attack[78]. We used the HOMER tool to perform motif matching in the transcription factor footprint region to identify potential binding sites of transcription factors.

3. Inference of the regulatory relationship between TFs and target genes

We first constructed an initial TF-TG network, which consists of three parts, transcription factor to motif, motif to footprint, and footprint to its neighbouring genes (within 500k base pairs around the TSS.) The scoring of TFs to their TG links can be expressed as:

$$\log_{10}(W) + \log_{10}(H) - 10^{-6} \cdot D$$

Where W is the Wellington footprint probability score calculated by the Wellington-bootstrap method, H is the motif purity score calculated by the HOMER tool, and D is the distance between the footprint and the target gene TSS[71].

We then constructed a dynamical system model to further refine the regulatory network. Dictys used stochastic differential equations to represent the transcriptional dynamics of each cell[71].

$$dx_i(\tau) = \left[\alpha_i + \sum_{j \neq i} \beta_{ji} x_j(\tau) - x_i(\tau)\right] d\tau + \sum_k \sigma_{ik} dW_k(\tau)$$

Where $x_i(\tau)$represents the logarithmic expression level of gene i at the assumed time τ, $x_j(\tau)$represents the logarithmic expression level of transcription factor j at the assumed time τ. α_i is the intercept term, and the coefficientβ_{ji} represents the regulatory effect of transcription factor j on target gene i.

4. Visualisation

The Dictys method provides visualisation of GRN sub-networks using a force-directed layout. Taking B cells as an example, we mapped the sub-regulatory network of the transcription factor GATA3.

4 Results

4.1 scMEGA

4.1.1 The result of data preprocession

The Fig. 1 shows the violin plot of scRNA-seq dataset before and after quality control. The three indicators “nFeature_RNA”, “nCount_RNA” and “percent.mt” in the figure represent the

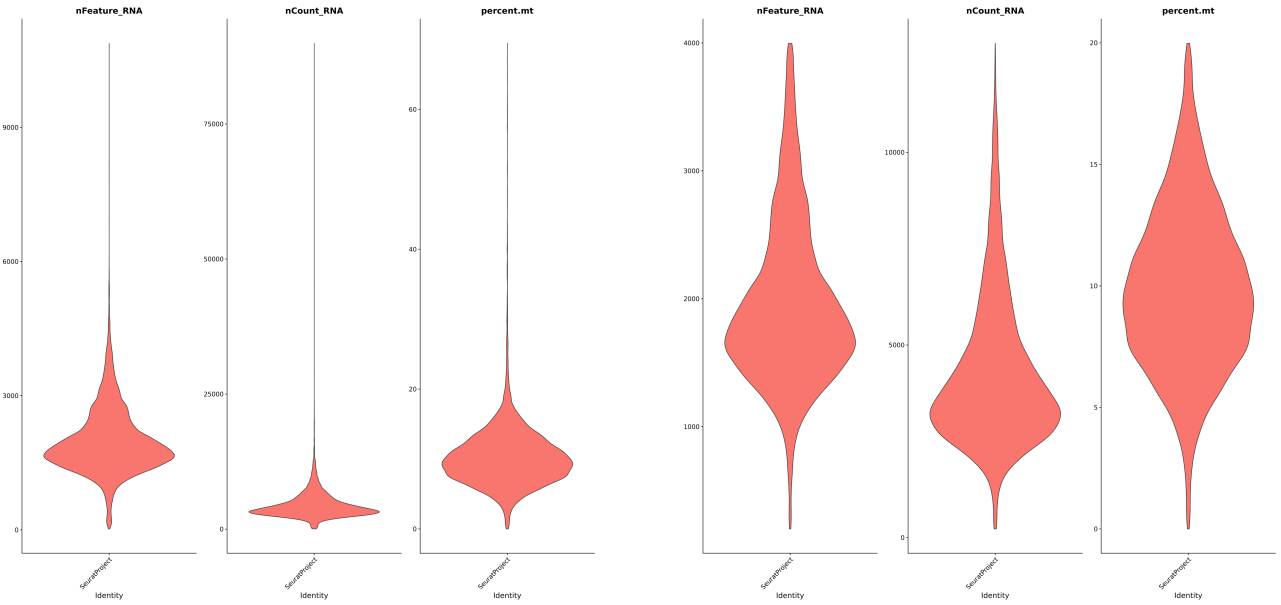

Fig. 1. Violin plots of scRNA-seq datasets before and after quality control.

three quality control indicators mentioned in the method section: the number of genes in the cell, the total number of RNA molecules and the proportion of mitochondrial genes. Violin plot is a visualization tool that combines the advantages of box plot and density plot. The width of the left-right symmetrical “drumsticks” of the violin plot reflects the density of the data at that point, and the vertical range of violin plot reflects the distribution range of the data. Before the quality control, there are obviously more extreme values in the dataset, and these cells are low quality cells. After completing the quality control operation, the distribution of indicators such as the number of genes and the total number of RNA molecules in the cells became more concentrated, the quality of the dataset was significantly improved.

