

Fig. S1

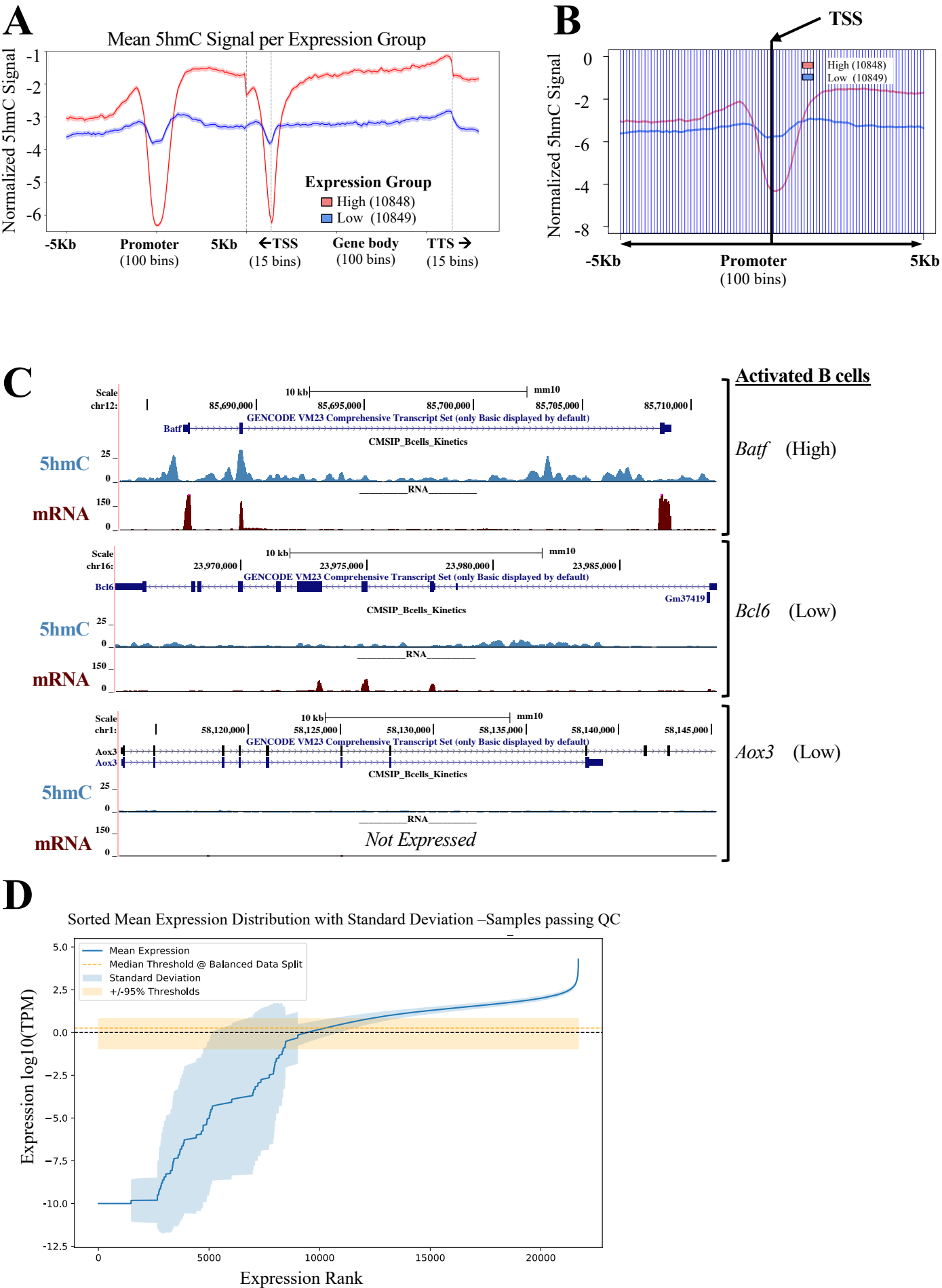
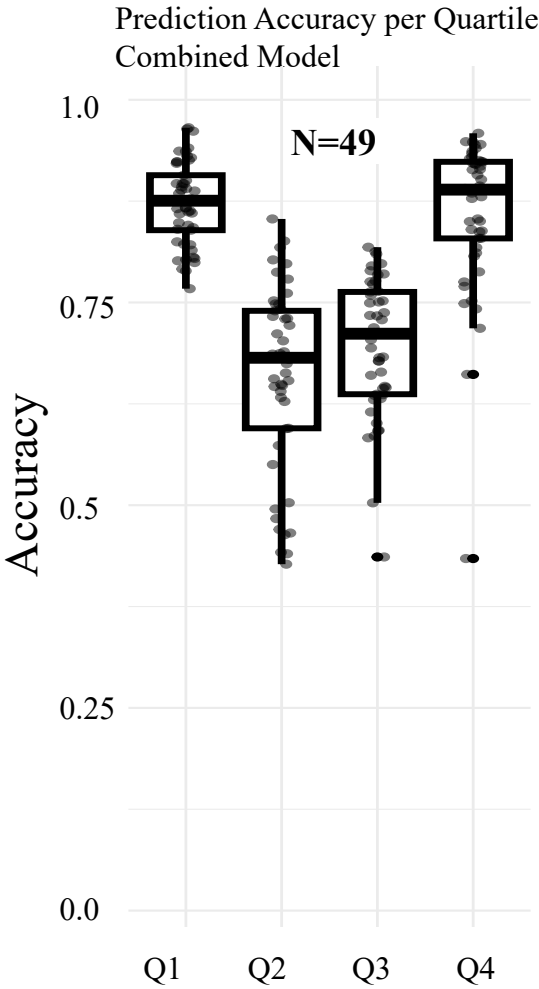


