



AIM: an accurate and explainable model for ATAC to GEX translation and pathway analysis

Quang H. Nguyen¹ · Hoang V. Tran¹ · Huu Tien Nguyen¹ · Phuong T. M. Le² · Phi Le Nguyen¹ · Binh P. Nguyen³

Received: 14 May 2024 / Accepted: 23 February 2025
© The Author(s) 2025

Abstract

The development of multimodal technologies has enabled the simultaneous measurement of various cellular modalities, such as chromatin accessibility (ATAC), gene expression (GEX), and surface protein abundance in single cells. However, the lack of multimodal datasets requires the development of robust algorithms that can translate data between different modalities. In this study, we present AIM, a framework for accurate and interpretive multimodal translation, specifically designed for the conversion of ATAC data into GEX profiles. AIM introduces a novel two-tier modeling architecture. The upper tier captures the global relationships between ATAC and GEX, generating an initial estimate of gene expression. The lower tier performs a finer-grained analysis by modeling inter-chromosomal interactions to refine the generated GEX representation. This modular structure enhances both the accuracy and adaptability of AIM. Additionally, an integrated attention mechanism provides interpretability by highlighting critical chromatin regions influencing specific gene expressions. Our experimental results demonstrate that AIM achieves state-of-the-art performance, with a per-chromosome RMSE of 0.2206, outperforming existing approaches (0.2232). Furthermore, the attention maps generated by AIM offer a pathway analysis capability, uncovering biologically significant gene-gene interactions such as ARHGAP24-ARAP2 and SYK-PAX5. These findings validate AIM's effectiveness not only as a data translation tool but also as a platform for deriving mechanistic insights into gene regulatory dynamics.

Keywords Single cell · Explainable AI · Pathways analysis · Modality prediction · ATAC · GEX

Quang H. Nguyen and Hoang V. Tran have contributed equally to this work.

✉ Binh P. Nguyen
binh.p.nguyen@vuw.ac.nz

Quang H. Nguyen
quangnh@soict.hust.edu.vn

Hoang V. Tran
hoang.tv861999@gmail.com

Huu Tien Nguyen
tien.nh205033@sis.hust.edu.vn

Phuong T. M. Le
phuongtm.ump@vnu.edu.vn

Phi Le Nguyen
lenp@soict.hust.edu.vn

¹ School of Information and Communication Technology, Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi 100000, Vietnam

² University of Medicine and Pharmacy, Vietnam National University, 144 Xuan Thuy, Hanoi 100000, Vietnam

1 Introduction

Data from single-cell technologies such as chromatin accessibility [1], gene expressions [2], and protein abundance [3] provide insights that are of great value in cellular regulation research, most notably in classifying cancer cells [4, 5]. On one hand, individual data types furnish unique insights into the regulation of cellular processes. Conversely, these modalities also elucidate their interrelation. For instance, chromatin accessibility delineates the physical accessibility of enhancers, promoters, and insulators to DNA, thereby orchestrating gene expression. The extent of gene expression reciprocally influences the abundance of surface proteins. Hence, elucidating this nexus among diverse modalities stands to enrich our comprehension of the intricate regulatory network governing gene expression within cellular systems.

³ School of Mathematics and Statistics, Victoria University of Wellington, Kelburn Parade, Wellington 6012, New Zealand

Genes seldom operate in isolation; rather, they function within intricate networks or pathways to execute biological functions [6]. Specifically, when external stimuli prompt changes in the expression or activity of a gene product, it can trigger subsequent modifications in the expression or activity of other gene products. This process gives rise to a phenomenon known as regulatory pathways. Such an interconnected cascade of events can regulate various levels of activity. Understanding these regulatory pathways is crucial for deciphering the molecular complexities underlying both normal physiological processes and diseases. Single-cell technologies have facilitated the discovery of pathways using Gene Set Enrichment Analysis (GSEA) [7, 8]. However, the identification of pathways necessitates potential candidate gene sets, which must be enriched and organized into pathway sets. Selecting these candidate gene sets demands considerable expertise and time, while testing them with all conceivable gene sets is even more time-consuming. Regrettably, there is currently no tool available to expedite the identification of potential candidate gene sets.

In recent years, thanks to the advancement in technology, experimental methods to measure multiple modalities at the same time, such as chromatin accessibility, gene expressions, and surface protein abundance in the same cell, have gradually become popular. Methods such as CITE-seq [9] allow simultaneous quantification of gene expression and surface protein abundance, and SNARE-seq [10] allows simultaneous quantification of gene expression and chromatin accessibility. Multi-omic single-cell data offers a comprehensive view of cellular biology by simultaneously measuring various molecular features within individual cells. This approach enables researchers to uncover complex relationships between different layers of cellular information, such as genomics, transcriptomics, epigenomics, proteomics, and metabolomics. Compared to single-omic single-cell data, which only captures one type of molecular information, multi-omic data provides a more holistic understanding of cellular states and functions. Although technology has made it possible to quantify multiple modalities simultaneously, current multimodal datasets are still limited, and single-modality data are much more common. In particular, there are many datasets that carry important information, notably those related to cancer [11, 12]. These datasets are very difficult to collect and reproduce with multi-omics single-cell technology. Therefore, building models that can integrate and convert single-modality single-cell datasets into multi-modality datasets is crucial for unlocking the full potential of single-omic single-cell datasets. Understanding this, Luecken *et al.* [13] posed the problem of predicting the feature of one modality from that of another modality, such as from ATAC to GEX. From there, the single-modal datasets can be integrated and converted to multimodality. Currently, there are many studies solving the problem of modality pre-

diction, such as BABEL [14], which builds two encoders and two decoders corresponding to each modality. These encoder-decoder sets are trained to encode two modalities into the same latent space and then use the corresponding decoder to predict the desired modality. MuItiVI [15], Polar-Bear [16], and scMM [17] have quite similar architectures to BABEL, but they use the Variational Autoencoder (VAE) [18] architecture instead of an autoencoder like BABEL and some changes in the training process. The scMoGNN model [13] builds a giant graph using the entire input data set and then passes it through the GCN network to encode each node corresponding to each cell in the input data. The encoded node vectors are then passed through a fully-connected network to predict the desired modality. The Cajal model [19] uses a simple fully-connected (FC) model but has unique data enrichment methods that increase the accuracy of the model.

Despite numerous attempts to address the translation challenge from ATAC to GEX, existing research suffers from two primary deficiencies: (1) *inadequate modeling of inter-chromosome relationships*, and (2) *lack of interpretability*. Existing methodologies universally treat each modality (i.e., ATAC and GEX) as a gene sequence, concentrating on the transition of gene sequences from ATAC to GEX. However, empirical findings indicate that genes residing on the same chromosome typically exhibit more robust interconnections than those located on separate chromosomes. Significantly, several studies have observed clusters of genes activated concurrently on a single chromosome rather than undergoing independent activation [20–22]. Therefore, the incorporation of models elucidating intra-chromosomal associations presents a prospective avenue for enhancing the accuracy of translating from ATAC to GEX. In addition, current methodologies emphasize the generation of GEX representations but fail to adequately address the necessity of interpreting the resulting outcomes. This oversight compromises the persuasiveness of the outcomes, consequently constraining the pragmatic utility of the proposed solutions.

In this study, we introduce AIM (stands for **A**ccurate and **E**xpl**I**n**A**ble **M**odel), a novel approach to derive GEX representation from ATAC data, effectively addressing the two aforementioned challenges. To tackle the initial challenge concerning inter-chromosome relationship modeling, AIM adopts a two-layer architecture consisting of two modules: *modality transformation* and *chromosome refinement*. Initially, the modality transformation layer aims to capture the general relationship between ATAC and GEX data and roughly convert the entire ATAC sequence into a GEX representation. Subsequently, the chromosome refinement phase involves 22 sub-models, each tailored to a specific chromosome. The objective of the chromosome refinement layer is to fine-tune the chromosome representations generated by the modality transformation, thereby enhancing the accuracy of the overall translation process. In the meantime, to enhance

the interpretability of the model, we incorporate an attention mechanism within the chromosome refinement layer. This mechanism facilitates the analysis of relationships between genes within both the ATAC and GEX sequences for each chromosome. By directing attention between vectors representing genes as well as between genes and chromatin accessibility, we can elucidate the interplay and dependencies among these elements.

Our main contributions are as follows:

- We introduce a novel two-layer framework for translating ATAC to GEX. The first layer models the global relationships between chromatin accessibility and gene expression, while the second layer focuses on capturing fine-grained inter-chromosomal dependencies, enhancing the accuracy of GEX predictions.
- We incorporate an attention mechanism to improve the interpretability of the model. This design highlights critical chromatin regions influencing gene expression, providing a transparent view of the model's decision-making process.
- Our model identifies key patterns learned during training, including the prevalence of ATAC features along gene segments on chromosomes and the tendency of genes to be expressed in clusters. By grouping genes with high pathway interaction potential by cell type and chromosome, the model reveals biologically meaningful clusters, many of which align with established findings, demonstrating the utility of our approach for pathway discovery.

2 Related work

In this section, we provide an overview of previous endeavors aimed at addressing the ATAC to GEX prediction problem.

Wu *et al.* [14] suggests using an autoencoder model consisting of two encoders and two decoders. These encoders and decoders are built entirely using fully-connected layers. In which two encoders are responsible for encoding ATAC and GEX data into latent space, respectively. The latent space of both encoders has the same number of dimensions. The remaining two decoders are responsible for decoding the latent space vector back to the original ATAC and GEX data. From two encoders and two decoders, they can combine four combined models, including GEX encoder–GEX decoder, ATAC encoder–ATAC decoder, GEX encoder–ATAC decoder, and ATAC encoder–GEX decoder. In which the pair GEX encoder–GEX decoder and ATAC encoder–GEX decoder are calculated loss using Binary Cross Entropy, the remaining two pairs use negative binomial loss. During the training process, data will be passed through all four encoder-decoder pairs mentioned above, then loss-synthesized and backpropagation performed. MultiVI [15]

and scMM [17] have the same idea of using variational autoencoders. Their model also includes two encoders and two decoders. These encoders and decoders are built entirely on fully-connected layers. The two encoders encode the ATAC and GEX data into two corresponding distributions. The two distributions are then averaged, and using this averaged distribution, the two decoders reconstruct the original ATAC and GEX vectors. The only difference between the two models proposed by MultiVI [15] and scMM [17] is that they use different distributions for the model. MultiVI [15] uses a negative binomial distribution, and scMM [17] uses a zero-inflated negative binomial distribution. Polarbear [16] introduces an approach involving two independent autoencoder models. One autoencoder encodes the ATAC vector into a latent space and decodes it back to reconstruct the ATAC vector. The other follows a similar process, encoding the GEX vector into the latent space and decoding it back to reconstruct the GEX vector. Both models share a latent space of identical dimensions and are trained independently. After training, the model parameters are frozen. By leveraging the shared latent space dimensions, the decoder blocks of the two models can be interchanged, enabling the creation of two cross-modal models: one translating ATAC inputs into GEX outputs and the other translating GEX inputs into ATAC outputs.