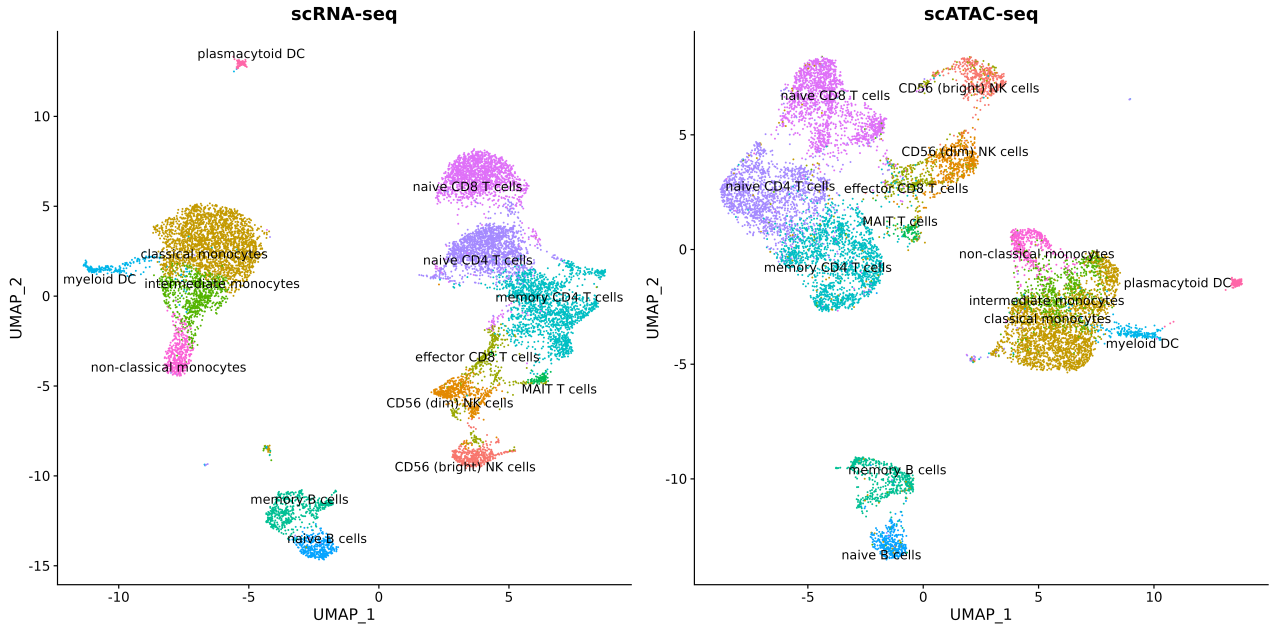

Fig. 2. Cell distribution and annotation results for scRNA-seq and scATAC-seq datasets.

The Fig. 2 shows the distribution of scRNA-seq and scATAC-seq datasets after dimensionality reduction analysis using UMAP. Each point in the figure represents a cell, and the colours and labels indicate different cell types. We successfully divided all cells into 14 categories and completed the cell annotation. After completing all data processing and integration, we obtained a Seurat object with 11,414 samples.

4.1.2 Visualisation of GRN

Finally, we successfully screened 92 active transcription factors and 2,169 genes to construct GRNs. and we filtered the genes with correlation greater than 0.5 to visualise GRNs.

