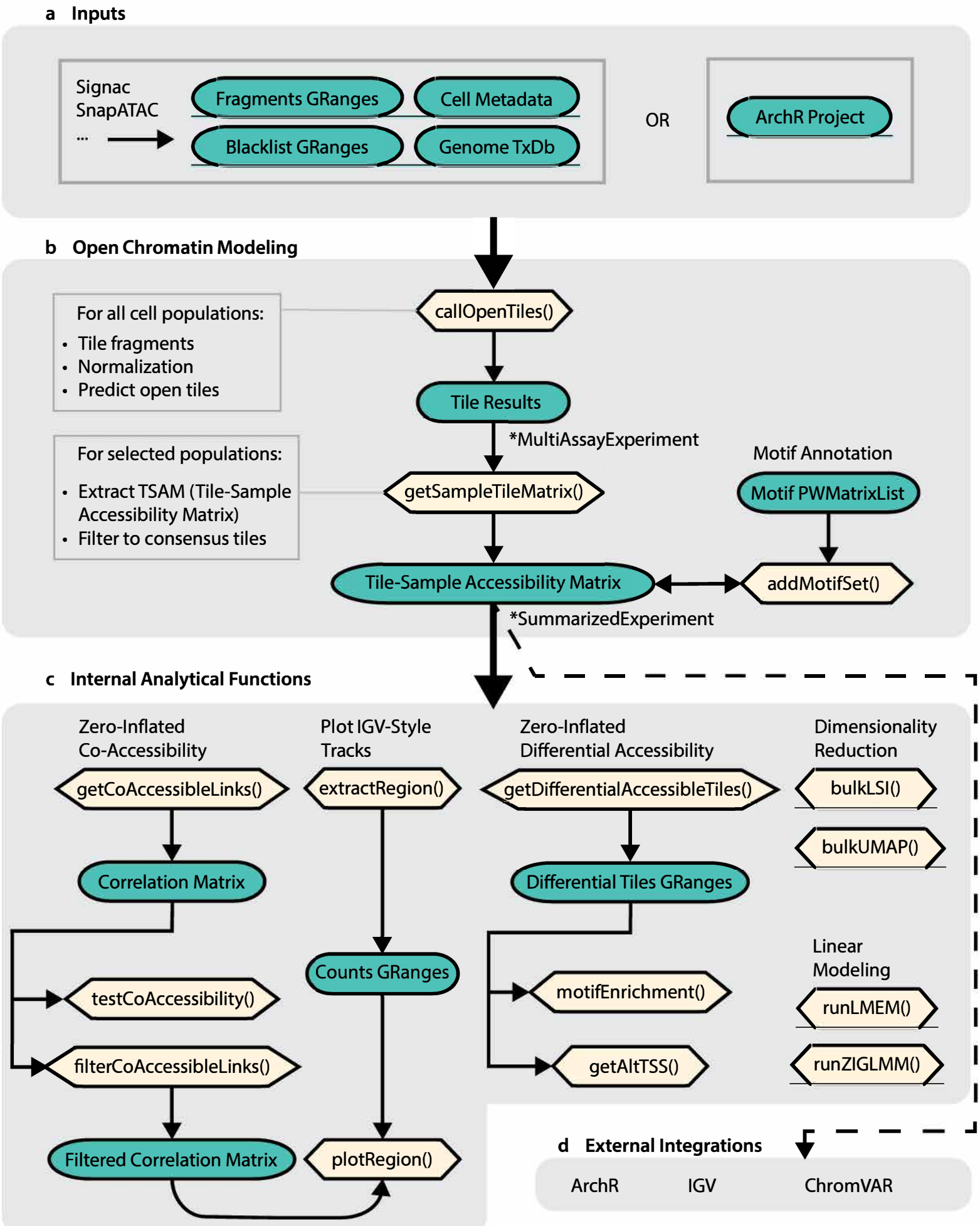
Supplementary Information

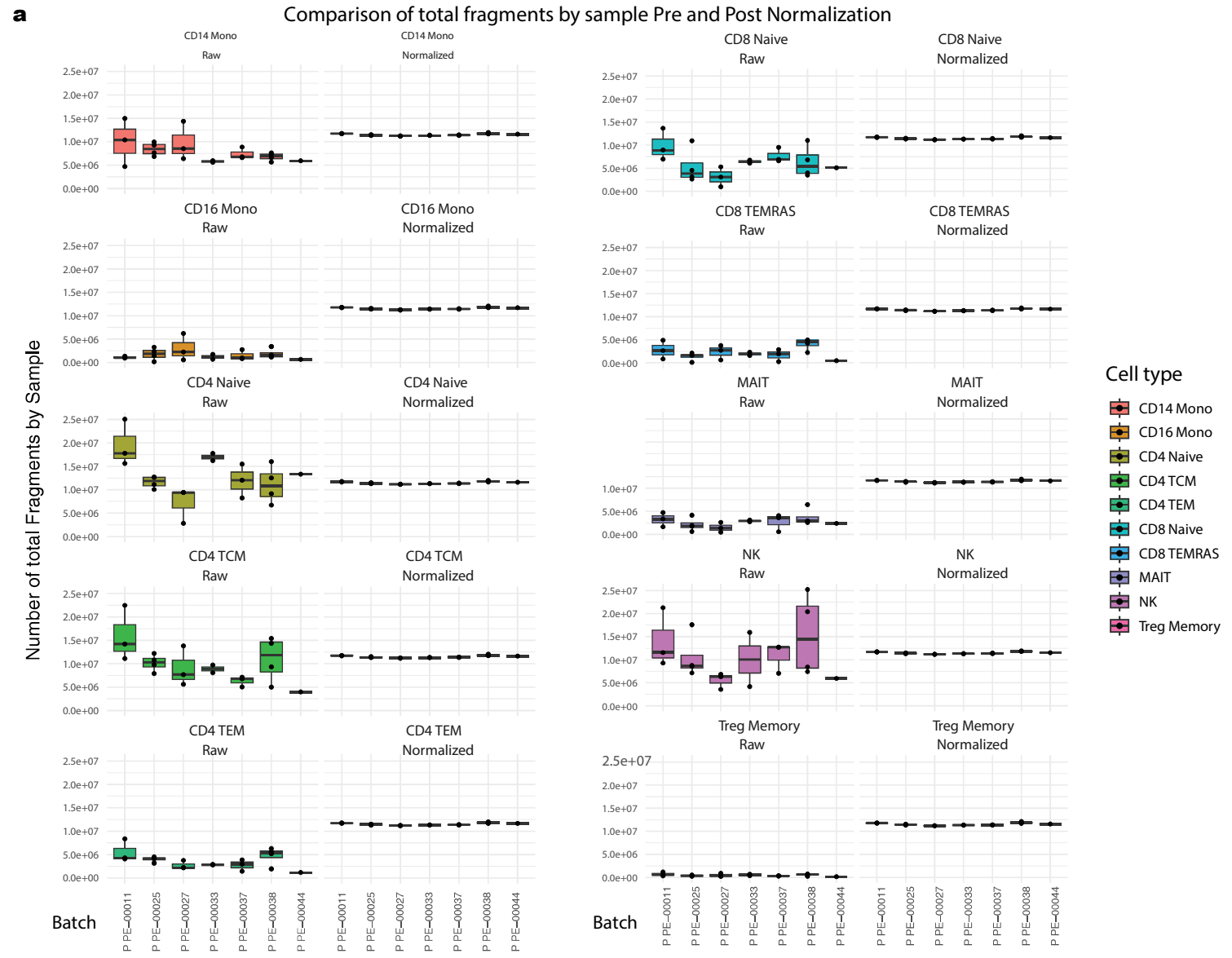
MOCHA's advanced statistical modeling of scATAC-seq data enables functional genomic inference in large human cohorts

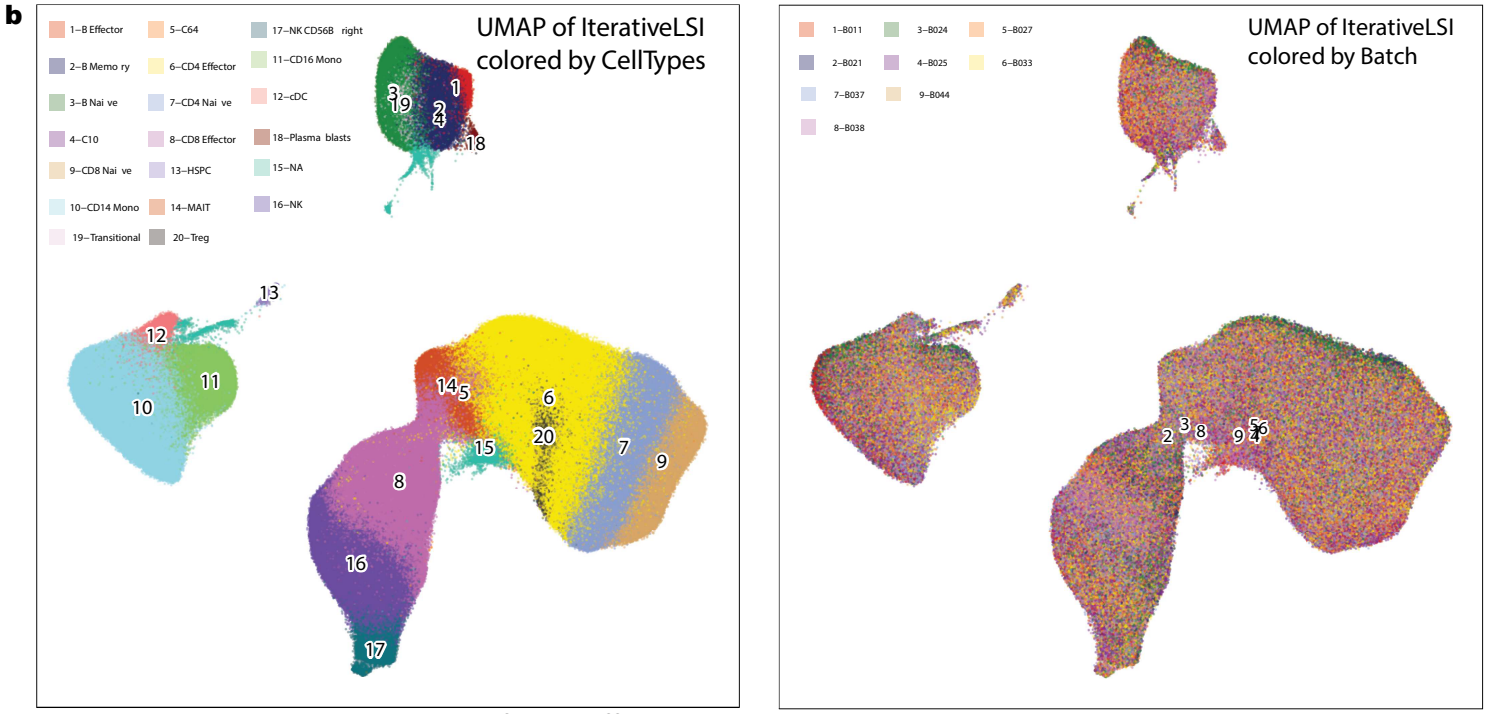
Supplementary Figure S1-S18
Supplementary Table 1-2



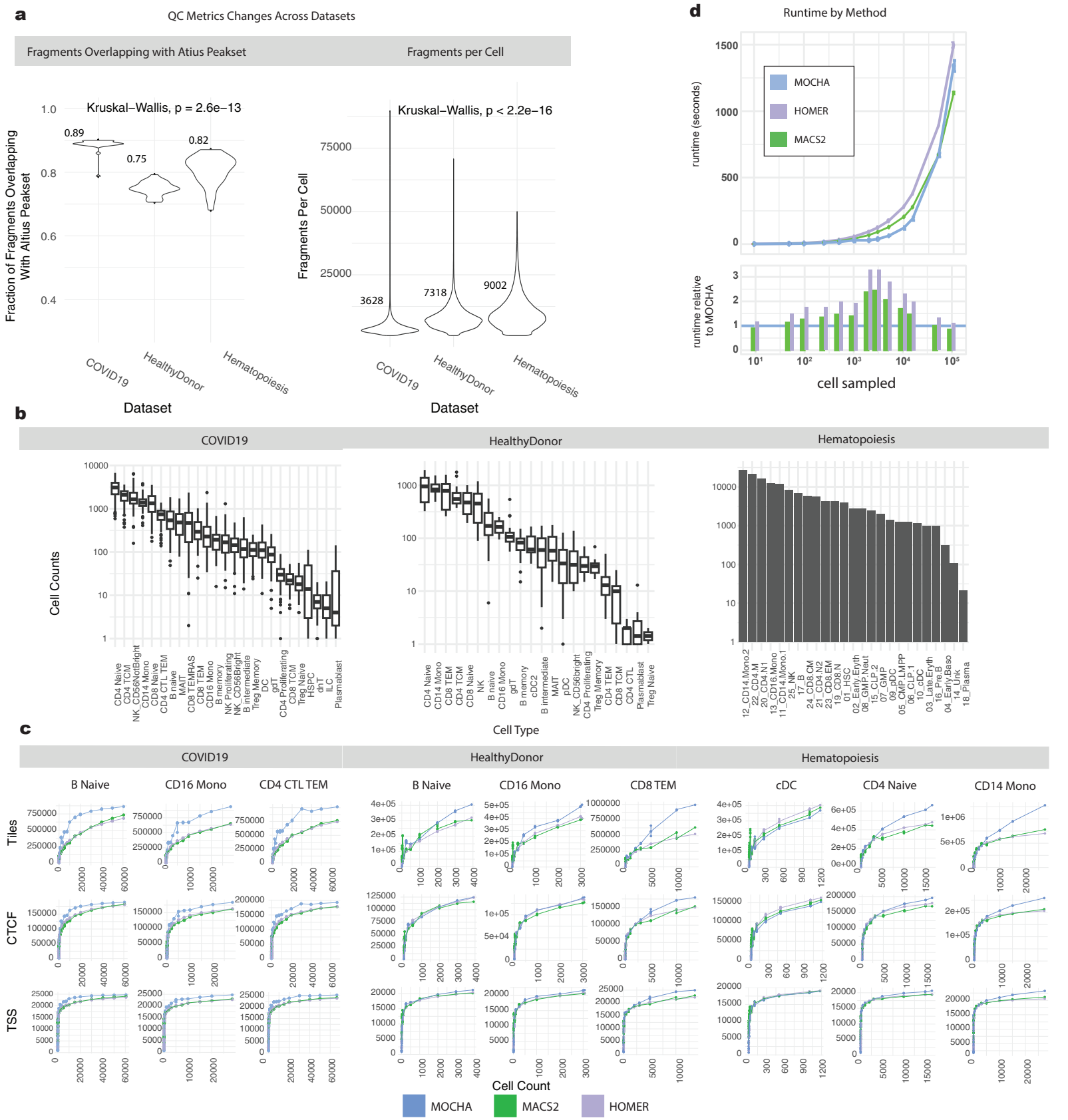
Supplementary Fig. 1. MOCHA's technical workflow schematic.
Schematic of the MOCHA R package workflow functions (in yellow) and objects (green). **a**, MOCHA takes inputs from an ArchR project or collections of input files from ATAC-seq analysis software. **b**, Core functions of MOCHA and result objects. **c**, Downstream analyses supported by MOCHA with functions. **d**, MOCHA enables additional downstream analyses available in external software. Figures were generated using Adobe Illustrator.