Luecken *et al.* [19] setup a series of models with the highest results in the ATAC to GEX prediction problem. Among those models, the authors propose to use GCN [23] for the scMoGNN model [13] to embed cells and thereby predict the desired modality. What is special about this model is that they do not consider each cell as an independent vector but build a two-layer graph with each cell corresponding to a cell node and each feature corresponding to a feature node. The edges between the cell node and the feature node of the graph are constructed based on the value of the feature in the vector of that cell. The author explains that this approach helps to provide more high-order interaction information between cells, thereby improving the results of the model. The graph, after being built, will be passed through GCN layers to be able to embed the features of the nodes. The cell node features are then passed through the FC layers to be able to predict the desired modality. The disadvantage of this method is that the graph after construction is too large, consuming a lot of resources to calculate, but the results are not really groundbreaking. The model that Dengk *et al.* [19] proposed is relatively simple. Before training, GEX and ATAC data are dimensionally reduced using truncatedSVD. The data after dimension reduction will be normalized, and then a Kernel Ridge Regression model with an RBF kernel will be used to predict ATAC data to GEX. Wen *et al.* [19] uses a multi-model ensemble strategy to achieve better results. Initially, the ATAC and GEX data will also be dimensionally reduced using truncatedSVD. The data after dimension reduction will

be divided into 10 small folds. Perform training with nine folds and the remaining one fold is used for validation one at a time until all 10 folds have been used for validation once, thereby obtaining 10 different models. Ensemble 10 models together and take the average result of the 10 models. This process is performed with two different models, a k-neighbors regressor and a random forest regressor, to get two outputs. These two outputs will then be combined by multiplying the weight of 0.45 with the output of the random forest regressor model plus 0.55 times the output of the k-neighbors regressor model. Novel *et al.* [19] uses deep learning networks to predict ATAC to GEX. The original data will also be preprocessed through a few mathematical transformations and then dimensionally reduced using truncatedSVD. It will finally be passed through a network containing four hidden layers with 2048, 2048, 2048, and 512 nodes, respectively. The Cajal model [19] used a ranking feature method that reduces the data dimensionality, then adds to the data hand-crafted features to enrich the data, and then passes through the FC layers to predict the desired modality. Specifically, with more than 116k features, ATAC will initially be ranked by cell type using the rank feature method provided by the scanpy library. This method compares the expression of features between cell types so that, for each cell type, a ranking of the most expressed to the least expressed features is obtained. From there, for each cell type, the author selects the 100 highest-ranked features and combines them. Then add the cell's total count, median, and standard error per batch at the end of each vector. The final vector will be passed through the FC layer to predict modality. Although this method is simple, the results are the best currently.

3 Preliminaries

In this section, we begin by formulating the problem, followed by a description of the dataset utilized in this study.

3.1 Problem formulation

Gene expression (GEX) data represents the quantity of mRNA expressed by genes. A GEX representation of a cell is denoted by a t -dimensional vector $\mathbf{y} = \{y_1, y_2, \dots, y_t\}$, where each feature y_i is a real number, i.e., $y_i \in \mathbb{R}$, indicates the mRNA expression value of the gene corresponding to that feature. Each feature will have metadata that tells us which gene that feature represents.

ATAC data reflects the physical accessibility of chromatin regions, facilitating gene expression. An ATAC representation of a cell is represented by a vector $\mathbf{x} = \{x_1, x_2, \dots, x_k\}$, where each feature $x_i \in \mathbb{R}$ represents physical access to the chromatin site represented by that feature. Each feature

Table 1 Details of the data set used in this study

Source dimension	116490
Target dimension	13431
Train cells	42492
Test cells	1000
Train batches	10
Test batches	3

will have metadata that tells us which chromatin accessibility location that feature represents.

The ATAC-to-GEX translation problem entails generating the GEX representation corresponding to a given ATAC data. Specifically, the objective is to construct a model f that takes an ATAC data $\mathbf{x} = \{x_1, x_2, \dots, x_k\}$ as input and produces a GEX vector $\hat{\mathbf{y}} = f(\mathbf{x})$ such that $\hat{\mathbf{y}}$ closely resembles $\mathbf{y}$ (the GEX corresponding to ATAC $\mathbf{x}$). The aim is to minimize the discrepancy between the generated GEX representation and the true GEX data. The problem can be mathematically formulated as follows:

Input: ATAC representation $\mathbf{x} = \{x_1, x_2, \dots, x_k\}$.

Output: GEX representation $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_t\}$.

Objective: Minimize:

$$RMSE = \sqrt{\sum_{i=1}^t \frac{(\hat{y}_i - y_i)^2}{t}}. \quad (1)$$

3.2 Dataset

Original dataset. The single-cell dataset utilized in this study is derived from the bone marrow samples of 12 healthy donors. Specifically, half of the samples were subjected to measurement via the 10× Multiome Gene Expression and Chromatin Accessibility tool, while the remaining half underwent measurement using the 10× 3' Single-Cell Gene Expression kit in conjunction with Feature Barcoding, complemented by the BioLegend TotalSeq B Universal Human Panel v1.0. Notably, the dataset under consideration in this investigation pertains solely to the output generated from the former half of samples, employing the 10× Multiome Gene Expression and Chromatin Accessibility tools. This dataset comprises both source data (ATAC data) and target data (GEX data). Each source data vector corresponds to a target data vector, with both vectors quantified from a single cell. The specifics of the cell set are delineated in Table 1.

Additional data collection and preprocessing. We partition the dataset into 23 subsets, each representing one of the 23 chromosome pairs. This partitioning reduces data dimensionality and mitigates potential cross-chromosomal interference, based on the hypothesis that genes and ATAC features on the same chromosome are more likely to interact due to their spatial proximity. The segmentation of the

ATAC dataset is straightforward, as each ATAC feature inherently includes its chromosomal location. However, dividing the GEX dataset, which provides only gene names, requires additional preprocessing. To address this, we retrieve detailed gene location information using the Entrez library [24], leveraging the NCBI database [25] to map each gene to its corresponding chromosome. This strategy ensures consistent and accurate dataset partitioning, crucial for the algorithm's chromosome-specific processing modules.

The data retrieval process is outlined as follows:

1. For each gene present in the dataset, we formulate a search query comprising the name of that gene, and the designated search database is GENE. Additionally, we impose supplementary criteria to constrain the search results to genes within the human gene set (*Homo sapiens*). The outcome of this process yields a roster of gene IDs.
2. Utilizing the obtained list of gene IDs, we proceed to retrieve pertinent data about each gene to ascertain the assembly version information. Each assembly version encompasses crucial details, including the name of the chromosome housing the gene, the starting and ending positions of the gene within the chromosome, and the directionality of the gene strand.
3. Acknowledging the existence of numerous assembly versions, each potentially harboring distinct chromosomal and positional data, we opt to filter and exclusively employ the most recent assembly version, namely GRCh38.p14. This strategic selection ensures data consistency and reflects the latest available information.

Due to the absence of discovery or study for certain genes in the dataset, information could only be collected for 11,670 out of the total 13,431 genes present. A summary of the dataset particulars is outlined in Table 2. It is noteworthy that no dataset is available for chromosome 23, as none of the Chromosome Accessibility (CA) genes within the provided dataset are situated on this chromosome.

4 Proposed method

This section details AIM, our proposed model. Section 4.1 presents an overview of AIM, followed by an in-depth explanation of the Modality Transformation module in Sect. 4.2 and the Chromosome Refinement module in Sect. 4.3.

4.1 Overview

AIM employs a hierarchical architecture comprising two layers: Modality Transformation and Chromosome Refinement as shown in Fig. 1. The Modality Transformation layer's objective is to model the overall relationship between

ATAC and GEX. It takes the original ATAC data as input and generates the complete GEX representation. Meanwhile, the Chromosome Refinement layer aims to capture interactions between genes within each chromosome to refine the GEX vector generated by the ATAC data, thereby enhancing overall accuracy. Specifically, there are 22 Chromosome Refinement models, each tailored to a specific chromosome. These models receive subsets of the GEX representation generated by the Modality Transformation layer and produce more accurate GEX representations for their respective chromosomes.

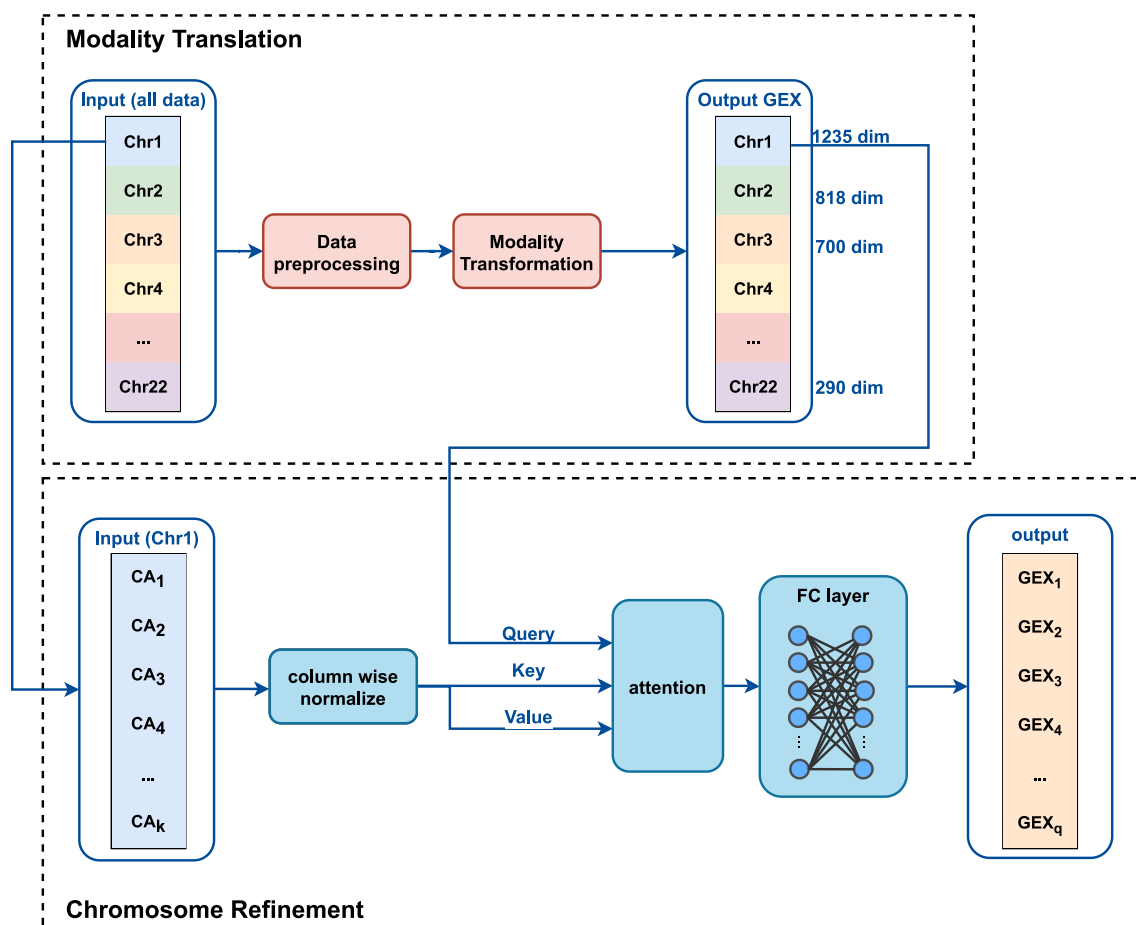
The reason for using this dual model architecture is that the Modality Transformation has applied dimensionality reduction techniques to the ATAC data, reducing the 116,490 ATAC values to the top-ranked features with approximately 1506 features. Thus, the number of features for each chromosome used in Modality Transformation has been significantly reduced. For example, chromosome 1 has 11,497 ATAC values, but only 139 values are used by the Modality Transformation block. The reason Modality Transformation employs dimensionality reduction techniques is that the number of ATACs across all chromosomes is very large (116,490). Additionally, we assume that the CAs on the same chromosome will influence gene expression more than the CAs on different chromosomes because these regions may be related to promoters, enhancers, and corresponding transcription factors.

Therefore, we have proposed adding a Refinement Model for each chromosome. This model uses all ATACs on that chromosome (for instance, the Refinement Model for chromosome 1 uses the full 11,497 ATACs instead of just the 139 ATAC values used by the Modality Transformation) to perform the Attention mechanism with each GEX value of each gene on this chromosome (the detailed Attention mechanism is described in Fig. 4), thereby helping to more accurately refine the GEX values of the genes on this chromosome.

Overall, the flow of the model will be executed as follows. Each input ATAC (all data from 22 chromosomes) will be processed through two streams. The first stream is Modality Translation, where the data will be processed and then passed through a neural network to predict the GEX vector (all data for 22 chromosomes). The second stream is Chromosome Refinement, where the input ATAC will be divided into smaller sets corresponding to each chromosome. The GEX output predicted from the first stream by Modality Translation will also be divided into corresponding sets for each chromosome. Then, the input ATAC and output GEX (from Modality Translation) corresponding to each chromosome will be used as inputs for the respective Chromosome Refinement model, resulting in a refined GEX output with higher accuracy.