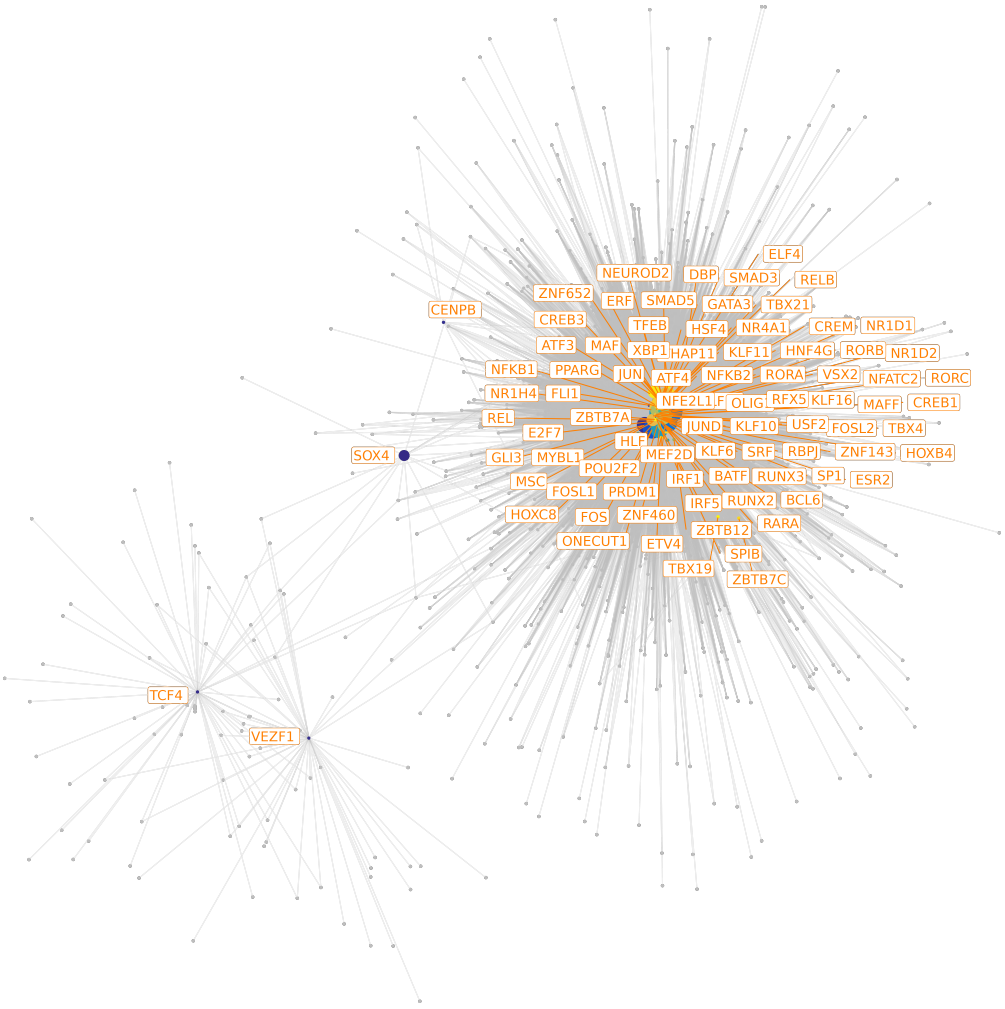

Fig. 3. Gene regulatory network of PBMC dataset mapped by scMEGA method.

The Fig. 3 shows the visualization of the gene regulatory network of the PBMC dataset. Each node in the figure represents a transcription factor or gene; the edge represents the regulatory relationship between genes, and nodes with more linking lines have a greater role in the gene regulation process. For example, Transcription Factor 4 (TCF4) is an important

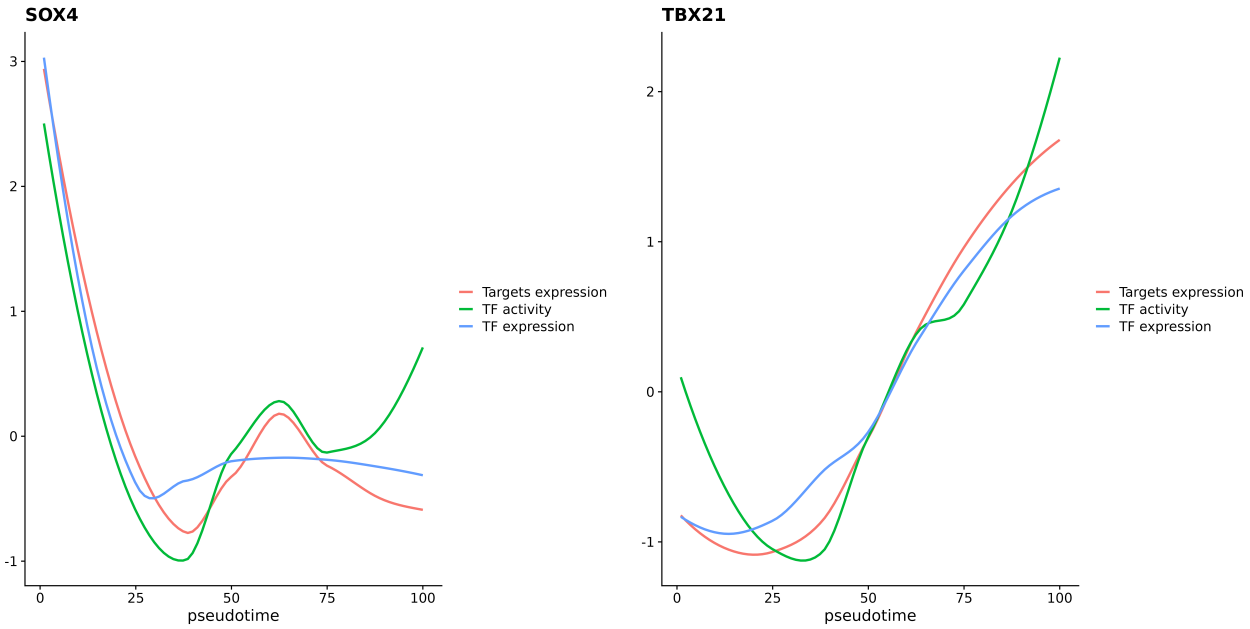

Fig. 4. Changes in the expression of transcription factors and target genes with pseudo time.

transcription factor that regulates foetal neural development during pregnancy by initiating neural differentiation through binding to DNA[79]. We found in our constructed GRN that the COL18A1 gene is its target gene, which is consistent with the prediction in the Harmonizome database[80].