Supplementary Fig. 2. Technical details for developing MOCHA's analytical modules.
a, Large difference in sequencing depth per sample observed in the full COVID19 dataset (n=91). The distribution of the number of fragments per sample in the CD16 monocytes are shown separately for COVID+ (n=69) and COVID- samples (n=22). **b-c**, Distributions of coefficients of variation (CVs) of the pseudo-bulk fragment counts at 2,230 cell-type invariant CCCTC-binding factor (CTCF) sites in the COVID19 dataset (n=91) before (Raw) or after normalization by the total number of fragments per cell type per sample (MOCHA), the total cellular abundance per sample (By Total Cell #), or the total number of fragments per sample. Pseudo-bulk fragment counts were calculated for individual tiles per cell type and sample. For each CTCF site, CV was calculated (**b**) within each cell type across samples (n=91) or (**c**) across cell types (n=25) within each sample. **d-i**, Based on data in the COVID19X dataset (n=39). **d**, Number of tiles in CD16 monocytes that were commonly open to at least a targeted fraction of samples. Data of COVID+ samples (n=17) and COVID- samples (n=22) were analyzed separately. The smooth curves and the shaded bands are the Loess fitting curves and the corresponding 95% confidence intervals. The vertical dashed line indicates the fraction threshold (20%) used. **e**, Heatmap of pseudo-bulk accessibility in the tile-sample accessibility matrix (TSAM) of CD16 monocytes with tiles in Chromosome 1 only. Tiles were sorted by their percentage of zeros across samples. **f**, Histogram of percentage of zeros across samples as a function of tile log2(accessibility) value. The bar represents the mean value while the error bar represents the corresponding standard deviation. **g**, The distribution of accessibilities of all tiles revealing a bimodal distribution. Accessibility threshold was set near the higher mode. **h-i**, Exemplar histograms showing differences in accessibility between COVID+ and COVID- samples that arose from either difference in non-zero accessibilities without significant difference in the proportion of zeros (**h**), or difference in the proportion of zeros without significant difference in non-zero accessibilities (**i**). Source data provided in **Source Data Supplementary Fig. 2**. Generated using Adobe Illustrator.