Table 2 Dataset details after division

Chromosome number	Source dimensions	Target dimensions	Chromosome number	Source dimensions	Target dimensions
1	11497	1235	12	5937	668
2	9123	818	13	2587	215
3	7241	700	14	3938	417
4	4443	454	15	3989	376
5	5573	554	16	4350	527
6	7152	645	17	6256	716
7	5684	559	18	2192	175
8	4605	408	19	5386	865
9	4789	478	20	3463	331
10	5449	464	21	1426	123
11	5833	652	22	2887	290

**Fig. 1** Overview of two-layer architecture of AIM. The Modality Transformation module the overall relationship between ATAC and GEX, while the Chromosome Refinement module is responsible for capture interaction between genes within each chromosome

4.2 Modality transformation

We adopt the ATAC-to-GEX translation model proposed in [19] as it represents the state-of-the-art in the field. The Modality Transformation module comprises two parts: *data preprocessing* and *GEX generation* as shown in Fig. 2. The former pursues two objectives: reducing the dimensionality of the input data and generating handcrafted features to enrich the information. Meanwhile, the latter comprises multiple fully-connected layers tasked with converting from the latent vector space to the GEX representation.

Data Preprocessing. Let $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ represent the set of ATAC vectors of interest, where $\mathbf{v}_i \in \mathbb{R}^T$ denotes the i -th ATAC consisting of T features. The data is partitioned into k batches denoted by $\{B_1, B_2, \dots, B_k\}$. The metadata of the data includes information about batches. Specifically, the metadata indicates that the data was obtained from samples provided by 12 donors. Each donor contributed multiple samples from various locations. The training data will consist of 10 batches from 9 donors, while the test data will comprise 3 batches from 3 donors.

The data preprocessing phase comprises four steps: (1) selecting the most important features, (2) computing the standard deviation of each batch, (3) determining the median of each batch, and (4) calculating the mean of each ATAC vector. The results of these steps are then concatenated to form the final preprocessed data, which is subsequently fed into the GEX generation module to produce the GEX representation. In the first step, for each cell type, the 100 highest-ranked ATAC features are selected. Consequently, with 20 cell types, a 1506-dimensional vector is obtained, from which 494 features that coincide with other features are removed. Moreover, in order to achieve a balance between the quantity of ATAC features and handcrafted features, each value computed in steps (2), (3), and (4) is replicated 50 times before being combined. As a result, the resulting vectors from the preprocessing stage encompass 1656 dimensions.

GEX Generation. The preprocessed ATAC data is subsequently inputted into a Multilayer Perceptron network, which comprises 5 fully-connected layers with layer sizes of 320, 360, 620, 440, and 13431, respectively. The input layer implements a dropout rate of 0.8. The activation function utilized is ReLU, and the model undergoes training for 38 epochs.

4.3 Chromosome refinement

Network Architecture. There are 22 specialized Chromosome Refinement models, each designed to capture the interconnections within individual chromosomes. These models take as input the GEX representation produced by Modality Transformation specific to that chromosome and refine it into a more precise GEX representation. Moreover, the Chromo-

some Refinement models employ an attention mechanism to emphasize the relationships between genes within the ATAC and those within the GEX. Figure 3 illustrates the structure of a Chromosome Refinement model. The inputs to the model include the ATAC vector and the GEX vector predicted by the Modality Translator model on the corresponding chromosome. The ATAC vector is feature-normalized on the entire dataset before feeding into the model. The normalized ATAC and the GEX are fed through the multi-head attention network, with the ATAC acting as Keys and Values, and the GEX serving as the Query. The output of the attention network is then passed through two fully-connected layers with 1024 and 2048 nodes, respectively, before going through the output layer to predict the final output. Each fully-connected layer uses a dropout layer with a dropout factor of 0.3. During the model training process, we use 10% of the data from the training set for validation and use the early stopping method along with the reduced learning rate on plateau to stop training the model when the learning results do not improve to avoid overfitting. We adopt the mean square error (MSE) for the loss function, which is represented as follows:

$$Loss = \frac{1}{B} \sum_{j=1}^B \sum_{i=1}^t \frac{(\hat{y}_i^{(j)} - y_i^{(j)})^2}{t}, \quad (2)$$

where $\hat{y}$ is the predicted vector, y is the corresponding label vector, t is the number of features of both vector $\hat{y}$ and y . B represents the batch size, j is the index representing each data point within the batch. $\hat{y}$ and y have the form $\hat{y}^{(j)} = y_1, y_2, \dots, y_t$ with $y_i \in \mathbb{R}$, $y^{(j)} = y_1, y_2, \dots, y_t$ with $y_i \in \mathbb{R}$ respectively.

Attention Mechanism. Let us denote the Query, Key, and Value of the Attention module as Q , K and V , respectively; $Q \in \mathbb{R}^{d_q}$, $K \in \mathbb{R}^{d_k}$, and $V \in \mathbb{R}^{d_v}$. For each tuple of Q , K and V , we perform a linear transformation using the weight parameters of the model W_q , W_k , W_v as follows.

$$Q' = QW_q^T, \quad K' = KW_k^T, \quad V' = VW_v^T. \quad (3)$$

After that, we determine the scale dot product attention using the following formula.

$$\begin{aligned} & \text{ScaleDotProductAttention}(Q, K, V) \\ &= \text{softmax} \left(\frac{Q'K'^T}{\sqrt{d_k}} \right) V'. \end{aligned} \quad (4)$$

As depicted in Fig. 4, the scaled dot product attention, denoted as $\text{softmax} \left(\frac{Q'K'^T}{\sqrt{d_k}} \right)$, computes the dot product for every possible pair of queries and keys. Essentially, this value indicates the relevance of genes to chromatin accessibility locations. The outcome of this computation is the attention

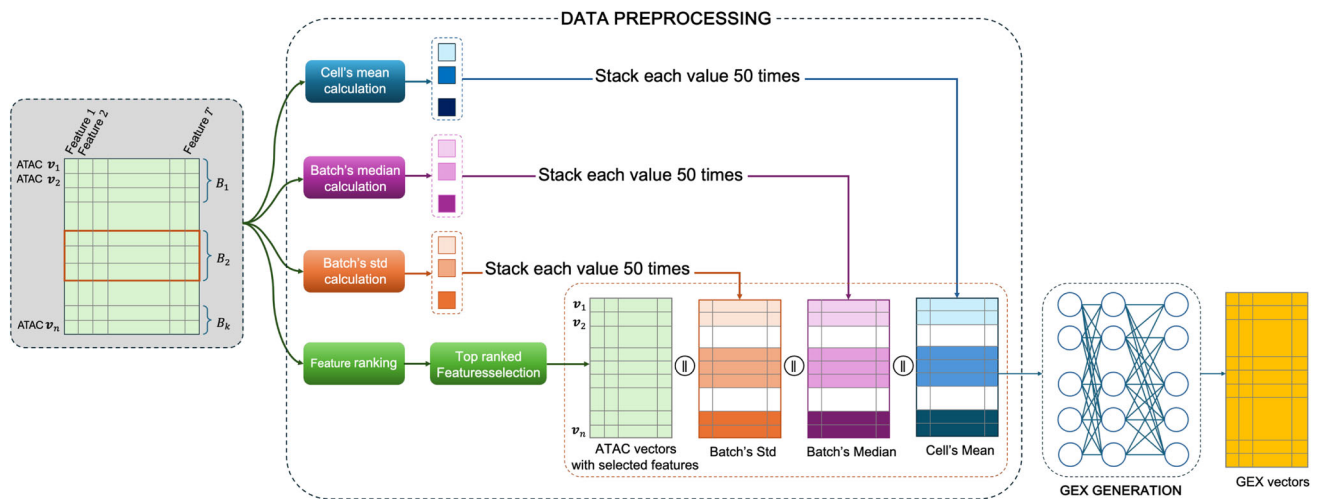
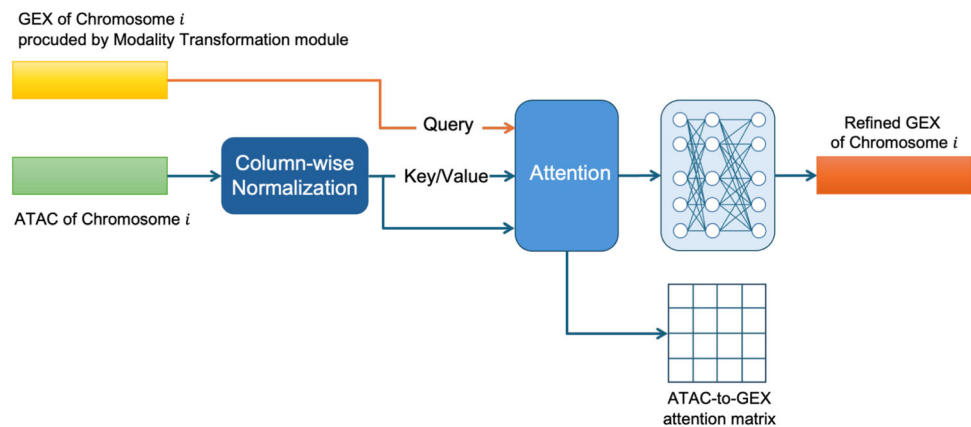


Fig. 2 Structure of the modality transformation module. The Data Preprocessing step is tasked with reducing the dimensions of the input and creating hand-crafted features, whereas the GEX generation step utilizes a fully-connected network to produce GEX representations

Fig. 3 Structure of the chromosome refinement module



matrix, which represents the attention relationship from each gene to each chromatin accessibility location. Multiplying this matrix by the value matrix V' yields the weighted mean vector based on attention. This weighted mean vector preserves the same number of dimensions as the query. In more detail, each GEX value in the Query vector will be multiplied by each Chromatin Accessibility value in the Key vector to obtain the value representing the attention of GEX to those CA positions. As a result, for each GEX value, we get an attention vector that shows the influence of each CA on that GEX. This vector is then softmaxed and multiplied with the Value vector to obtain the GEX value after being refined by attention.

From Eq. (3), the formula for the multi-head attention neural network will have the form:

$$\begin{aligned} \text{Multihead}(Q, K, V) \\ = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_n)W^o, \end{aligned} \quad (5)$$

where $\text{head}_i = \text{ScaleDotProductAttention}(QW_i^q, KW_i^k, VW_i^v)$, $W_{1..h}^q \in \mathbb{R}^{d_k}$, $W_{1..h}^k \in \mathbb{R}^{d_k}$, $W_{1..h}^v \in \mathbb{R}^{d_v}$ and $W^o \in \mathbb{R}^{(h \cdot d_v) \times d_{out}}$ indicates the parameters that can be learned during the training of the network.

5 Evaluation

This section presents the model's results and compares them to some current state-of-the-art models. In addition, we present some of our findings on the attention weight generated by the extended explainable model.

5.1 Experiment settings

This study aims to determine if local gene relationships on chromosomes affect the accuracy of the problem. We are specifically interested in determining whether the separation of genes on chromosomes influences the outcome of the problem.