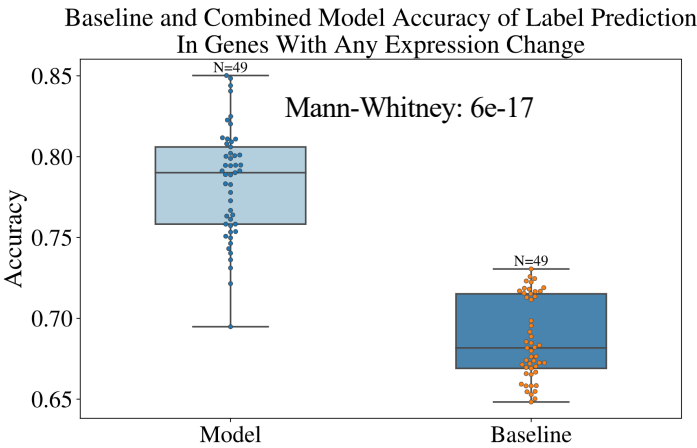
Fig. S1: Generation of input features and labels. (A) Composite plot of average 5hmC normalized signal per bin in each expression group across an aggregate gene body and promoter-associated bins. Shadows around the mean is the 95% Confidence Interval (CI). The CI was calculated by multiplying the standard error by the t-value corresponding to 97.5% of the t-distribution, with degrees of freedom of 10847 for Class 1 and 10848 for Class 2. (B) Expanded view of the TSS. Vertical lines correspond to division into bins with blue lines representing a fixed size bin of 100 bp. (C) Illustration of 5hmC enrichment signal (blue tracks) across three levels of mRNA expression (dark-red tracks) in Activated B cells. *Batf* is highly expressed (Top) and has a strong 5hmC signal distribution; in contrast, *Aox3* is not expressed and has no 5hmC signal (Bottom). *Bcl6* is expressed at low levels and has low 5hmC enrichment (Middle). (D) Variance of gene expression for each ranking (highest to lowest TPM on x-axis) across the 49 samples. Y axis represent $\log_{10}(\text{TPM})$ values and mean expression per ranking is plotted by a solid blue line and with standard deviation shown as lighter blue shade. Dotted horizontal black line corresponds to 1 TPM (or $\log_{10}(\text{TPM})=0$). Dotted red horizontal line and horizontal shade surrounding it correspond to median TPM thresholds used across different samples.