We also plotted the changes in the activity and expression of the transcription factors SOX4 and TCF4, as well as the expression of the target genes with pseudo time. We can see from the Fig. 4 that these two transcription factors activity and expression are basically in line with the trend of change in the expression of the target genes.

4.1.3 Time and memory consumption

The final time consumption of the scMEGA method was 498 seconds, and the peak memory usage was 8444 MiB. All the codes for inferring GRNs by this method can be found at github (`https://github.com/Zilong-j/GRN-benchmark`).

4.2 Pando

We also obtained a Seurat object with 11,414 cells after completing data processing and integration, which we eventually used to construct a gene regulatory network containing 115 transcription factors and 55 target genes.

4.2.1 Selection of transcription factor modules

The Pando method identified 773 transcription factor modules in the gene regulatory network. For example, the transcription factor module of the transcription factor ATF3 and the target ETV6 contains 8 genes and 20 transcription factors, and the number of genomic regions regulated by the transcription factors is 20. In addition, Pando also provides quality assessment criteria for the modules.

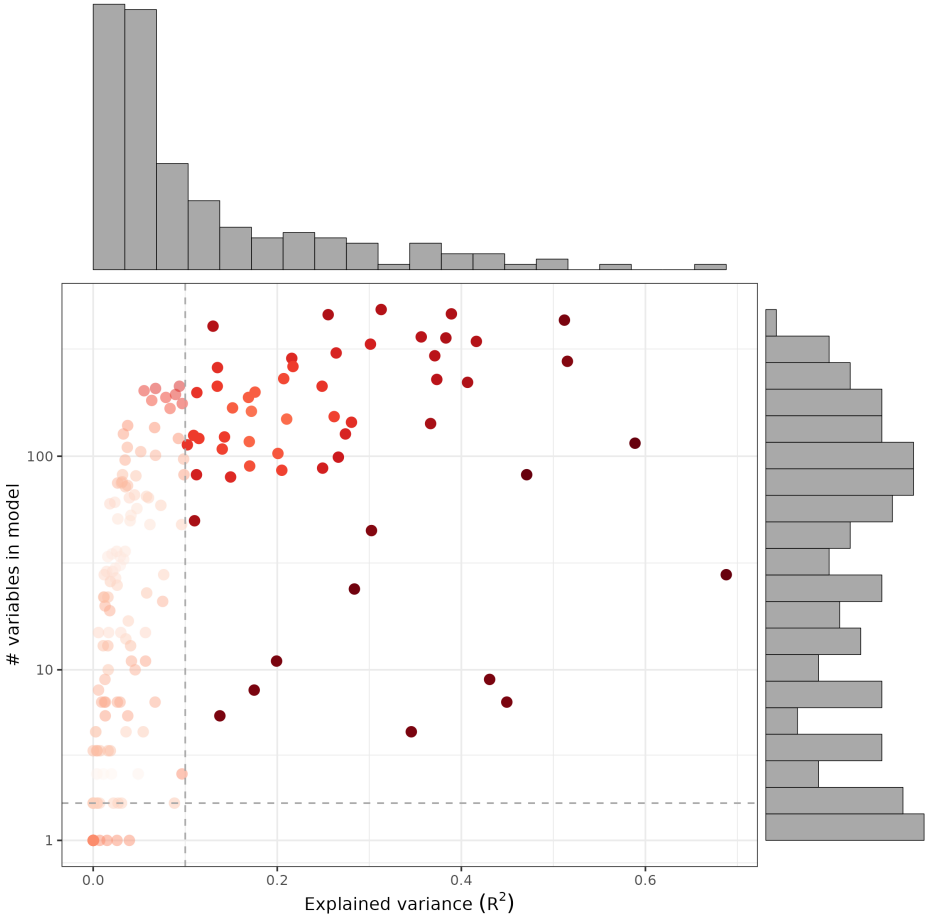

Fig. 5. The quality assessment criteria for the transcription factor modules. The figure consists of a scatter plot and a histogram. The vertical axis of the scatter plot represents the total number of variables in the transcription factor regulatory module, which includes the number of target genes, the number of transcription factors, and the number of regions regulated by the transcription factors. The abscissa is the explained variance of the transcription factor module on the expression of the target gene, which reflects the explanatory power of the transcription factor module. The dashed line is the threshold for filtering transcription factor modules. In addition, the histograms reflect the distribution of explained variance and the number of variables in the module, respectively.

As shown in the Fig. 5 above, each point in the figure represents a transcription factor module, the vertical axis reflects the number of variables in the module, and the horizontal axis reflects the degree to which the module explains the variation in the overall gene expression data. We can find that when the number of variables in the module is larger, its ability to explain the overall gene regulatory network is more likely to be stronger. We followed the filtering rules mentioned in the Method section and successfully filtered out the transcription factor modules with higher explanatory power and mapped out the gene regulatory network.