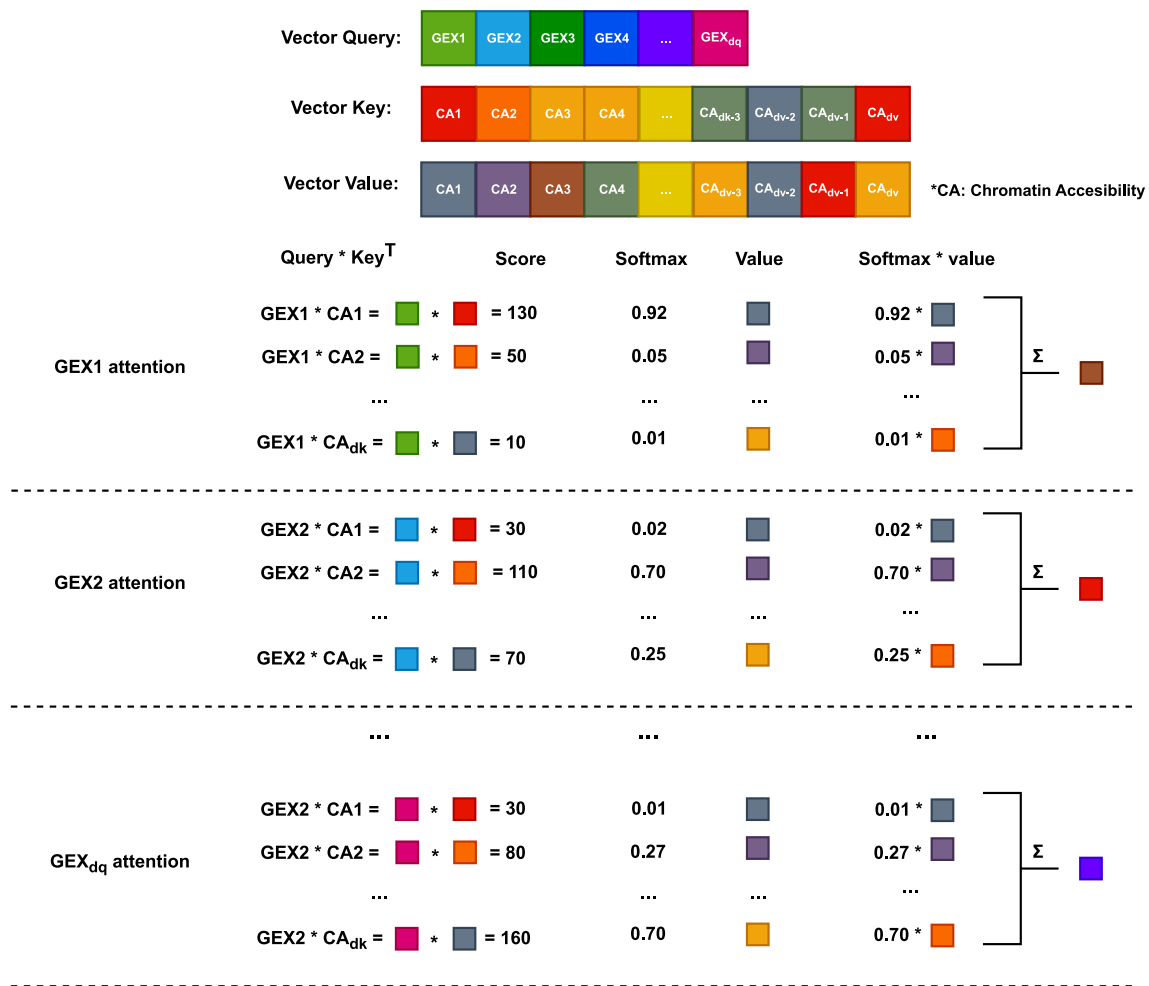


Fig. 4 Illustrates of the process for calculating attention of each gene in the Query to chromatin accessibility in the key

The data set for this study contains 42,492 training samples and 1000 test samples. The input vector has 116490 dimensions, while the output vector has 13431 dimensions. In addition, this data set was split into subsets corresponding to each chromosome, as described in Sect. 3. We used the RMSE metric, which is presented in Eq. 1, to measure the performance of our model.

For comparison and evaluation against our proposed model, we use the baseline model developed by Cajal *et al.*, which won the ATAC-to-GEX prediction task organized by Luecken *et al.* [13]. The Cajal model uses a complex data pre-processing strategy, followed by a neural network with four hidden layers of 320, 360, 620, and 440 nodes, respectively. We also compare our method with several other models that have produced commendable results in the same competition.

We implemented our model on a computing platform that included an NVIDIA GeForce 1080 GPU in addition to an AMD Ryzen 5 1600 CPU and 16GB of RAM. With the PyTorch framework (version 1.13.1) utilizing CUDA (version 11.6) for accelerated computation, the implementation

was done with Python 3.7. We also used a number of libraries, such as NumPy, Scanpy, and Scikit-learn, to help with different parts of data processing.

5.2 Comparison with existing approaches

This section provides a comparative analysis of the results achieved by the Chromosome Refinement model and the most advanced methodologies currently available. Expanding on the structure described in Sect. 4, we integrate a Chromosome Refinement model block to utilize the local information of genes on the same chromosome, thereby improving the predictive abilities of our approach. The results of the Chromosome Refinement model testing process show that the model can inherit the predicted results from the Modality Translator model, thereby improving the final predictive efficiency. As can be seen in Table 3, our proposed model has significantly improved performance compared to the four models with the best results at the present time. All the data here is 11670 genes whose information we found

Table 3 Result of the chromosome refinement model on the test set

Model	RMSE	
	All Chr data	Chr1
Cajal <i>et al.</i> [19]	0.2232	0.2156
Dengk <i>et al.</i> [19]	0.2294	0.2216
Wen <i>et al.</i> [19]	0.2469	0.2338
Novel <i>et al.</i> [19]	0.2631	0.2465
Ours	0.2206	0.2132

Bold text highlights the best results among the compared models

from the NCBI dataset. It should be noted that we cannot compare our model with that of scMoGNN [13] as we were unable to reproduce their results. However, their report indicates that their model's performance was not superior to Cajal *et al.*'s. Therefore, it can be inferred that our proposed model outperforms that of scMoGNN.

In addition, the Chromosome Refinement model not only improves predictive efficiency on Chr1 but on all 22 models corresponding to 22 chromosome pairs. Table 4 shows that the model achieves a significantly smaller RMSE value than the current state-of-the-art, the per-chromosome Cajal model. In addition, all the *t*-statistic values are less than -1.960 , showing that the improvement from chromosome refinement is statistically significant.

6 Attention weight analysis and finding potential pathway groups

In this section, we propose some analyses along with how we search for groups with potential pathways and perform them on the B cell group.

6.1 Marker gene analysis

Before conducting the attention weight analysis between genes with chromatin accessibility locations, we conducted a gene marker analysis of each cell type in the dataset. We used the Scanpy library [26] as our primary tool. We utilized the `rank_genes_groups` function, employing the "t-test_overestim_var" method (which overestimates the variance of each group) and grouping the data by cell type (cell type being the metadata indicating the specific cell type to which the cells in the dataset belong). This function accepts the GEX dataset as an input and produces a gene ranking for each cell type. This ranking comprises the names of genes and their corresponding scores. A higher score indicates a greater likelihood of gene expression in that particular cell type. Upon executing the aforementioned function, we observed that genes belonging to the same groups, such as CD8+ T, CD8+ T Naive, CD4+ T Naive, CD4+ T Activated within the T cell group, exhibit remarkably similar levels

of gene expression. Consequently, we organized these cells into distinct groups and conducted gene marker analyses by changing the grouping option from cell type to cell group. Once we have acquired those rankings, we proceed to choose the top 20 genes with the highest scores for each cell group. These selected genes serve as marker genes for their respective cell groups. The groups of cell types are listed below:

- Group T cell: CD8+ T, CD8+ T Naive, CD4+ T Naive, CD4+ T Activated
- Group B cell: Naive CD20+ B, Transitional B, B1 B
- Group Prog cell: Lymph prog, MK/E prog, G/M prog
- Group Mono cell: CD14+ Mono, CD16+ Mono
- Group Blast cell: Proerythroblast, Erythroblast, Normoblast
- In addition, the remaining cells are considered to belong to independent groups.

Figure 5 lists the top 20 predicted gene markers for each cell group that we obtained after performing gene marker analysis. These gene markers will then be used to analyze the attention weight obtained from the model.

6.2 Attention weight analysis

6.2.1 Attention matrix

The Chromosome Refinement model is not only effective in predicting but also provides a lot of valuable information regarding the relationship between Gene and chromatin accessibility from the model's attention weight. For each cell in the dataset have ATAC input of the form $X = x_1, x_2, \dots, x_n, x_{dk} \in \mathbb{R}$ and GEX input from modality translation model of the form $Y = y_1, y_2, \dots, y_{dq}, y_i \in \mathbb{R}$, the chromosome refinement model produces the output of a predicted GEX vector Y and an attention weight matrix A with $Y' = y'_1, y'_2, \dots, y'_{dq}, y'_i \in \mathbb{R}^T$ and $W =$

$\begin{bmatrix} w_{1,1} & \dots & w_{1,dk} \\ \dots & \dots & \dots \\ w_{dq,1} & \dots & w_{dq,dk} \end{bmatrix}$, $w_{i,j} \in \mathbb{R}$, dk and dq are the dimensions of the input and output vectors of the model, respectively. $w_{i,j}$ is the probability that the expression of gene i is related to the j^{th} chromatin accessibility position. Figure 6 visualizes attention from genes to chromatin accessibility on chromosomes. With the goal of analyzing attention weight for each cell type, we perform the analysis of the attention matrix on the group of cells by averaging the attention matrix of the cells in each group. From that, with each cell group, we get a new attention weight matrix of the same size as the original that carries attention information for the whole group of cells. With 10 different groups of cells, each group of cells is divided into 22 pairs of chromosomes. For each

Table 4 The RMSE and *t*-statistic values of the chromosome refinement model for each chromosome in the test set compared to that of the Cajal model

	Proposed model	Cajal model	<i>t</i> -statistic		Proposed model	Cajal model	<i>t</i> -statistic
Chr1	0.2132	0.2156	−26.544	Chr12	0.2206	0.2231	−23.041
Chr2	0.2443	0.2479	−32.441	Chr13	0.2610	0.2632	−11.859
Chr3	0.2312	0.2348	−28.126	Chr14	0.2265	0.2295	−23.031
Chr4	0.2378	0.2400	−12.416	Chr15	0.2370	0.2386	−11.057
Chr5	0.2457	0.2488	−21.746	Chr16	0.2054	0.2062	−5.297
Chr6	0.2418	0.2452	−24.700	Chr17	0.2008	0.2037	−29.148
Chr7	0.2246	0.2275	−24.373	Chr18	0.2416	0.2435	−11.517
Chr8	0.2319	0.2342	−17.703	Chr19	0.1616	0.1652	−39.741
Chr9	0.2127	0.2148	−20.080	Chr20	0.1971	0.1987	−7.171
Chr10	0.2305	0.2337	−26.042	Chr21	0.2203	0.2223	−9.478
Chr11	0.2164	0.2170	−2.271	Chr22	0.1787	0.1816	−13.245

Bold text highlights the best results among the compared models

Table 5 Correlation of marker genes in the same chromosome. Chromosomes 2, 3, 7, 8, 10, 11,..., 22 have less than two marker genes so they are not listed in this table

Chromosome	Gene marker 1	Gene marker 2	Correlation value
Chr1	FCRL1	RALGPS2	0.9594
Chr4	ARHGAP24	BANK1	0.9560
	ARHGAP24	CAMK2D	0.9893
	BANK1	CAMK2D	0.9843
Chr5	CD74	EBF1	0.9992
Chr6	HLA-DRA	KHDRBS2	0.9763
	HLA-DRA	COL19A1	0.9668
	HLA-DRA	BACH2	0.8392
	HLA-DRA	PLEKHG1	0.9703
	KHDRBS2	COL19A1	0.9980
	KHDRBS2	BACH2	0.8027
	KHDRBS2	PLEKHG1	0.9737
	COL19A1	BACH2	0.8013
	COL19A1	PLEKHG1	0.9622
	BACH2	PLEKHG1	0.7557
Chr9	PAX5	ZCCHC7	0.9814

pair of chromosomes, we have an average attention matrix, from which we have 220 average attention matrices.