Fig. S2

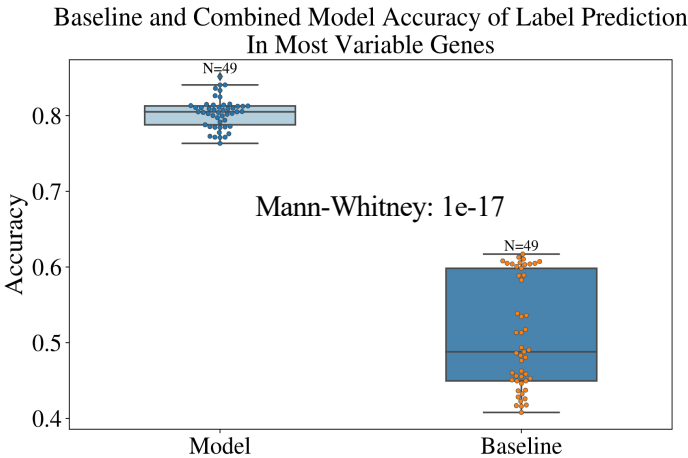
A



B



C



D

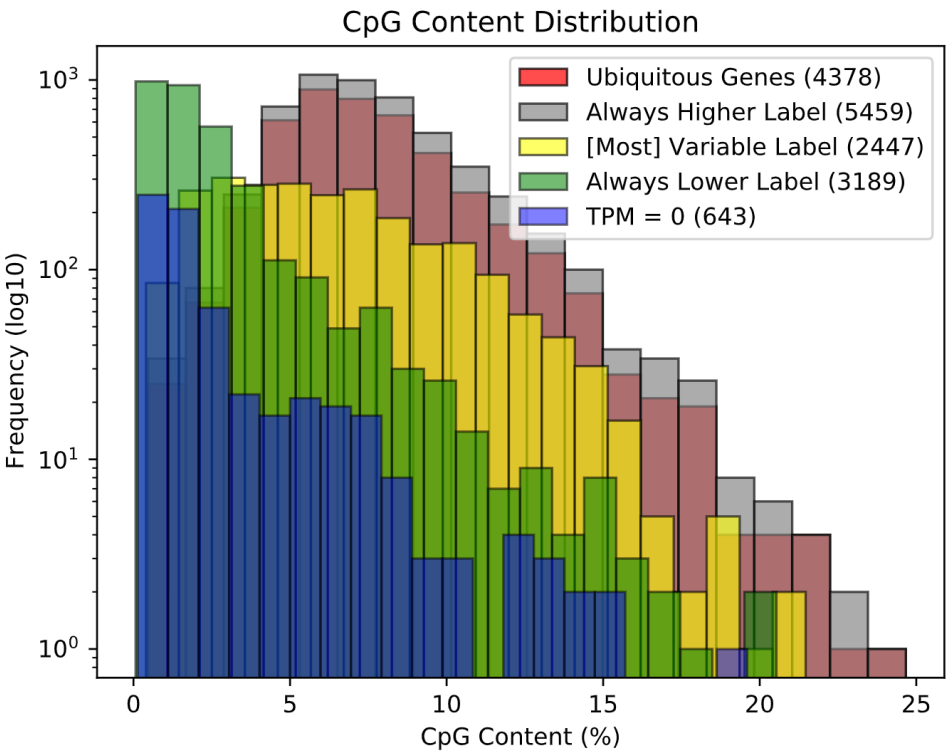
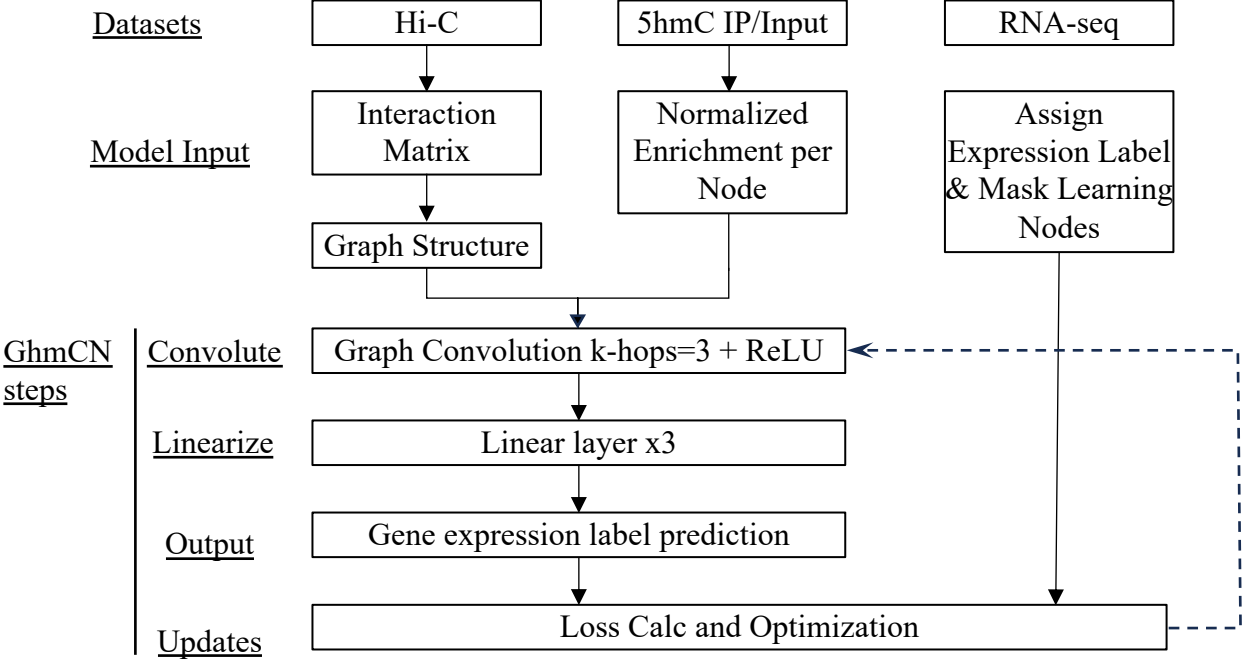