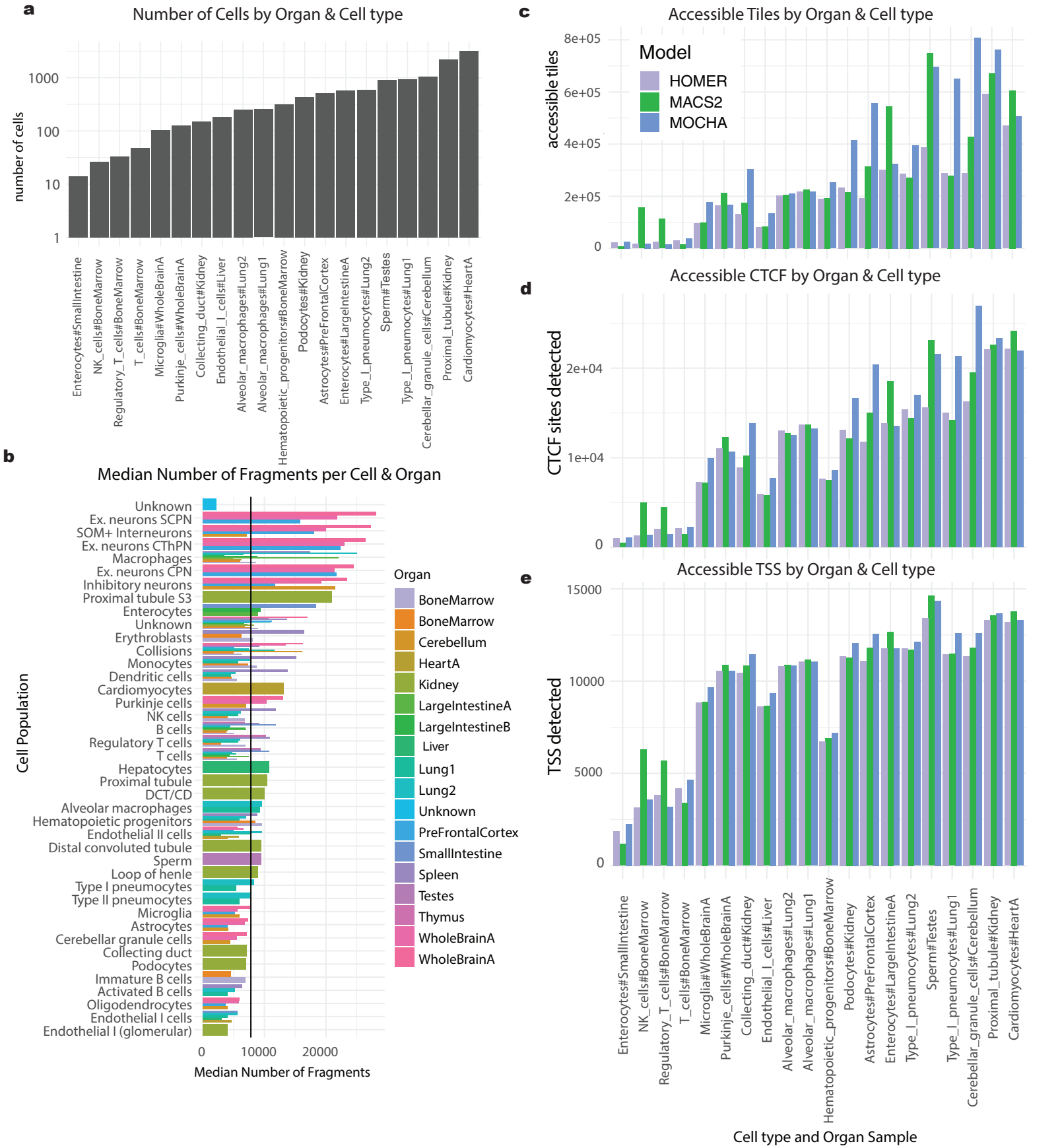




Supplementary Fig. 3. Assessing impact of batch effects on the COVID Longitudinal Dataset.