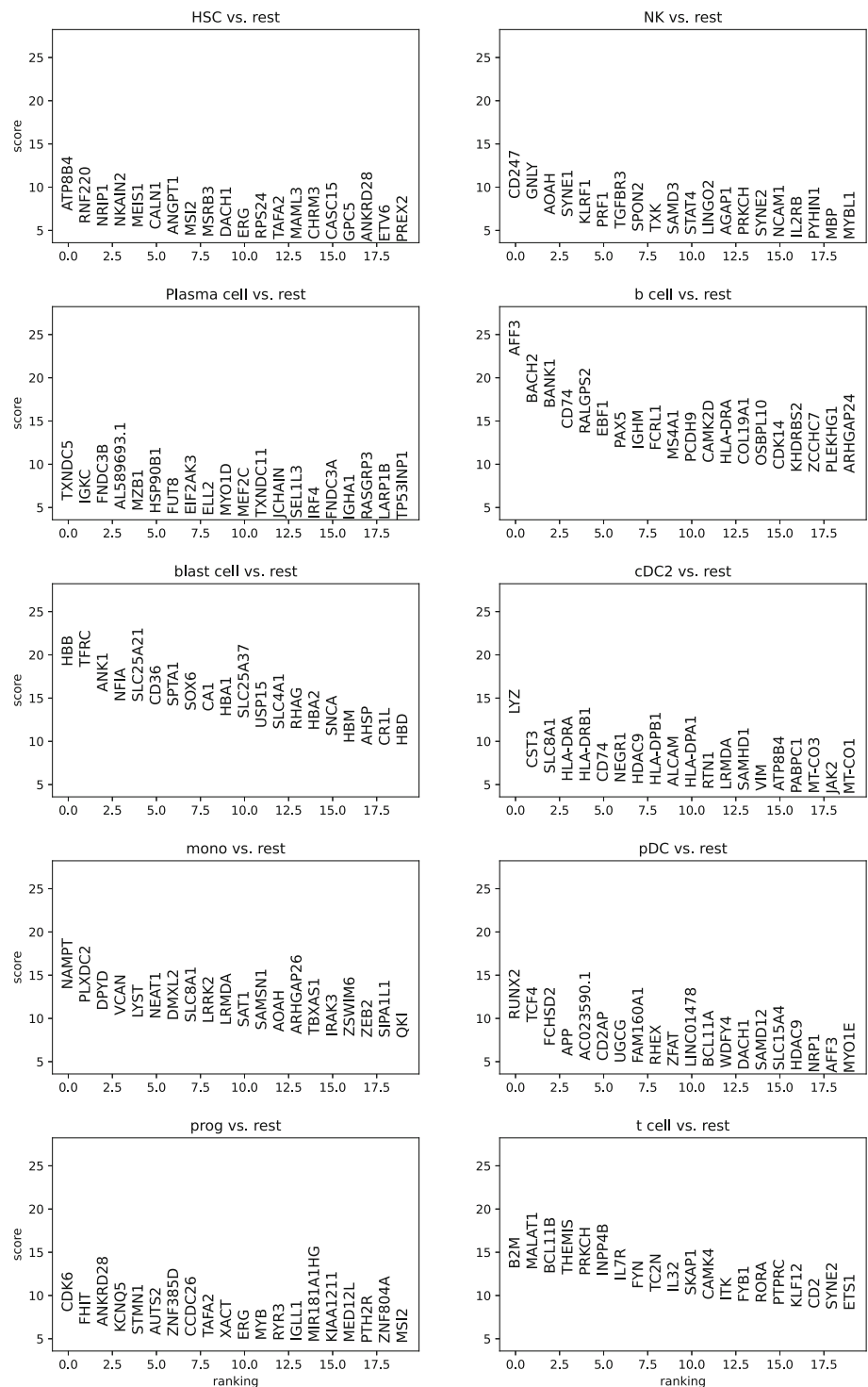
6.2.2 Attention weight analysis

In this section, we will only propose analytical directions and perform analyses on the B cell group. The main goal of this analysis is to answer these questions: which chromatin accessibility (CA) locations in the same chromosome most impact the expression of marker genes: those on or near genes, or those outside genes? Do all CAs have an equal effect, or do just a few CAs have the most significant effect on marker gene expression?

To answer the question of which chromatin accessibility (CA) locations in the same chromosome most impact the

expression of marker genes: those on or near genes, or those outside genes? First, we define CAs located on or near genes as CAs whose start and end points intersect or lie inside the start and end points of the gene. In addition, to include promoter sites of the gene, the start points of the gene will be extended by 1500 nucleotides. Then we will calculate the percentage of CAs in the genes in the top *k*%. The pseudocode for this computation is presented in Algorithm 1. The algorithm input is a list of attention matrices, a list of gene markers, and a list of CAs in genes. where the list of CAs in the gene is obtained from Algorithm 2. Specifically, for each gene marker in group B cell, we extract the attention vector of that gene from the attention matrix, rearranging them in descending order of attention value. From there, we take the top *k*% of positions with the highest attention (*k* gradually

Fig. 5 Score evaluates gene markers of genes on each cell type



increases from 0.1 to 100) and then consider how many of those $k\%$ of CAs with the highest attention are located on or near genes. Figure 7 depicts the % of CAs located on or near genes in the top $k\%$ of CAs with the highest attention when k increases from 0.1 to 100.

It can be seen from Fig. 7 that gene markers like OSBPL10, CDK14, AFF3, etc. have a graph showing that the percent-

age of CAs located on the gene does not change and only fluctuates around 50% when k increases from 0.1 to 100, because this is the ratio of CAs in the gene on the corresponding chromosome. Meanwhile, gene markers such as CD74, EBF1, CAMK2D, BANK1, ARHGAP24, RALGPS2, and FCRL1 have a very large value in the k range from 0.1 to 20. This is clearly shown by the fact that, in the range of top

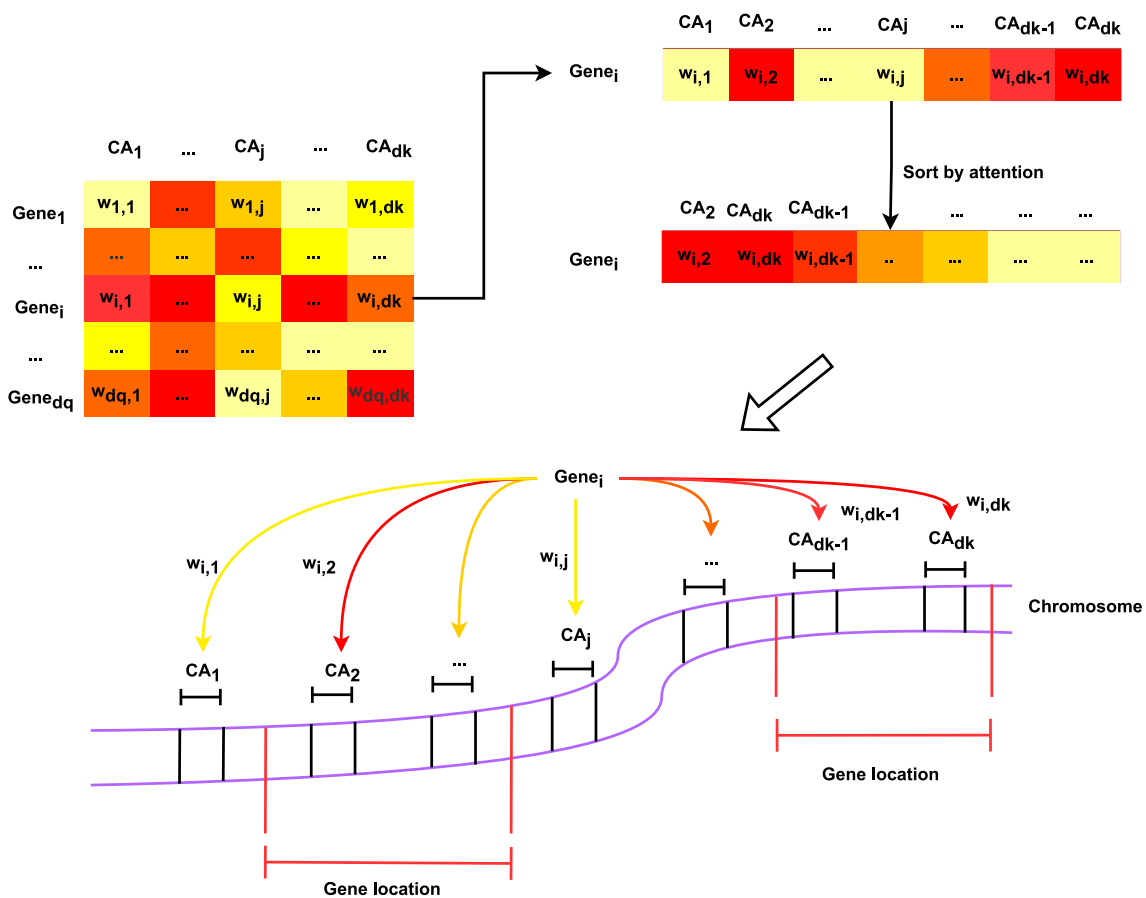


Fig. 6 Illustrates the attention matrix and visualize attention from Gene to chromatin accessibility on chromosomes. Some gene marker give a lot more attention to CAs located in gene

0.1–20%, the CAs that receive the most attention, accounting for over 90–70% of them, are located in or near the gene. This indicates that CAs located on or near genes have a great influence on the gene expression of these marker genes. For a more specific example, with the gene marker EBF1, in the top 4% of CAs have the highest attention value (a total of 5573 CAs), i.e., 223 positions. Of these, up to 95%, or 211 CAs, are located on or near genes in the same chromosomes. This proves that these positions play a very important role in regulating gene expression for this gene marker. It is possible that these positions are promoters, enhancers, or other important regulatory positions for the above gene markers that require further experimental research to analyze.

The next question is whether all CAs have an equal effect, or if only a few CAs have the most significant effect on marker gene expression. The pseudocode for this computation is presented in Algorithm 3. Specifically, we also perform attention vector extraction for each gene marker and sort in descending order according to the attention value. We then take the sum of the attention coefficients of the $k\%$ of top CAs with the highest attention; k runs from 0.1 to 100. Figure 8 depicts the total attention of the top $k\%$ CAs with the highest atten-

tion. It can be seen that the special genes BACH2, AFF3, ZCCHC7, MS4A1, and PAX5 appear to only pay attention to a small number of CA positions when the total attention of the top 10% of CAs with the highest attention of these genes has occupied from 70% to nearly 100% of all attention. This suggests that there are genes that are actually only interested in a small number of specific CAs. These positions may play an important role in regulating the expression of the above gene markers. Experimental studies are needed to obtain more information about these special locations.

6.3 Finding potential pathways group

To search for genes with potential pathways or co-expression, we assume that genes with the same attention vector are likely to co-express each other or have certain pathway interactions. To evaluate the “similarity” of two attention vectors, we use the Pearson Correlation Coefficient parameter. The Pearson Correlation Coefficient value of two attention vectors $G = g_1, g_2, \dots, g_{dk}$ and $H = h_1, h_2, \dots, h_{dk}$ with $g_i, h_i \in \mathbb{R}, i = 1, 2, \dots, dk$ is calculated by Eq. 6, where dk is the length of the attention vector, $\bar{G}$ and $\bar{H}$ is mean of

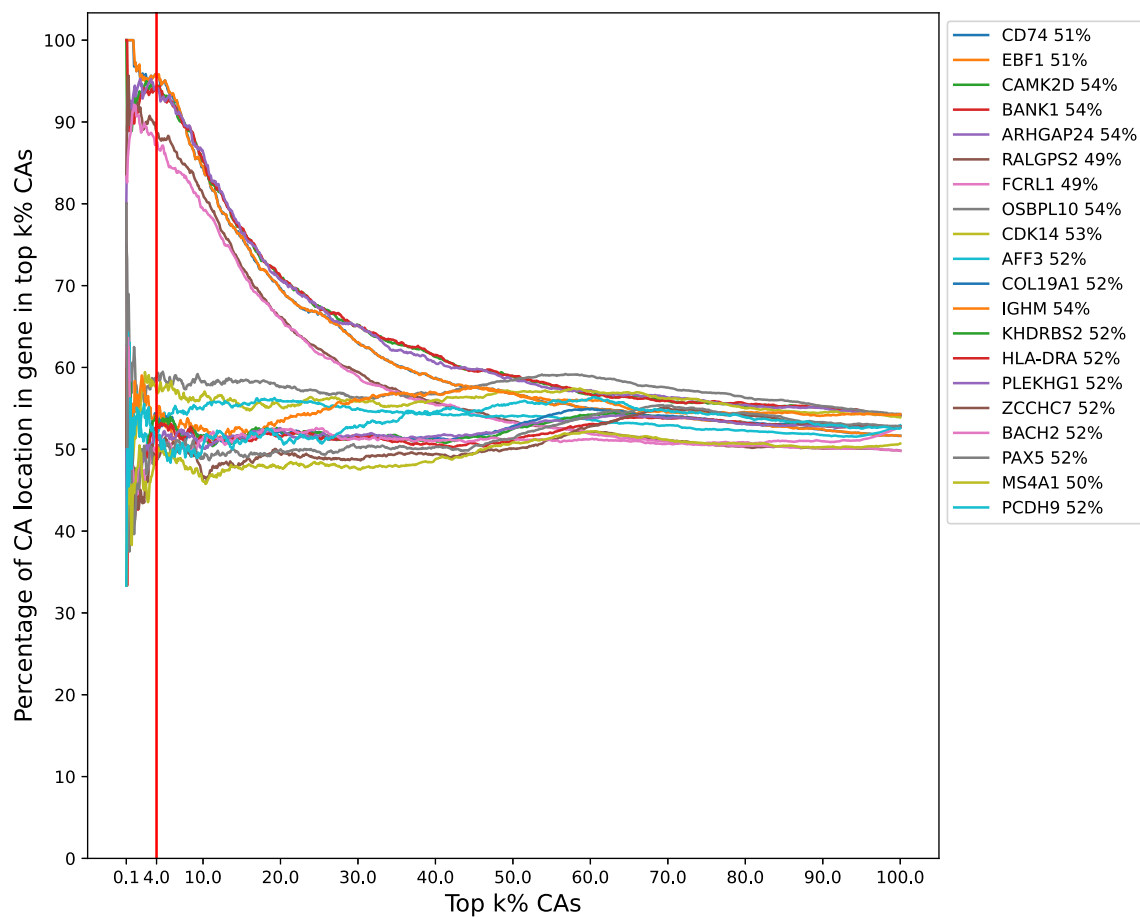


Fig. 7 Represents the change in the percentage of CA locations in genes in the top k% of CAs with the highest attention. k varies from 0.1 to 100. The percentage next to the gene names in the legend section represents

