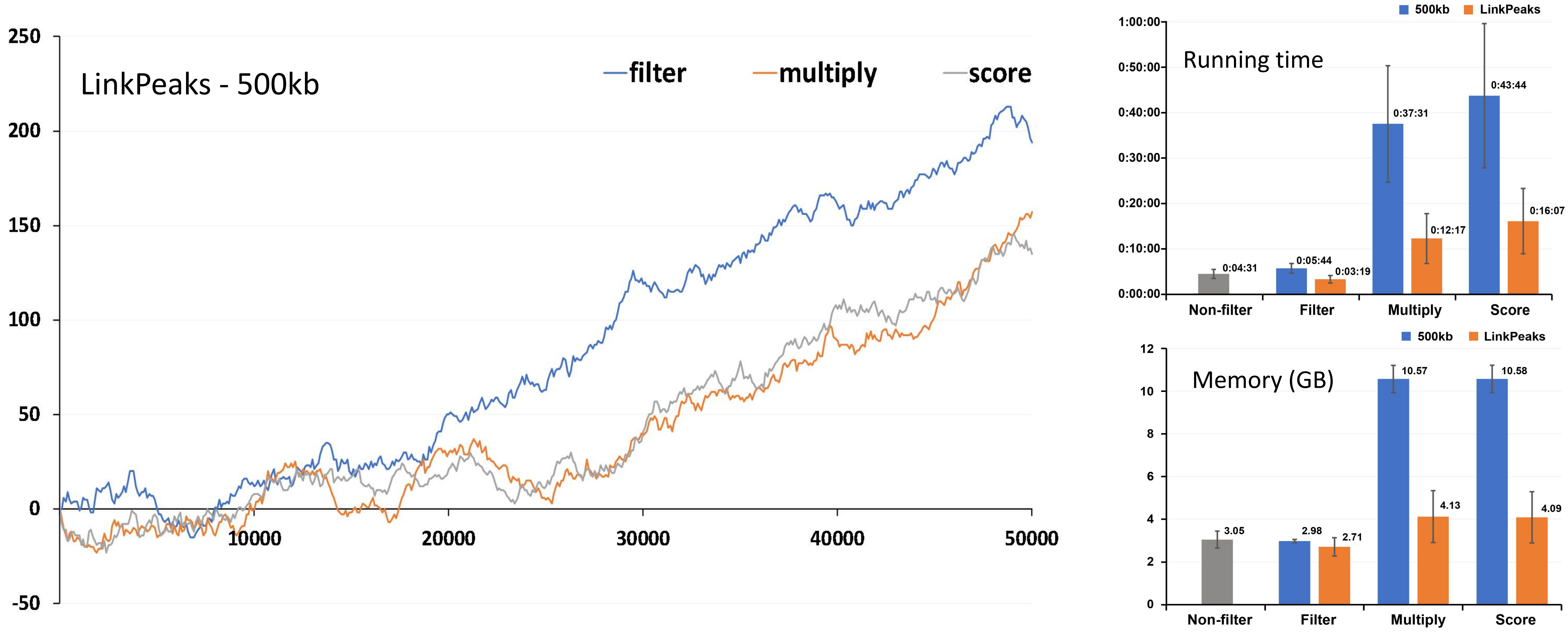


Fig 20. Model performance comparison between peak-gene pairing based on location only (500kb) and based on correlation strength (LinkPeaks)

Other Details

- **Improvements** in other details:
  - whether to use **motifmatchr** or **FIMO** for **peak-TF pairing (motifmatchr)**
  - whether to use **all gene-peak pairs (with correlation both positive and negative)** that pass the p.value test or only preserve those with **positive correlation** (use all gene-peak pairs)
  - what the **best thresholds** for **p.value and score** of gene-peak pairing are (0.05, 0.05)
  - whether to use **q.value** or **p.value** to filter out TF-peak pairs (p.value)
  - whether to use **normalized data** or **count data** to build a tree-based model (count data)
  - …………