a, Boxplots of pseudo-bulk total fragment counts for 8 representative cell types across batches before (left panel) and after (right panel) MOCHA normalization. Each point shows the total number of fragments per sample for a given cell type. Each boxplot corresponds to an individual batch and displays the median (centerline), the first and third quartiles (the lower and upper bound of the box), and the 1.5x interquartile range (whiskers) of the data. Data included 10 randomly selected COVID+ samples and 10 randomly selected COVID samples from the COVID19X dataset (n=39). **b**, Uniform Manifold Approximation and Projection (UMAP) plots of single cell data from the full COVID19 dataset (n=91), colored by cell type identity (left panel) or batch (right panel). Minimal batch effects are visible. Source data are provided in **Source Data Supplementary Figure 3_1** (normalization data), and **Source Data Supplementary Figure 3_2** (UMAP coordinates with cell type and batch IDs). Figures were generated using Adobe Illustrator.



Supplementary Fig. 4. Dataset characteristics and benchmarking on open chromatin identifcation during downsampling.
a, Quality control (QC) metrics across three datasets, as measured by the percentage of fragments that overlap with the Altius peakset (left) and the number of fragments per individual cell (right). The corresponding median values are indicated for each dataset. The three datasets were significantly different on these two QC metrics (P = 2.6x10-13 and P < 2.2x10-16, respectively; Kruskal-Wallis test). **b**, The number of cells per cell type across the three datasets. Each boxplot displays the median (centerline), the first and third quartiles (the lower and upper bound of the box), and the 1.5x interquartile range (whiskers) of the data. **c**, Head-to-head comparison between MOCHA, MACS2, and HOMER on numbers of detected tiles (top), tiles overlapping with CTCF sites (middle), and overlapping with TSSs (bottom) as functions of sampled cell count in three representative cell types from each of the three datasets. The three datasets are COVID19 (n=91, left), HealthyDonor (n=18, middle), and Hematopoiesis (treated as n=1 sample, right). CTCF: CCCTC-binding factor; TSS: transcription starting site. **d**, Line plots indicating the run time (in seconds) required to identify open chromatin from scATAC-seq data as a function of the number of analyzed samples (top panel). Barplots showing the relative runtime of MACS2 (green) and HOMER (light purple) to MOCHA (bottom panel). The blue horizontal line at 1 in the bottom panel marks the MOCHA reference runtime. Source data are provided in **Source Data Supplementary Fig. 4-1** and **4-2**. Figures were generated using Adobe Illustrator.



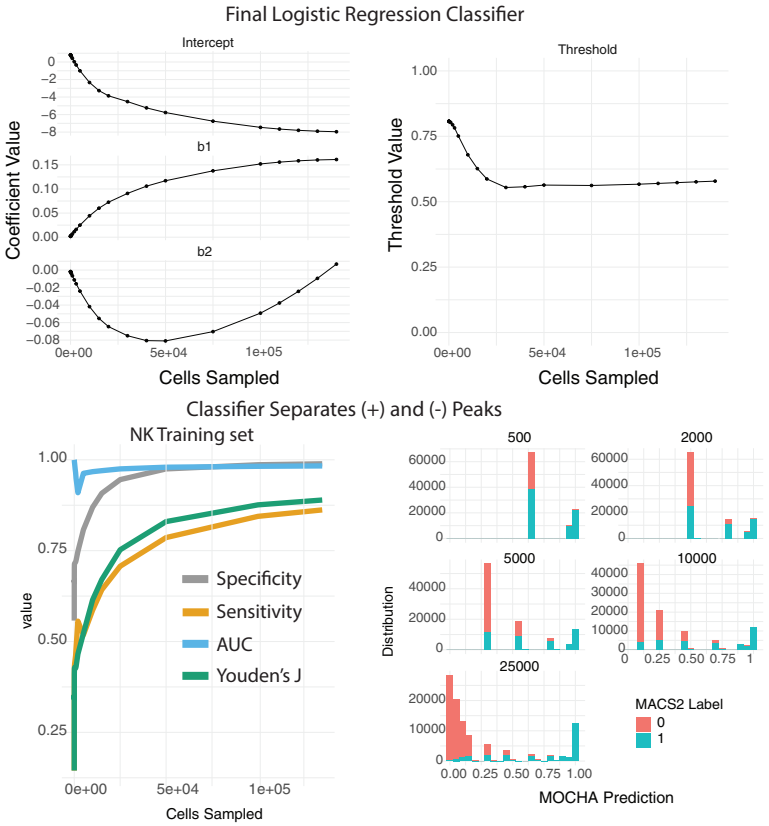
Supplementary Fig. 5. Benchmarking on mouse pan-organ data.