the default value of the percentage of CAs located on the gene in the chromosome containing that gene marker

vector G and H , respectively. This Pearson correlation coefficient has a value ranging from -1 to 1 . -1 represents that the two vectors G and H have absolutely no linear correlation with each other, and 1 represents that these two vectors have a strong correlation with each other. Genes with higher correlations with gene markers indicate that genes have the potential to co-express or pathways with gene markers. This is clearly shown by the fact that gene markers located on the same chromosome have a very high correlation with each other, reaching values of 0.8 or more (Table 5). From there, for each gene marker, we calculate the correlation value of all genes on the same chromosome to that gene marker. Then, we filter out genes with a correlation value of 0.8 or more to the gene marker and form groups (Figure 9). The pseudocode for this computation is presented in Algorithm 4. During the analysis process, gene markers on the same chromosome almost always have a correlation value of 0.8 or more, and the groups found are almost identical. Therefore, we grouped these groups together within a chromosome and

obtained potential gene groups according to each chromosome.

$$r_{GH} = \frac{\sum_{i=1}^{dk} (g_i - \bar{G})(h_i - \bar{H})}{\sqrt{\sum_{i=1}^{dk} (g_i - \bar{G})^2} \sqrt{\sum_{i=1}^{dk} (h_i - \bar{H})^2}}. \quad (6)$$

In this section, we will verify the hypothesis that genes with high correlations with gene markers have potential pathways with that gene marker. First, we perform a gene ranking maker as described in 6.1 to find gene markers for the B cell group and select 20 gene markers with the highest ranking. We observed that most marker genes exhibit a very high correlation with each other, with values typically reaching 0.95 or higher. Notably, the BACH2 marker gene, located on chromosome 6, is the only exception, displaying a lower correlation with other marker genes (Table 5). Then we ran Algorithm 4 with the 20 found gene markers and group of genes with high correlation with the gene markers found presented in Table 6. In the gene groups found, we found

Fig. 8 Visualize the total attention of the top k% CAs with the highest attention, k runs from 0.1 to 100

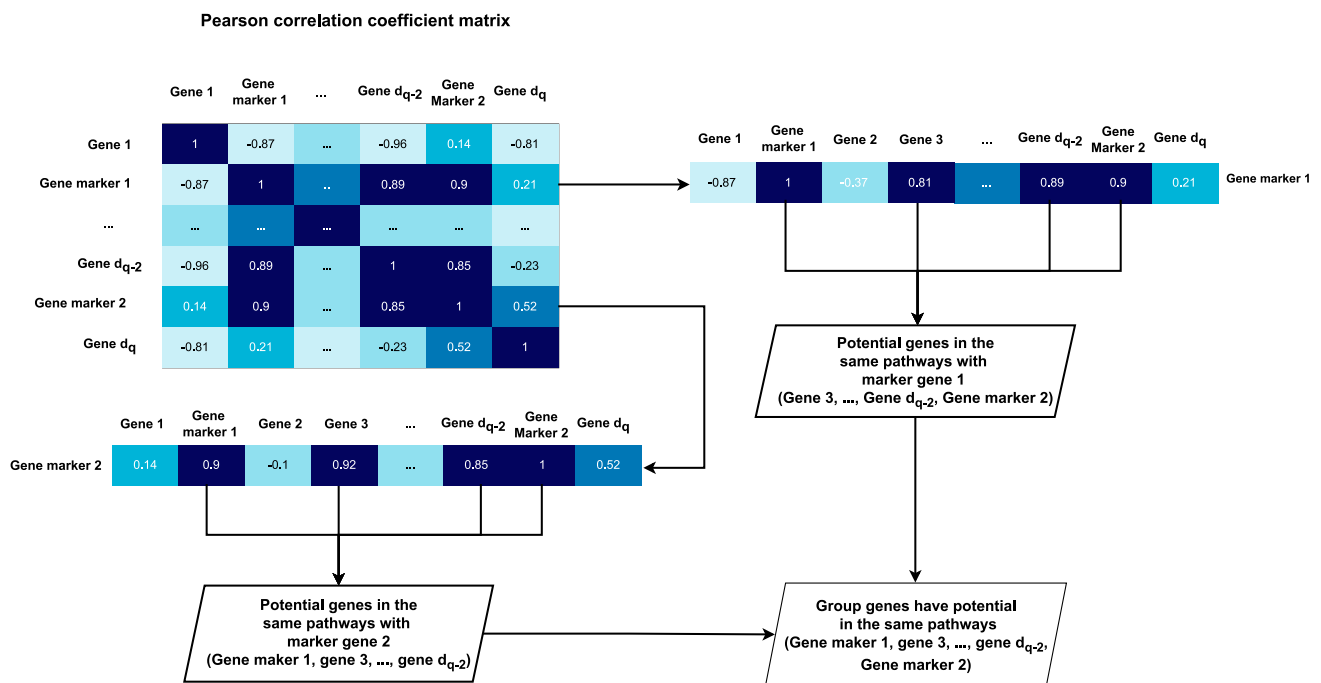
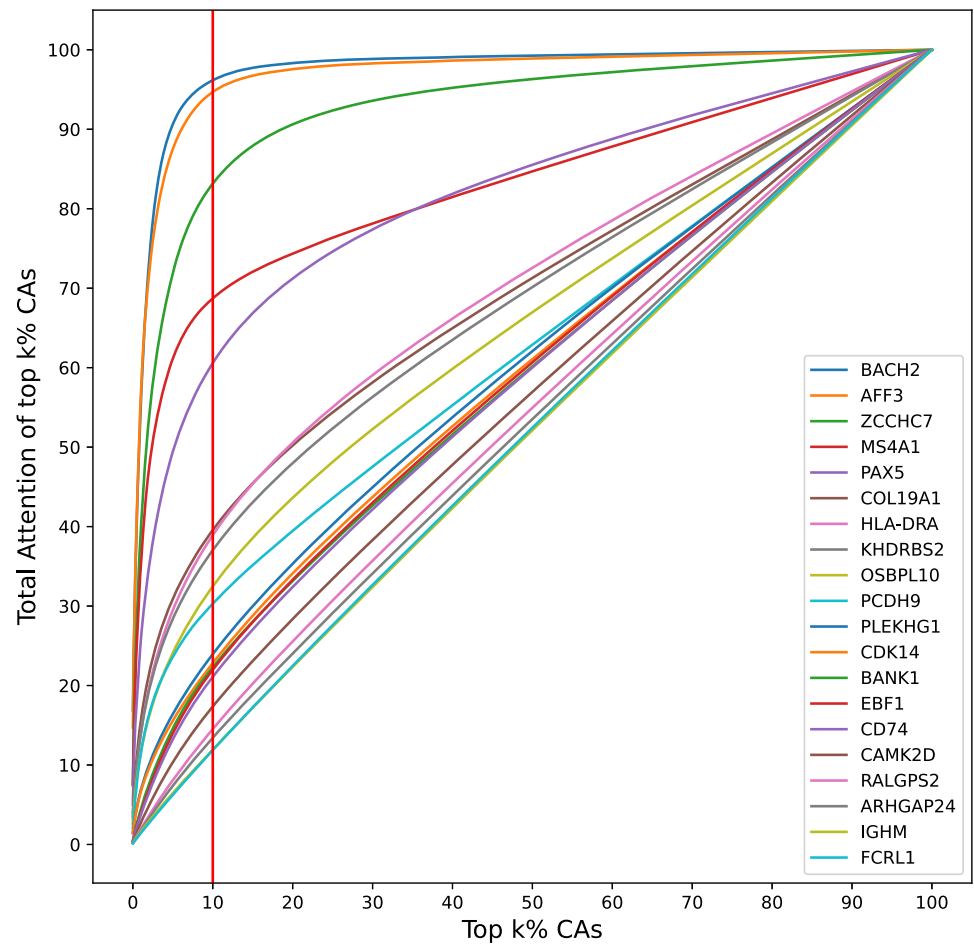


Fig. 9 Illustrates the process of finding groups of genes with potential pathways with marker genes in the same chromosome

Table 6 Group of genes with high correlation with gene markers of B cell group on each chromosome.

Chromosome 1	FCRL1 , RABGAP1L , RERE, VAV3, PRDM2, TXNIP, RPS27, DPYD, CD55, RPL11, COP1, EIF4G3, PDE4B, MIR181A1HG, SMYD3, RALGPS2, LYST, ASH1L, RCSD1, JUN, PTPRC, JAK1, LAPTM5, FAF1, ARID4B, RPS8, PRRC2C, ZNF644, AKT3, H3F3A, MACF1, USP24, SRSF11, TUT4
Chromosome 2	AFF3 , ARHGAP15, ANKRD44, PRKCE, MGAT5
Chromosome 3	OSBPL10 , ST6GAL1, LPP, ITPR1, BBX, UBE2E2, TBC1D5, TGFB2, ACAP2, RASA2, ZBTB20, RPL32, MBNL1, FOXP1, RPL35A, TRA2B, RPSA, CBLB, CCNL1, ATP11B, SUCLG2, RBM6
Chromosome 4	ARHGAP24 , CAMK2D , BANK1 , MARCH1, LRBA, PPP3CA, TMEM131L, ATP8A1, CCSER1, STIM2, RPL34, SEL1L3, ANKRD17, FBXW7, RHOH, BANK1 , FRYL, TBC1D1, NFKB1, RPS3A, BMP2K, ARAP2, RAP1GDS1, ANTXR2, CNOT6L, RPL9, AFF1, TET2, GAB1, PDS5A, TAPT1, MANBA, SCFD2
Chromosome 5	CD74 , EBF1 , MEF2C, SSBP2, PDE4D, DOCK2, LIX1-AS1, NIPBL, FAM172A, TNFAIP8, NR3C1, CDC42SE2, ARL15, TRIO, RPS23, MAP3K1, FBXL17, PARP8, RPS14, ADAM19, AP3B1, SNX2, ERBIN, ZSWIM6, DMXL1, RICTOR, RPL37, CYFIP2, COMMD10, IQGAP2, MARCH3, RAPGEF6, DUSP1, ANKHD1, ARHGAP26, CCNH, MAN2A1, WDR70, AFF4, CAST, HNRNPH1, CERT1, VCAN
Chromosome 6	HLA-DRA , KHDRBS2 , PLEKHG1 , COL19A1 , BACH2 , CDKAL1, EZR, HLA-DRB1, LMBRD1, RPS18, ZFAND3, CD83, BTBD9, HLA-DQA1, RIPOR2, HLA-B, SNX9, STX7, EEF1A1, JARID2, HLA-A, RPS12, ELOVL5, UTRN, PHIP, HLA-E, PTPRK, ZNF292, CCND3, PNISR, HLA-C, MAP3K5, ARID1B, SENP6, ASCC3, NEDD9, HLA-DPB1, BCLAF1, RNGTT, FARS2, CEP85L, PHF3, QKI, GMD5, UST
Chromosome 7	CDK14 , IMMP2L, IKZF1, ELMO1, HDAC9, AUTS2, EXOC4, JAZF1, SRPK2, KMT2E, ACTB, COG5, KMT2C, BRAF
Chromosome 9	PAX5 , KDM4C, STRBP, RFX3, DOCK8, SYK, RPS6, RAPGEF1, RPL7A, RIC1, TLE4, PIP5K1B, UHRF2, CEMIP2
Chromosome 11	MS4A1 , ZCCHC7 , PACS1, CD44, ATM, SIK3, UVRAG, TTC17, FAU, RPLP2, SWAP70, SESN3, BIRC3, STIM1, RPS3, FCHSD2, PPP6R3, RAB30, FLI1, PAN3, PHF21A, RPS25, RPL27A, HBB, PICALM, ELP4, KDM2A, RRAS2, MALAT1, FBNP4, CELF1, KMT2A, RSF1, RPS13, APLP2
Chromosome 13	PCDH9 , FOXO1
Chromosome 14	IGHM , HIF1A, RAD51B, FOXN3, RPS29, SIPA1L1, PPP2R5C, HSP90AA1, HNRNPC, SRSF5

Genes highlighted in bold in groups are gene markers located on that chromosome

that many genes were predicted in the top 100 gene markers (Table 7), proving that these genes are very likely to co-express with gene markers. In addition, in the gene groups found, there are many genes that have been researched to prove that they have pathways to marker genes (Table 8) in that group, according to GeneAnalytics [27].