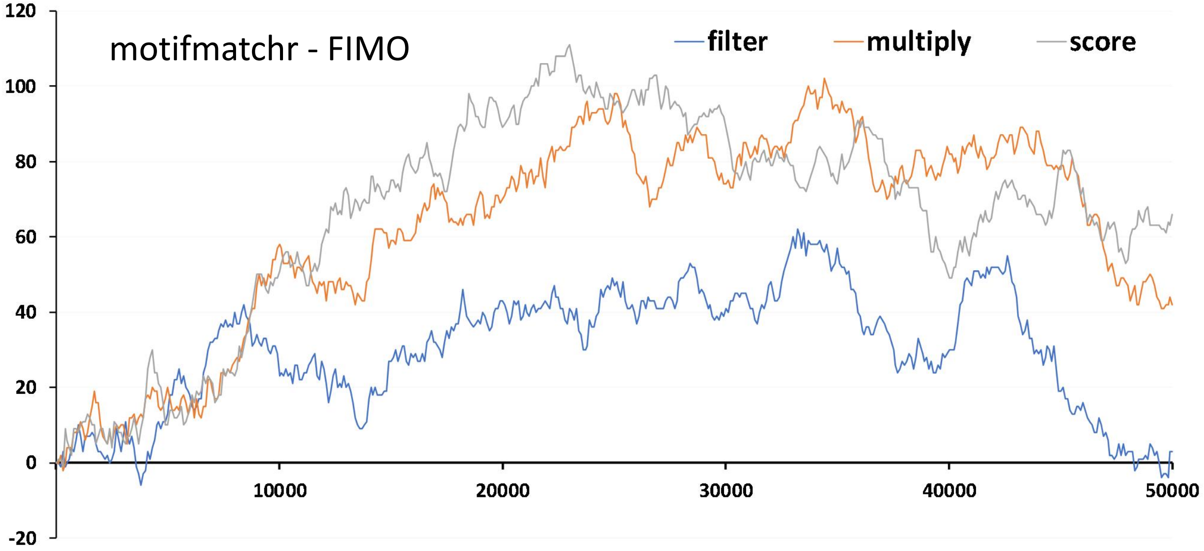


Fig 21. Model performance comparison between peak-TF pairing based on motifmatchr and FIMO

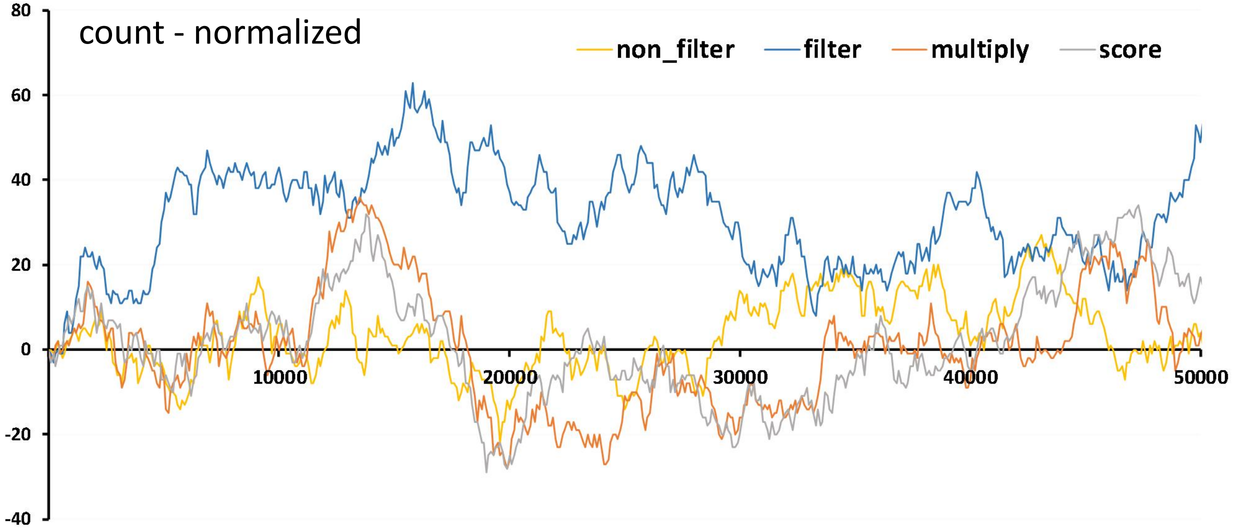


Fig 22. Model performance comparison for normalized data and count data

Results using new data: BMMC

- **Data**: BMMC (bone marrow mononuclear cell)
- model 1 **non-filter** > model 2 **filter**, model 3 **multiply** and model 4 **score**