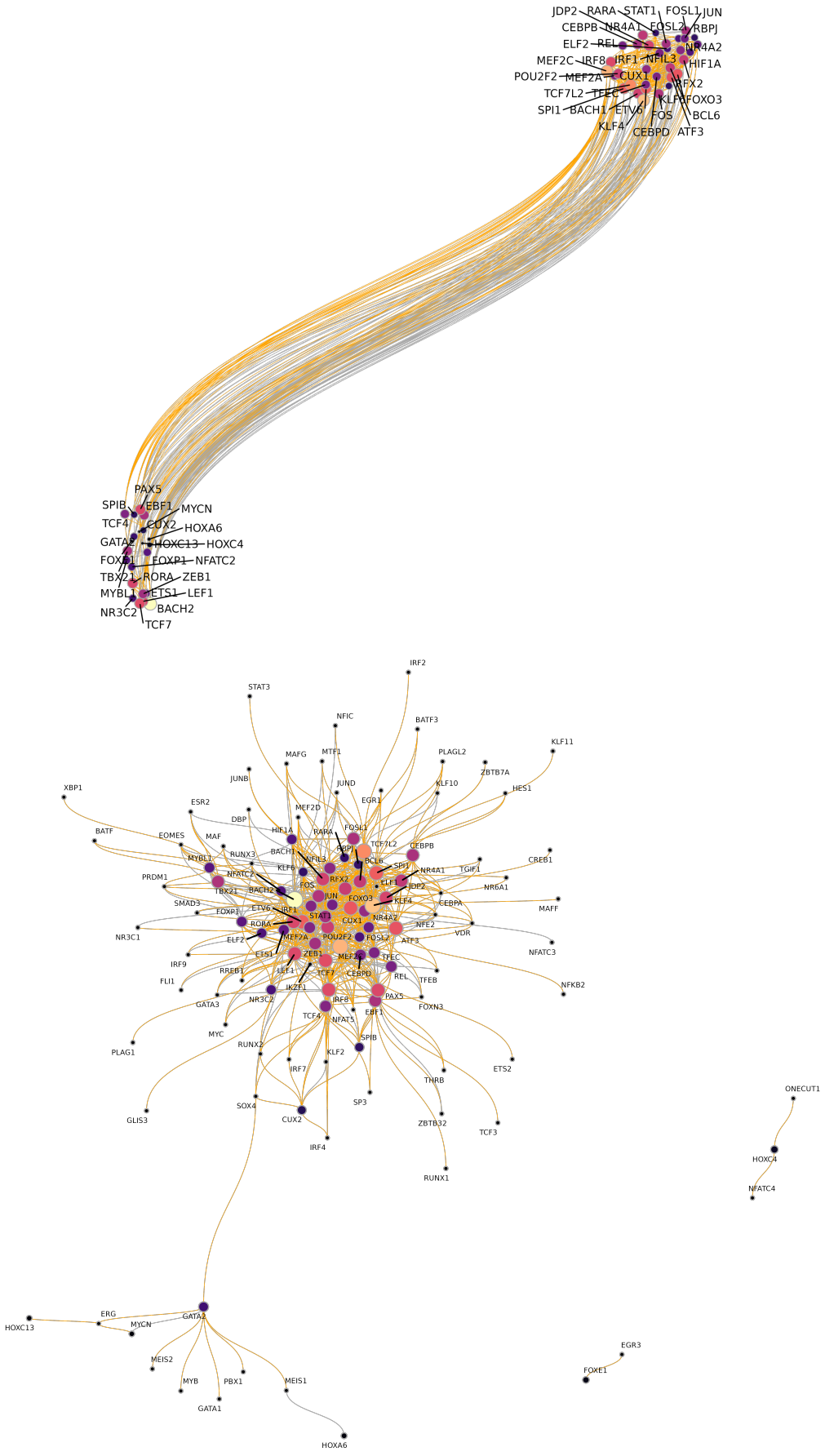

Fig. 6. Gene regulatory network of PBMC dataset mapped by Pando method. These two figures show the linear and force-directed layouts of the GRN, respectively. Each node in the graph represents a gene or transcription factor, and the colour and size of the nodes reflect their centrality in the graph. The edges represent regulatory relationships, and the colour of the edges reflects the type of regulatory relationship, where orange edges represent activation and grey edges represent inhibition.