Regarding the group of marker genes of B lymphocyte cells that we analyzed, including RALGPS2, FCRL1, BANK1, CAMK2D, ARHGAP24, CD74, EBF1, OSBPL10, CDK14, AFF3, COL19A1, IGHM, KHDRBS2, HLA-DRA, PLEKHG1, ZCCHC7, BACH2, PAX5, MS4A1, PCDH9, our results are in line with previous studies underscoring the essential functions of these genes in B lymphocyte activity. For instance, FCRL1 may play an important role in the activation of human B-cell responses and has the potential to serve as a target for immunotherapy of B-cell-related disorders [28]. The expression of FCRL1 occurs on both naive and memory B cells and stands out for its distinctive intracellular co-activation potential [29].

In addition, the role of the CD74 gene in B cell activity cannot be ignored. CD74 is a gene that encodes a transcriptional regulatory factor in normal B cells. It plays a direct role in shaping B cell components by controlling the survival of mature B cells [30]. Furthermore, the activated CD74 protein can interact in the cytoplasm with transcription factors, specifically PAX5, to form a complex. This complex binds

to chromatin at a significantly higher frequency compared to its binding in CLL cells. Consequently, a large portion of the genes bound by this complex experience downregulation in expression within malignant cells. Moreover, the CD74-ICD:PAX5 complex binds to promoter regions of the tumor-suppressor gene DMTF1, leading to a reduction in its expression through transcription inhibition. These discoveries may unveil new therapeutic pathways implicated in oncogenic transformations and potential targets for future medical interventions. In our study, predictions of the interactions of these genes have also been mentioned in Table 6.

Some typical studies on the above pairs of pathways can include the results of the pair of pathways between ARHGAP24 and ARAP2, which were pointed out by Muller et al. [31]. ARHGAP24 is one of the gene markers we found, and ARAP2 is one of the potential gene pathways with ARHGAP24 found through the attention matrix, in which ARAP2 has a correlation value of 0.87 with ARHGAP24. These two genes, which are part of the RHO (RAS homolog) family, are essential components of the RAS superfamily of guanine nucleotide-binding proteins [32]. For nearly two decades, Rho family GTPases have been acknowledged as vital signal-transducing proteins with significant roles in various cellular processes such as cell division, survival, migration, and adhesion. Current understanding indicates that Rho family GTPases transmit signals

Table 7 The genes in the high correlation group were found to be in the top 100 predicted marker genes of B cell

	Gene name	Top marker gen predicted	Gene name	Top marker gen predicted
Chr1 group	RCSD1	52		
Chr2 group	PRKCE	38	MGAT5	39
Chr3 group	UBE2E2	30	FOXP1	73
	ITPR1	58		
Chr4 group	MARCH1	44	TAPT1	82
	STIM2	94		
Chr5 group	MEF2C	48	LIX1-AS1	42
	SSBP2	41	TRIO	62
	PDE4D	54	ADAM19	25
Chr6 group	HLA-DRB1	21	HLA-DQA1	71
	STX7	53		
Chr7 group	HDAC9	31	AUTS2	91
Chr9 group	STRBP	32	SYK	29
Chr11 group	UVRAG	99	FCHSD2	35
	SWAP70	57	RAB30	78
	BIRC3	79		

Table 8 Pathway pairs in B cells have been studied and proven by GeneAnalytics [27]

chromosome	Marker gene	Gene interacts with
Chr1	FCRL1	PTPRC
	RALGPS2	VAV3
Chr4	ARHGAP24	ARAP2
	CAMK2D	RPS3A
Chr5	CD74	DOCK2, MAP3K1
	EBF1	HNRNPH1
Chr6	HLA-DRA	HLA-DRB1, HLA-DQA1, HLA-B, EEF1A1, HLA-A, HLA-E, HLA-C, HLA-DPB1
	KHDRBS2	QKI
	BACH2	FARS2
Chr7	CDK14	BRAF
Chr9	PAX5	SYK, TLE4

from antigen, chemokine, and cytokine receptors, along with adhesion molecules and pattern recognition receptors, serving as key hubs for intercommunication among different signaling pathways [33, 34].

With the HLA-DRA gene marker in chromosome 6, we also found that HLA-DRB1 has a correlation with HLA-DRA reaching 0.972 and has been proven to have a pathway interaction in literature [35]. HLA-DR [35] is a major histocompatibility complex (MHC II), which is expressed on the surface of various antigen-presenting cells such as B lymphocytes, Dendritic cells, and monocytes/macrophages and plays a central role in the immune and response systems by presenting peptides derived from extracellular proteins, especially those derived from pathogens, to T cells. A protein called a beta-chain made from the HLA-DRB1 gene binds

to an alpha-chain protein made from the HLA-DRA gene. These two proteins bind to form a functional protein complex known as the HLA-DR antigen-binding heterodimer. This complex presents foreign antigens to the immune system to trigger the body's immune response. Changes in the HLA-DR molecule have been shown to be associated with an increased risk of rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), and other diseases [36].

Regarding chromosome 9, we found that SYK has the potential to interact with the PAX5 gene marker since the correlation value between these two genes reached 0.899. Inagaki *et al.* [37] have reported this pair of pathways, where the interaction between the SYK and PAX5 genes regulates the differentiation of B cells into plasma cells. PAX5 is an important transcriptional repressor involved in B-cell devel-

Algorithm 1 Calculate percentage of CAs in gene for each marker gene

Require: *list_attention_matrix*: list of 22 attention matrix corresponding to 22 chromosome
Require: *list_gene_markers*: list gene markers
Require: *list_CAs_in_gene*: list CAs located in gene
Ensure: *percentage_of_CAs_in_gene*: Percentage of CAs in or gene for each marker gene
Initialize an empty dictionary: *percentage_of_CAs_in_gene* $\leftarrow \{\}$
for *gene_marker* **in** *gene_markers_list* **do**
 Load corresponding attention matrix: *attention_matrix* $\leftarrow$ load(*list_attention_matrix*, *gene_marker*)
 Extract attention vector for *gene_marker* and rearrange attentions in descending order: *attention_vector* $\leftarrow$ *attention_matrix*[*gene_marker*].sort_values(descending = True)
 for *k* **from** 0.1 **to** 100 **step** 0.1 **do**
 $h \leftarrow k \times \frac{\text{len}(\text{attention_vector})}{100}$
 Filter top $k\%$ CAs: *filtered_top_CAs* $\leftarrow$ *attention_vector*[:*h*]
 Get list name of CAs: *filtered_top_CAs_name* $\leftarrow$ *filtered_top_CAs.index*
 init variable *j* for counting: *j* $\leftarrow$ 0
 for *CA* **in** *filtered_top_CAs_name* **do**
 if *CA* **in** *list_CAs_in_gene* **then**
 j $\leftarrow j + 1$
 end if
 end for
 $\text{percentage} \leftarrow \frac{j}{h}$
 percentage_of_CAs_in_gene[*gene_marker*].append(*percentage*)
 end for
end for
return *percentage_of_CAs_in_gene*

Algorithm 2 Get list CAs located near or in gene for chromosome

Require: *list_CAs*: list CAs, each CA have start and stop location in chromosome
Require: *list_genes*: list genes, each gene have start and stop location in the chromosome
Ensure: *list_CAs_in_gene*: list CAs located in or near gene
Initialize list CAs in gene: *list_CAs_in_gene* $\leftarrow \{\}$
for *CA* **in** *list_CAs* **do**
 get start and stop location of CA: *CA_start*, *CA_stop* $\leftarrow$ *CA*[0], *CA*[1]
 for *gene* **in** *list_genes* **do**
 get start and stop location of gene: *gene_start*, *gene_stop* $\leftarrow$ *gene*[0], *gene*[1]
 if **not** (*CA_stop* < (*gene_start* + 1500) **or** *CA_start* > *gene_stop*) **then**
 append CA to list CAs in gene: *list_CAs_in_gene.append(CA)*
 end if
 end for
end for
return *list_CAs_in_gene*

Algorithm 3 Calculate total attention of top CAs for each marker gene

Require: *list_attention_matrix*: list of 22 attention matrix corresponding to 22 chromosome
Require: *list_gene_markers*: list gene markers
Ensure: *total_attention_of_CAs_in_gene*: Total attention of top CAs for each marker gene
Initialize an empty dictionary: *total_attention_of_top_CAs* $\leftarrow \{\}$
for *gene_marker* **in** *gene_markers_list* **do**
 Load corresponding attention matrix: *attention_matrix* $\leftarrow$ load(*list_attention_matrix*, *gene_marker*)
 Extract attention vector for *gene_marker* and rearrange attentions in descending order: *attention_vector* $\leftarrow$ *attention_matrix*[*gene_marker*].sort_values(descending = True)
 for *k* **from** 0.1 **to** 100 **step** 0.1 **do**
 $h \leftarrow k \times \frac{\text{len}(\text{attention_vector})}{100}$
 Filter top $k\%$ CAs: *filtered_top_CAs* $\leftarrow$ *attention_vector*[:*h*]
 total_attention $\leftarrow$ sum(*filtered_top_CAs*)
 total_attention_of_top_CAs[*gene_marker*].append(*total_attention*)
 end for
end for
return *total_attention_of_top_CAs*

Algorithm 4 Automatically find potential pathway group

Require: *list_attention_matrix*: list of 22 attention matrix corresponding to 22 chromosome
Require: *list_gene_markers*: list gene markers
Ensure: *potential_pathway_group*: Potential pathway group
Initialize an empty dictionary: *potential_pathway_group* $\leftarrow \{\}$
for *gene_marker* **in** *list_gene_markers* **do**
 Load corresponding attention matrix: *attention_matrix* $\leftarrow$ load(*list_attention_matrix*, *gene_marker*)
 Calculate Pearson correlation coefficient for all possible pairs: *correlation_matrix* $\leftarrow$ *attention_matrix*.corr()
 Extract vector correlation of *gene_marker*: *correlation_vector* $\leftarrow$ *correlation_matrix*[*gene_marker*]
 Set threshold: *threshold* $\leftarrow$ 0.8
 for *gene* **in** *correlation_vector.index* **do**
 if *correlation_vector*[*gene*] > *threshold* **then**
 potential_pathway_group[*gene_marker*].append(*gene*)
 end if
 end for
end for
return *potential_pathway_group*

opment. SYK induces tyrosine phosphorylation of PAX5, rendering it incapable of controlling and suppressing the growth of plasma cells. Instead, the process by which B cells differentiate into plasma cells is activated to produce antibodies that respond to foreign stimuli.

Our results also suggest that some pathways involved in gene expression on B cells can be further investigated, laying the foundation for future biological applications, such as immunotherapy in cancer treatment.