Fig. S2. Assessment of our predictive models and potential confounding factors. (A) Prediction accuracy of our combined model across different quartiles of gene expression. Quartiles (Q) were defined using the TPM-normalized gene expression. Q1 represents the bottom 25% of genes based on expression (per sample), whereas Q4 represents the 25% topmost expressed genes. Q1 and Q2 are genes labeled with “Low” expression state, whereas Q3 and Q4 are genes labeled with “High” expression state. **(B-C)** Prediction accuracies of our model versus the majority vote baseline for genes with any variability in their expression state across the 49 samples **(B)** and for genes that are most variable with at least one third of samples having a label opposite to the majority label **(C)**. **(D)** CpG content distributions for five different groups of genes classified with respect to their expression levels and states. We calculate CpG content as the number of CpG dinucleotides divided by half the length of the overall sequence.

Fig. S3

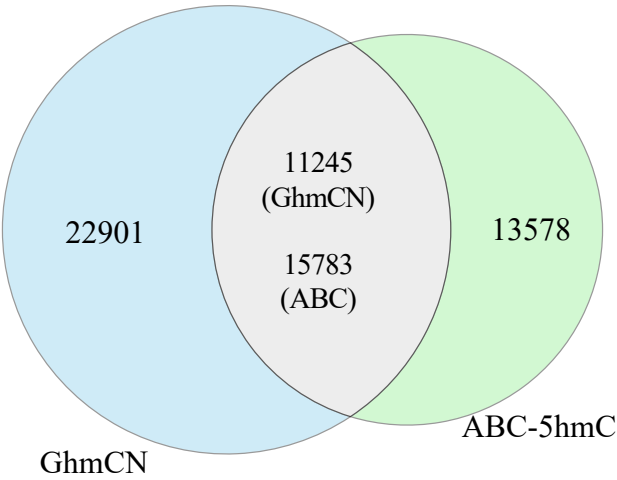
A

Flow diagram of GhmCN



B

Overlap of GNNExplainer top enhancers and ABC 5hmC (ATAC peaks) enhancers



C

Overlap of GNNExplainer top enhancers and ABC H3K27Ac (ATAC peaks) enhancers

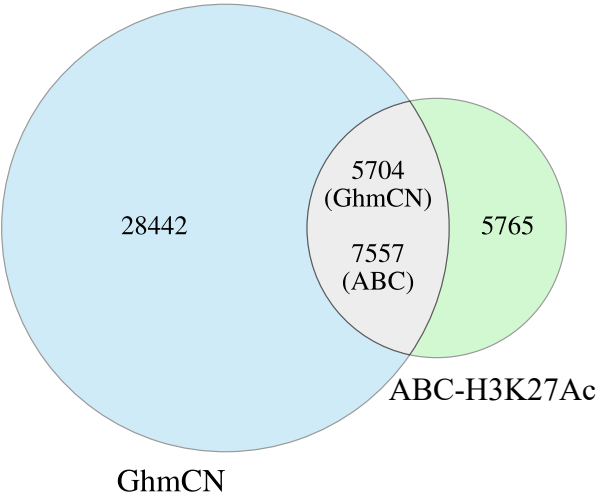
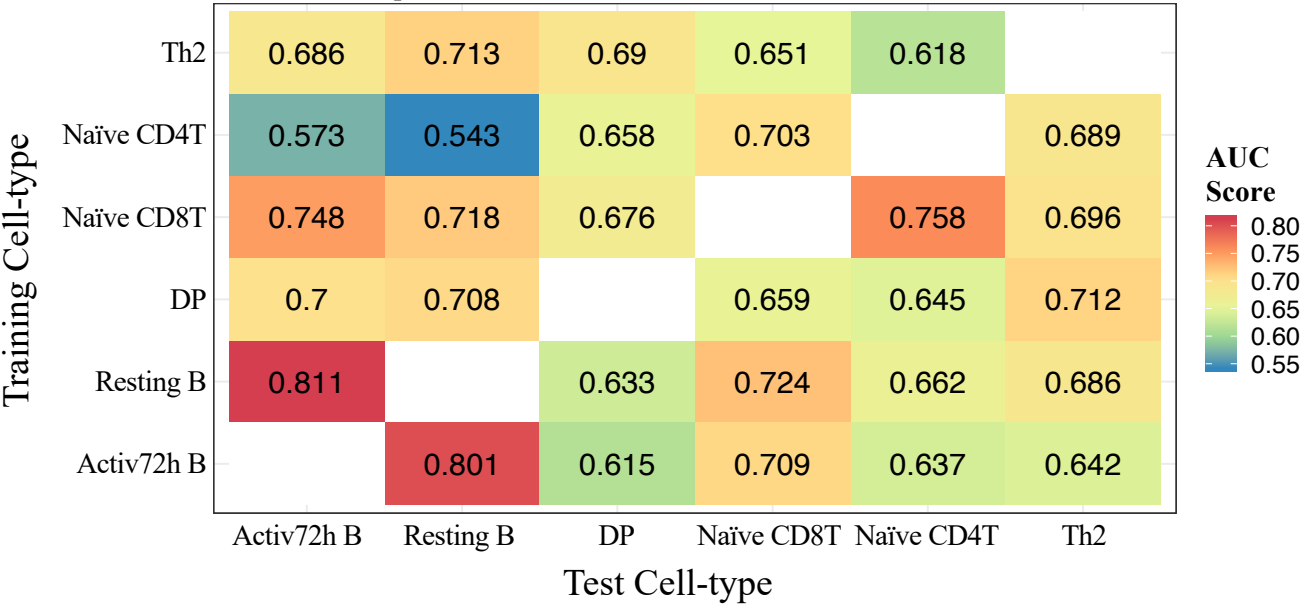


Fig. S3. Schematic representation of the GhmCN model. (A) Flow diagram of the GhmCN method and model architecture. (B) Overlap between regions predicted from GhmCN and from ABC-5hmC models for activated B cells. (C) Overlap between regions predicted from GhmCN and from ABC-H3K27ac models for activated B cells.