4.2.2 Visualization of GRN

The Pando package provides rich GRN visualisation tools. We used different drawing methods to visualize the gene regulatory network of PBMC. The upper image in Fig. 6 shows the gene regulatory network mapped after dimensionality reduction using the UMAP method, and the lower image shows the complete gene regulatory network. The layout method is Fruchterman-Reingold (FR) layout, which is a force-directed layout algorithm that determines the position of nodes by simulating physical forces[81]. Due to the complexity of the gene regulatory network, the GRN drawn by the UMAP method cannot provide much effective information intuitively, but the GRN with FR layout fully displays the regulatory relationship between all genes and transcription factors in the GRNs. For example, it can be seen from the right figure that the expression of GATA2 transcription factor activates the expression of SOX4 transcription factor and inhibits the expression of MYCN gene.

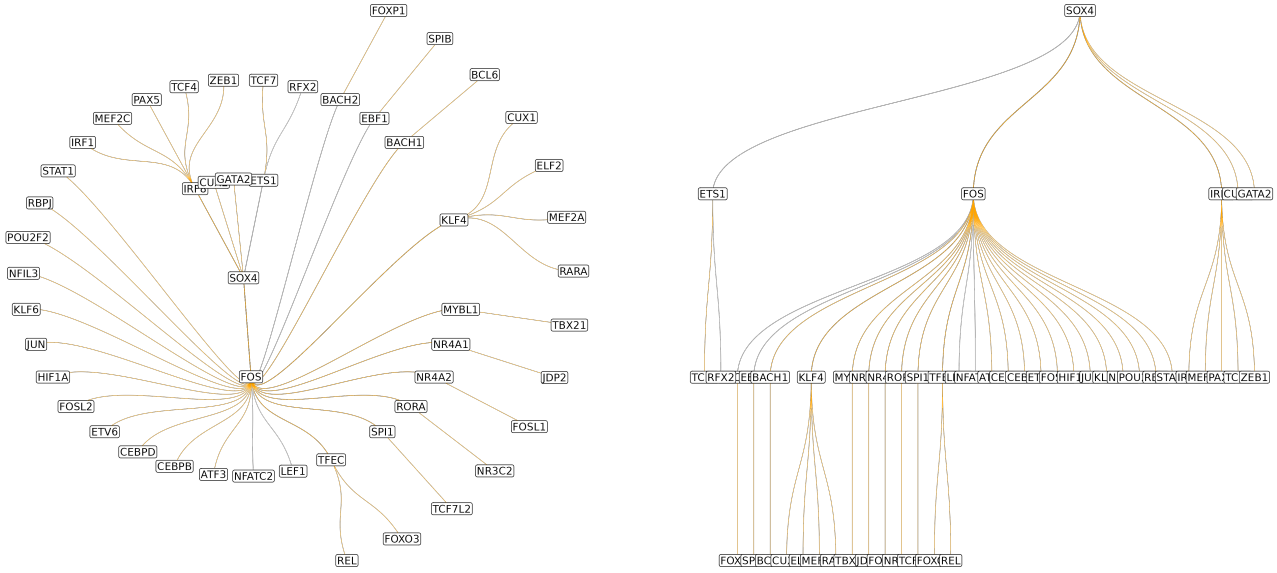

Fig. 7. Sub-regulatory network cantered on transcription factor SOX4. The left figure shows the sub-regulatory network centred on the transcription factor SOX4 in a circular layout, and the right figure shows the sub-regulatory network centred on the transcription factor SOX4 in a tree layout.

For complex GRNs, the output of the previous direct visualisation often does not provide much effective information because there are a large number of nodes in the centre of the graph. To solve this problem, Pando also provides a special method for drawing GRN subgraphs, and we have drawn a sub-regulatory network centred on the SOX4 transcription factor as an example. The Fig. 7 shows all the regulatory relationships involving the SOX4 transcription factor, and we can also judge the type of regulatory relationship from the colour of the straight line.

4.2.3 Time and memory consumption

The final time consumption of the Pando method was 234 seconds, and the peak memory usage was 6859 MiB. All the codes for inferring GRNs by this method can be found at github (`https://github.com/Zilong-j/GRN-benchmark`).

4.3 LINGER

Instead of directly outputting the GRN visualisation, the LINGER method generates matrix results on three regulatory relationships: TF binding (TF-RE), cis-regulation (RE-TG), and trans-regulation (TF-TG). The cell type GRN inference results are consistent with the cell level results, so we only present the cell level GRN results here.

4.3.1 TF binding potential

The first output is the cell-level TF binding potential, which is a matrix of 96533 rows and 451 columns. The Table 1 below shows only the first 5 rows and 9 columns. The row headers are the genomic regions of the RE, the column headers are the names of the transcription factors, and the values of the matrix represent the binding scores of the transcription factors in specific genomic regions. A binding score closer to 1 indicates a higher binding strength. For example, the binding score of the transcription factor SOX15 on the genomic region chr1:100035922-100040109 is 0.9997, which indicates that their binding affinity is very high.

Table 1. Matrix of the TF-RE binding score.