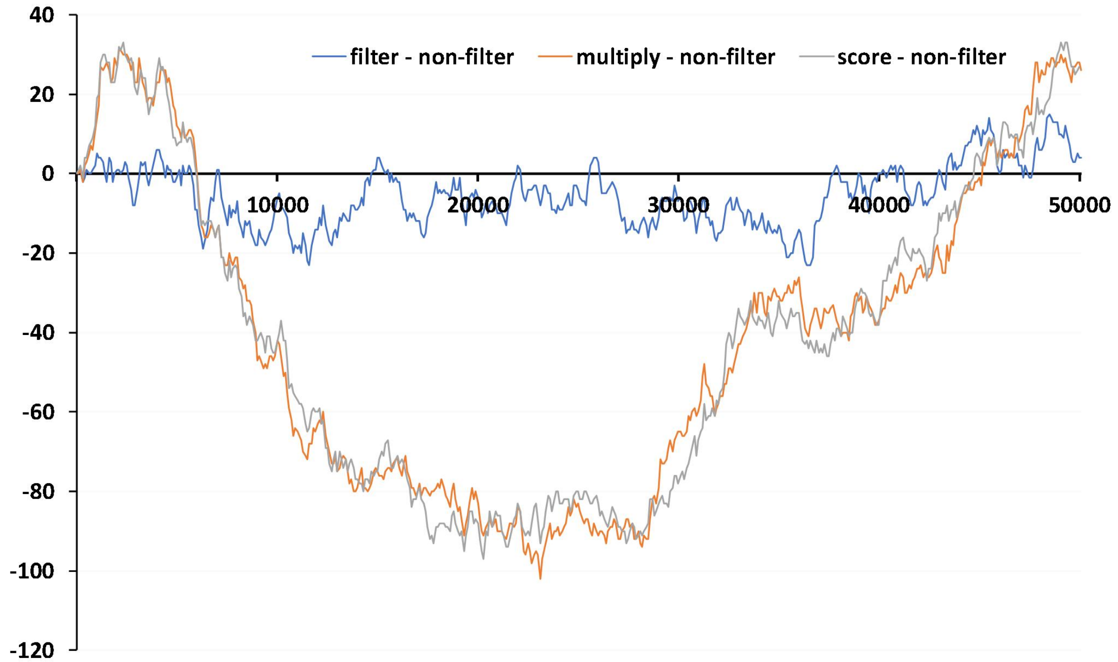


Fig 23. Model performance comparison for peak-gene pairing based on location only (500kb) and peak-TF pairing based on motifmatchr

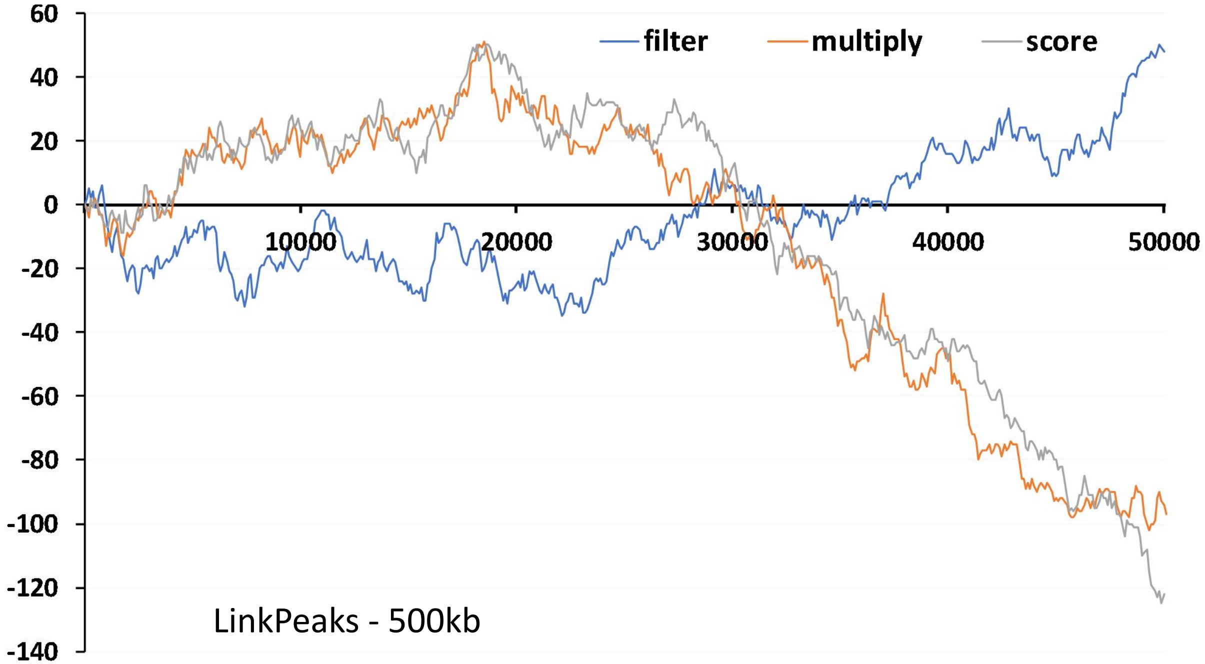


Fig 24. Model performance comparison between peak-gene pairing based on location only (500kb) and based on correlation strength (LinkPeaks)