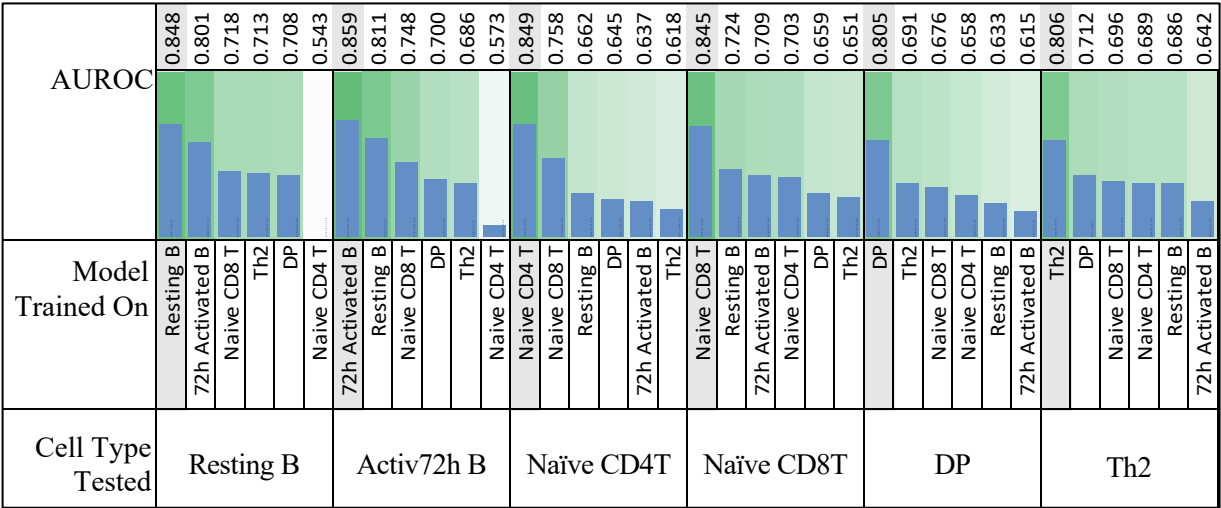
Fig. S4

A

AUC Scores of our models trained and tested in different cells
Cross-cell Comparisons



B



C

Gene expression (RNA-seq)

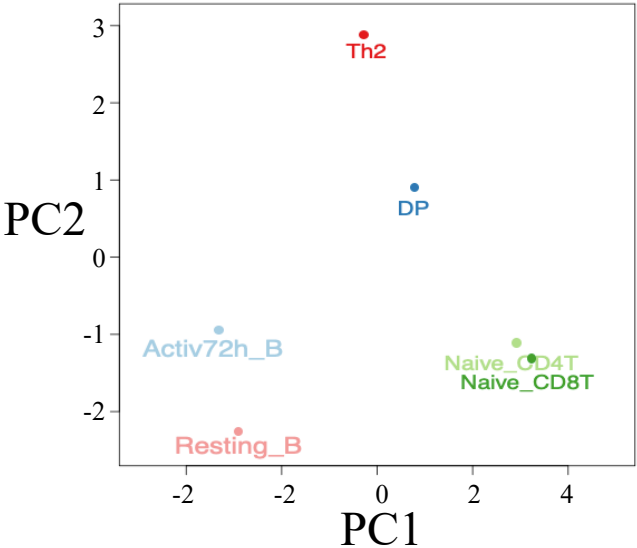


Fig. S4. Detailed analysis of the GhmCN model and cross-cell-type predictions. (A) AUC values from training GhmCN model on one sample and predicting gene expression in another. (B) Same AUC values as presented in (A) but organized differently. (C) PCA plot of RNA-seq data used in this study. The top 1000 most variable genes among all six samples were used for this analysis.