Genomic region	Transcription factors								
	AIRE	TWIST1	SOX15	FOXD2	NEUROD2	IRF1	PPARG	ALX3	IRF2
chr1:100028489-100029404	0.9986	0.9977	0.9991	0.9980	0.9985	0.9978	0.9974	0.9994	0.9973
chr1:100034436-100035279	0.9990	0.9980	0.9986	0.9981	0.9996	0.9968	0.9964	0.9994	0.9981
chr1:100035922-10004109	0.9993	0.9988	0.9997	0.9991	0.9997	0.9985	0.9978	0.9995	0.9986
chr1:100041493-100041927	0.9981	0.9972	0.9982	0.9973	0.9988	0.9941	0.9901	0.9990	0.9965
chr1:100046068-100047735	0.9993	0.9985	0.9996	0.9985	0.9996	0.9988	0.9983	0.9996	0.9987

4.3.2 Cis-regulatory network

The second output is the cell-level cis-regulatory network. LINGER outputs a matrix containing 1,529,212 pairs of regulatory relationships. The following Table 2 shows the first five rows of the matrix. The first column of the matrix represents the genomic region of the RE, the second column is the name of the target gene, and the value in the third column of the

table represents the cis-regulatory score of the RE for the TG. The higher the score, the greater the regulatory potential of the RE on the target gene.

Table 2. Cell-level cis-regulatory networks.

Region	Target gene	Cis score
chr1:100028489-100029404	AGL	2.67×10^{-6}
chr1:100028489-100029404	CDC14A	6.73×10^{-4}
chr1:100028489-100029404	DBT	2.76×10^{-5}
chr1:100028489-100029404	DPH5	7.53×10^{-5}
chr1:100028489-100029404	EXTL2	9.24×10^{-7}

4.3.3 Trans-regulatory network

The third output is the cell-level trans-regulatory network. The output result is a matrix with 14907 rows and 451 columns. The Table 3 shows the first 5 rows and 9 columns of the matrix. The row headers of the matrix are the names of the target genes, the column headers are the names of the transcription factors, and the values of the matrix represent the trans-regulatory potential of TF-TG. Trans-regulatory potential reflects the ability of transcription factors to indirectly regulate target genes through regulatory elements. This potential combines the results of TF binding potential and cis-regulatory potential.

Table 3. Cell-level trans-regulatory network.

	AHR	AIRE	ALX3	ALX4	AR	ARID3A	ARID3B	ARID5A	ARID5B
SAMD11	1.01×10^{-5}	4.11×10^{-6}	1.32×10^{-6}	7.10×10^{-7}	2.89×10^{-6}	3.68×10^{-6}	5.97×10^{-6}	6.45×10^{-6}	7.17×10^{-6}
NOC2L	4.74×10^{-5}	1.96×10^{-5}	5.82×10^{-6}	4.11×10^{-6}	1.33×10^{-5}	2.78×10^{-5}	2.42×10^{-5}	2.89×10^{-5}	1.85×10^{-5}
KLHL17	3.87×10^{-7}	9.23×10^{-8}	5.25×10^{-8}	3.64×10^{-8}	1.06×10^{-7}	2.94×10^{-7}	2.79×10^{-7}	2.04×10^{-7}	2.59×10^{-7}
PLEKHN1	1.93×10^{-6}	1.28×10^{-6}	8.38×10^{-6}	1.14×10^{-7}	3.82×10^{-6}	1.03×10^{-6}	7.70×10^{-7}	7.90×10^{-7}	1.22×10^{-6}
HES4	2.14×10^{-4}	1.67×10^{-5}	1.86×10^{-5}	4.28×10^{-6}	1.39×10^{-5}	4.09×10^{-4}	8.19×10^{-5}	4.60×10^{-5}	2.63×10^{-4}

4.3.4 Time and memory consumption

The final time consumption of the LINGER method was 11 hours and 8 minutes, and the peak memory usage was 22591 MiB. All the codes for inferring GRNs by this method can be found at github (`https://github.com/Zilong-j/GRN-benchmark`).

4.4 Dictys

4.4.1 Sub-regulatory network

The Dictys method outputs sub-regulatory network maps rather than entire GRNs. For example, we mapped a sub-regulatory network centred on the transcription factor GATA3 in B cells.

The central node in the Fig. 8 represents the transcription factor GATA3, the nodes around it are the target genes of the transcription factor GATA3, and the lines represent the regulatory relationship between GATA3 and the target genes. The subnetwork diagram generated by Dictys indicates the direction of the regulatory relationship, but does not specify the specific type of regulatory relationship.