To summarize, our study provides the following key contributions:

1. We proposed the segmentation of ATAC and GEX datasets based on chromosomes. Specifically, for ATAC data, the ATAC features are pre-annotated by their genomic locations and associated chromosomes, enabling straightforward grouping by these identifiers. For GEX data, we collected information on gene locations from the latest assembly version GRCh38.p14 in the NCBI database [25] to assign genes to their corresponding chromosomes. This process allows GEX features (genes) to be grouped by chromosome. This segmentation reduces dataset complexity, enabling models to bypass dimensionality reduction steps that often result in information loss during training. Furthermore, this approach leverages local chromosomal information more effectively, as numerous studies have demonstrated that genes on the same chromosome tend to influence each other more than those located on different chromosomes.
2. We proposed and developed the AIM architecture, consisting of two layers: Modality Transformation and Chromosome Refinement. The primary contribution lies in the Chromosome Refinement layer, which leverages comprehensive chromosome-specific data and the predictions from the Modality Transformation layer to refine these predictions with greater accuracy. Furthermore, through a novel approach to utilizing the attention layer, the model not only achieves improved predictive accuracy but also generates an attention matrix that reveals the relationships between ATAC and GEX data. These insights are significant and provide valuable explanations or clarifications regarding the interactions between these two data modalities.
3. We analyzed the attention matrices derived from the Chromosome Refinement layer to uncover insights into the relationship between GEX and ATAC data. We proposed a methodology for identifying gene groups with potential pathway interactions and applied this procedure to B cells. Among the gene groups identified with potential pathway interactions, many have been validated by prior studies, as cited in this work. This reinforces the credibility of the identified gene groups and highlights the need for further research to elucidate their interactions in greater detail.

7 Conclusion

This study introduced AIM, an algorithm designed to address the ATAC-to-GEX translation problem. AIM employs a two-layer architecture to model global relationships and inter-chromosomal interactions, aiming to improve the accuracy of GEX representation. Experimental results indicate that AIM achieves better performance compared to existing models. Additionally, the integration of an attention mechanism enhances interpretability, offering insights into poten-

tial gene sets with pathways or co-expression interactions. Future work will focus on further refining AIM's predictive capabilities and conducting a more detailed analysis of the attention weights. This will help uncover additional relationships between genes, pathways, and chromatin accessibility, contributing to the broader understanding of multimodal data integration.

Acknowledgements The work of BPN was supported by the Faculty Strategic Research Grant (FSRG) number 411494 at Victoria University of Wellington (VUW) and the Endeavour Fund (Smart Ideas) from the New Zealand Ministry of Business, Innovation and Employment (MBIE) under contract VUW RTVU2301.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Klemm SL, Shipony Z, Greenleaf WJ (2019) Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 20(4):207–220. <https://doi.org/10.1038/s41576-018-0089-8>
2. Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, Sexton DW, Holmes C (2013) Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol* 31(8):748–752. <https://doi.org/10.1038/nbt.2642>
3. Albert FW, Treusch S, Shockley AH, Bloom JS, Kruglyak L (2014) Genetics of single-cell protein abundance variation in large yeast populations. *Nature* 506(7489):494–497. <https://doi.org/10.1038/nature12904>
4. Nofech-Mozes I, Soave D, Awadalla P, Abelson S (2023) Pan-cancer classification of single cells in the tumour microenvironment. *Nat Commun*. <https://doi.org/10.1038/s41467-023-37353-8>
5. Kim B-H, Yu K, Lee PCW (2019) Cancer classification of single-cell gene expression data by neural network. *Bioinformatics* 36(5):1360–1366. <https://doi.org/10.1093/bioinformatics/btz772>
6. Ritchie MD, Davis JR, Aschard H, Battle A, Conti D, Du M, Eskin E, Fallin MD, Hsu L, Kraft P, Moore JH, Pierce BL, Bien SA, Thomas DC, Wei P, Montgomery SB (2017) Incorporation of biological knowledge into the study of gene-environment interactions. *Am J Epidemiol* 186(7):771–777. <https://doi.org/10.1093/aje/kwx229>
7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceed Natl Acad Sci* 102(43):15545–15550. <https://doi.org/10.1073/pnas.0506580102>

8. Mootha VK, Lindgren CM, Eriksson S, Puigserver P, Carlsson E, Ridderstråle M, Laurila Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC (2003) PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34(3):267–273. <https://doi.org/10.1038/ng1180>
9. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 14(9):865–868. <https://doi.org/10.1038/nmeth.4380>
10. Chen S, Lake BB, Zhang K (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 37(12):1452–1457. <https://doi.org/10.1038/s41587-019-0290-0>
11. Prazanowska KH, Lim SB (2023) An integrated single-cell transcriptomic dataset for non-small cell lung cancer. *Sci Data*. <https://doi.org/10.1038/s41597-023-02074-6>
12. Zeng J, Zhang Y, Shang Y, Mai J, Shi S, Lu M, Bu C, Zhang Z, Zhang Z, Li Y, Du Z, Xiao J (2021) CancerSCEM: a database of single-cell expression map across various human cancers. *Nucleic Acids Res* 50(D1):1147–1155. <https://doi.org/10.1093/nar/gkab905>
13. Luecken M, Burkhardt D, Cannoodt R, Lance C, Agrawal A, Aliee H, Chen A, Deconinck L, Detweiler A, Granados A, Huynh S, Isacco L, Kim Y, Klein D, DE KUMAR B, Kuppasani S, Lickert H, McGeever A, Melgarejo J, Mekonen H, Morri M, Müller M, Neff N, Paul S, Rieck B, Schneider K, Steelman S, Sterr M, Treacy D, Tong A, Villani A-C, Wang G, Yan J, Zhang C, Pisco A, Krishnaswamy S, Theis F, Bloom JM (2021) A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In: Vanschoren, J., Yeung, S. (eds.) *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1. Curran. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/158f3069a435b314a80bdc024f8e422-Paper-round2.pdf
14. Wu KE, Yost KE, Chang HY, Zou J (2021) BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceed Natl Acad Sci*. <https://doi.org/10.1073/pnas.2023070118>
15. Ashuach T, Gabitto MI, Jordan MI, Yosef N (2021) MultiVI: deep generative model for the integration of multi-modal data. *bioRxiv* <https://doi.org/10.1101/2021.08.20.457057>
16. Zhang R, Meng-Papaxanthos L, Vert J-P, Noble WS (2021) Semi-supervised single-cell cross-modality translation using Polarbear. *bioRxiv* <https://doi.org/10.1101/2021.11.18.467517>
17. Minoura K, Abe K, Nam H, Nishikawa H, Shimamura T (2021) A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Rep Methods* 1(5):100071. <https://doi.org/10.1016/j.crmeth.2021.100071>
18. Kingma DP, Welling M (2013) Auto-Encoding Variational Bayes. *arXiv*. <https://doi.org/10.48550/ARXIV.1312.6114>
19. Wen H, Ding J, Jin W, Wang Y, Xie Y, Tang J (2022) Graph neural networks for multimodal single-cell data integration. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '22*, pp. 4153–4163. Association for Computing Machinery, New York, NY, USA
20. Donisakeller H (1987) A genetic linkage map of the human genome. *Cell* 51(2):319–337. [https://doi.org/10.1016/0092-8674\(87\)90158-9](https://doi.org/10.1016/0092-8674(87)90158-9)
21. Barton NH (2010) Genetic linkage and natural selection. *Philos Trans R Soc B Biol Sci* 365(1552):2559–2569. <https://doi.org/10.1098/rstb.2010.0106>
22. Ott J, Wang J, Leal SM (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* 16(5):275–284. <https://doi.org/10.1038/nrg3908>
23. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks
24. Maglott D, Ostell J, Pruitt KD, Tatusova T (2010) Entrez gene: gene-centered information at NCBI. *Nucl Acids Res* 39:52–57. <https://doi.org/10.1093/nar/gkq1237>
25. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau D, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick B, Pruitt K, Sherry S (2021) Database resources of the national center for biotechnology information. *Nucl Acids Res* 50(D1):20–26. <https://doi.org/10.1093/nar/gkab1112>
26. Wolf FA, Angerer P, Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. <https://doi.org/10.1186/s13059-017-1382-0>
27. Ben-Ari Fuchs S, Lieder I, Stelzer G, Mazor Y, Buzhor E, Kaplan S, Bogoch Y, Plaschkes I, Shitrit A, Rappaport N, Kohn A, Edgar R, Shenav L, Safran M, Lancet D, Guan-Golan Y, Warshawsky D, Shtrichman R (2016) GeneAnalytics: an integrative gene set analysis tool for next generation sequencing, RNAseq and microarray data. *OMICS J Integr Biol* 20(3):139–151
28. Yousefi Z, Sharifzadeh S, Yar-Ahmadi V, Andalib A, Eskandari N (2019) Fc receptor-like 1 as a promising target for immunotherapeutic interventions of B-cell-related disorders. *Biomark Insights* 14:117727191988235. <https://doi.org/10.1177/1177271919882351>
29. Mamidi MK, Huang J, Honjo K, Li R, Tabengwa EM, Neeli I, Randall NL, Ponnuchetty MV, Radic M, Leu C-M, Davis RS (2023) FCRL1 immunoregulation in b cell development and malignancy. *Front Immunol*. <https://doi.org/10.3389/fimmu.2023.1251127>
30. David K, Friedlander G, Pellegrino B, Radomir L, Lewinsky H, Leng L, Bucala R, Becker-Herman S, Shachar I (2022) CD74 as a regulator of transcription in normal B cells. *Cell Rep* 41(5):111572. <https://doi.org/10.1016/j.celrep.2022.111572>
31. Müller PM, Rademacher J, Bagshaw RD, Wortmann C, Barth C, Unen J, Alp KM, Giudice G, Eccles RL, Heinrich LE, Pascual-Vargas P, Sanchez-Castro M, Brandenburg L, Mbamalu G, Tucholska M, Spatt L, Czajkowski MT, Welke R-W, Zhang S, Nguyen V, Rrustemi T, Trnka P, Freitag K, Larsen B, Popp O, Mertins P, Gingras A-C, Roth FP, Colwill K, Bakal C, Pertz O, Pawson T, Petsalaki E, Rocks O (2020) Systems analysis of RhoGEF and RhoGAP regulatory proteins reveals spatially organized RAC1 signalling from integrin adhesions. *Nat Cell Biol* 22(4):498–511. <https://doi.org/10.1038/s41556-020-0488-x>
32. Mosaddeghzadeh N, Ahmadian MR (2021) The Rho family GTPases: mechanisms of regulation and signaling. *Cells* 10(7):1831. <https://doi.org/10.3390/cells10071831>
33. Heasman SJ, Ridley AJ (2008) Mammalian Rho GTPases: new insights into their functions from in vivo studies. *Nat Rev Mol Cell Biol* 9(9):690–701. <https://doi.org/10.1038/nrm2476>
34. Tybulewicz VLJ, Henderson RB (2009) Rho family GTPases and their regulators in lymphocytes. *Nat Rev Immunol* 9(9):630–644. <https://doi.org/10.1038/nri2606>
35. Kim K, Bang S-Y, Yoo DH, Cho S-K, Choi C-B, Sung Y-K, Kim T-H, Jun J-B, Kang YM, Suh C-H, Shim S-C, Lee S-S, Lee J, Chung WT, Kim S-K, Choe J-Y, Nath SK, Lee H-S, Bae S-C (2016) Imputing variants in HLA-DR beta genes reveals that HLA-DRB1 is solely associated with rheumatoid arthritis and systemic lupus erythematosus. *PLOS ONE* 11(2):0150283. <https://doi.org/10.1371/journal.pone.0150283>
36. Ahmed I, Tamouza R, Delord M, Krishnamoorthy R, Tzourio C, Mulot C, Nacfer M, Lambert J, Beaune P, Laurent-Puig P, Lorient M, Charron D, Elbaz A (2012) Association between Parkinson's disease and the HLA-DRB1 locus. *Mov Disord* 27(9):1104–1110. <https://doi.org/10.1002/mds.25035>

37. Inagaki Y, Hayakawa F, Hirano D, Kojima Y, Morishita T, Yasuda T, Naoe T, Kiyoi H (2016) PAX5 tyrosine phosphorylation by SYK co-operatively functions with its serine phosphorylation to cancel the PAX5-dependent repression of BLIMP1: a mechanism for antigen-triggered plasma cell differentiation. *Biochem Biophys Res Commun* 475(2):176–181. <https://doi.org/10.1016/j.bbrc.2016.05.067>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.