- **>50% hvgs** will be filtered out in **model 2, model 3 and model 4** using **LinkPeaks** because they **have no ATAC peaks** paired with them
- **500kb** outperforms **LinkPeaks**
- Assumption: Models **involving gene-peak correlation & peak accessibility** is **highly dependent on data quality**

Verification: BMMC is more sparse than PBMC

- **Median** of **summation of count data** of genes(ATAC) in cells in **BMMC** << **Median** of **summation of count data** of genes(ATAC) in cells in **PBMC**
- **Frequency distribution histogram**: count summation (RNA & ATAC) of **PBMC** mostly distributed over **larger values** than **BMMC**

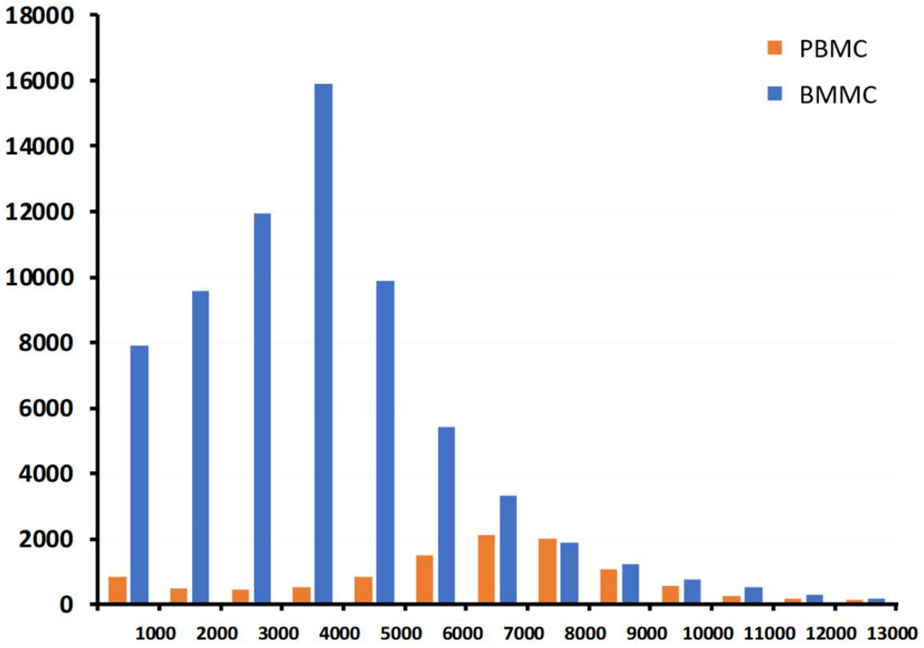


Fig 25. Frequency of count summation in cells of ATAC data

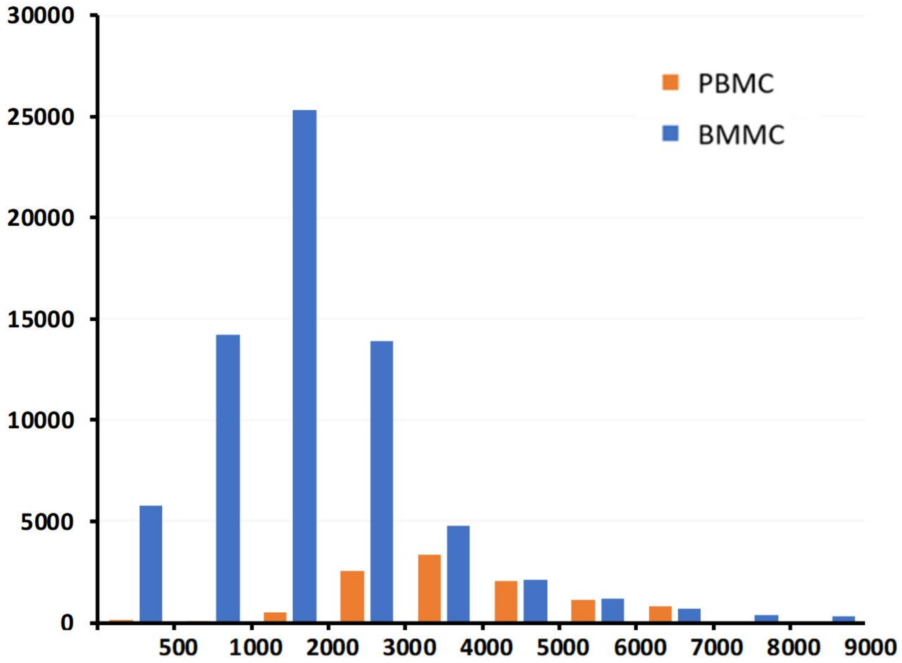


Fig 26. Frequency of count summation in cells of RNA count data

Thank you!