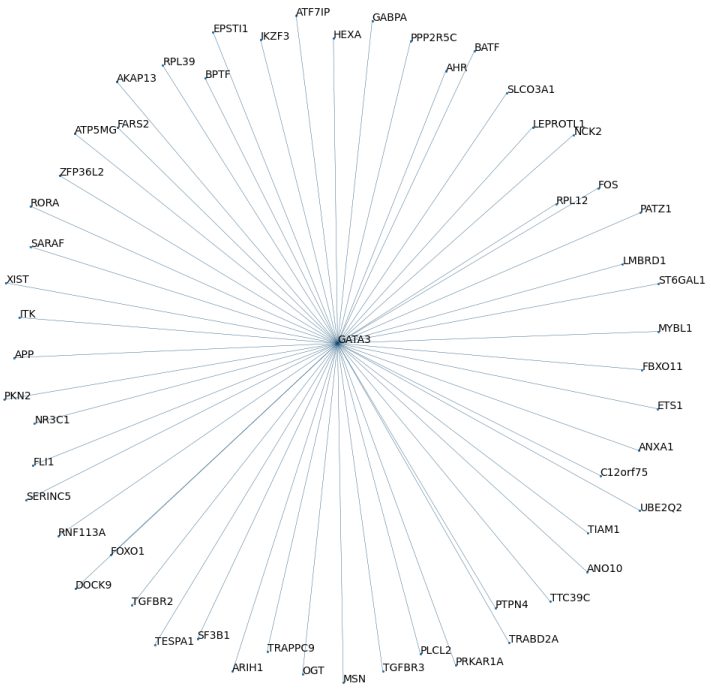

Fig. 8. The sub-regulatory network centred on the transcription factor GATA3.

4.4.2 Gene expression and regulation in different cells

Dictys can also compare the expression and regulation of genes in different cells. For example, we mapped the difference in gene expression and regulation between regulatory T cells and B cells. The top 20 genes are shown in Fig. 9.

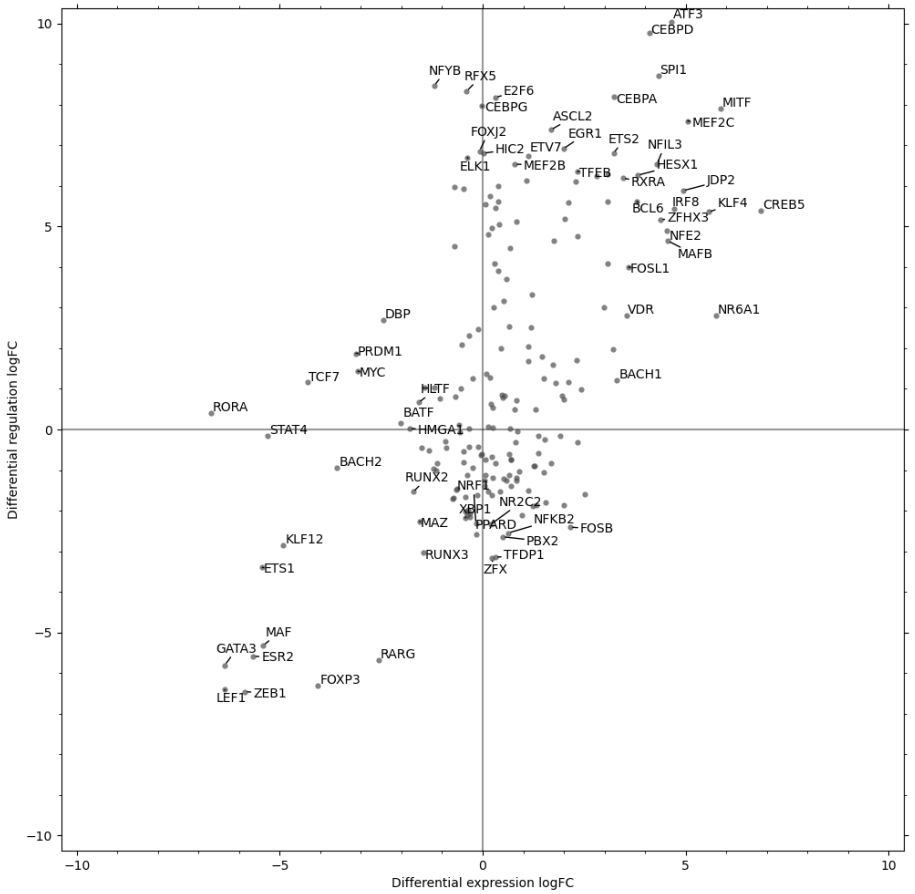

Fig. 9. Gene expression difference map. The horizontal and vertical axes of the figure are the log fold change of gene expression level and regulatory activity, respectively. The farther a gene's position is from the centre on the horizontal axis, the greater the difference in its expression levels between the two cells. And the farther a gene's position is from the centre on the vertical axis, the greater the difference in the regulatory activity of the gene in the two cells.

From Fig. 9, we can see that in B cells and T cells, the expression levels of genes such as RORA and GATA3 are quite different, and there is a large difference in the regulatory activity of genes such as ATF3 and CEBPD. This figure is important for us to identify the differences between GRNs of different cell types.

4.4.3 Time and memory consumption

The final time consumption of the Dictys method was 23 hours and 22 minutes, and the peak memory usage was 6408 MiB. All the codes for inferring GRNs by this method can be found at github (`https://github.com/Zilong-j/GRN-benchmark`).