a, Bar plots illustrating the total number of cells per organ and cell type in the mouse panorgan scATAC-seq dataset. **b**, Bar plots show the median number of fragments per cell by organ sample and cell type. Bar plots are color-coded by the organ sample. **c-e**, The bar plots provide the number of the accessible tiles (**c**), CTCF (CCCTC-binding factor) sites (**d**), and TSS (transcription starting site) sites (**e**) detected by MOCHA, MACS2, and HOMER, in blue, green, and light purple, respectively. Source data are provided in **Source Data Supplementary Figure 5**. Figures were generated using Adobe Illustrator.

a. Model training

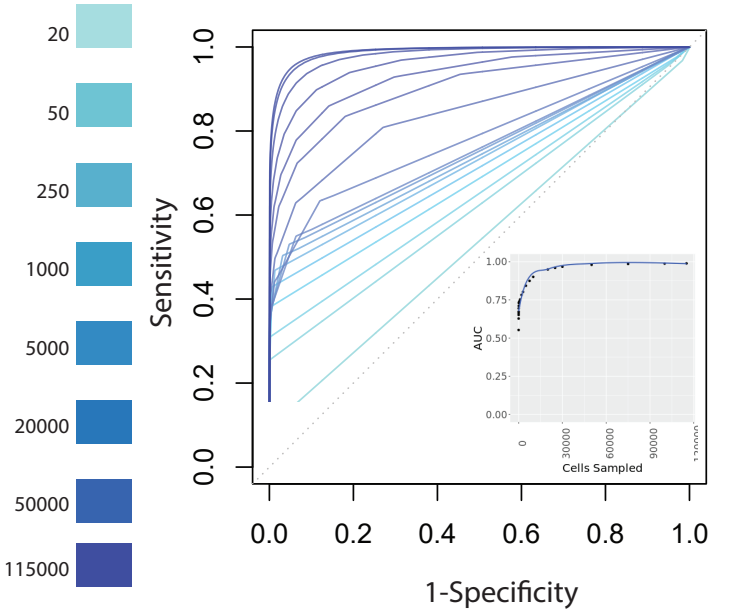
Algorithm 1 Training MOCHA's LRM

procedure MODEL TRAINING
  For $k \in \{5, 10, \cdots, 150000\}$
    Sample 10 groups of size k cells, $\mathbf{S_k} = \{S_{1,k} \cdots, S_{10,k}\}$
  procedure TRAINING($\mathbf{S_k}$)
    Train model on each subsample $S_{i,k} \in \mathbf{S_k}$
    procedure TRAINING($S_{i,k}$)
      Estimate, genome-wide normalized counts
      Calculate $\lambda_{i,k}^{(1)}, \lambda_{i,k}^{(2)}$ measures of intensity
      Train Logistic Regression Model using MACS2 label
      Extract Model Coefficients: $\beta_{i,k}$
      Estimate Optimal Threshold based on Youden Index: $T_{i,k}$
    return $\left(T_{i,k}, \beta_{i,k}\right)$
  return $\left(T_k, \beta_k\right) = \left(\text{Median}(T_{i,k}), \text{Median}(\beta_{i,k})\right)$ for $\mathbf{S_k}$
return Model Object$= \left[\left(T_5, \beta_5\right), \cdots, \left(T_{150000}, \beta_{150000}\right)\right]$
Final LRM: Smoothen and interpolate across $\left[\left(T_k, \beta_k\right)\right]_{k=5}^{150000}$ to generate final model

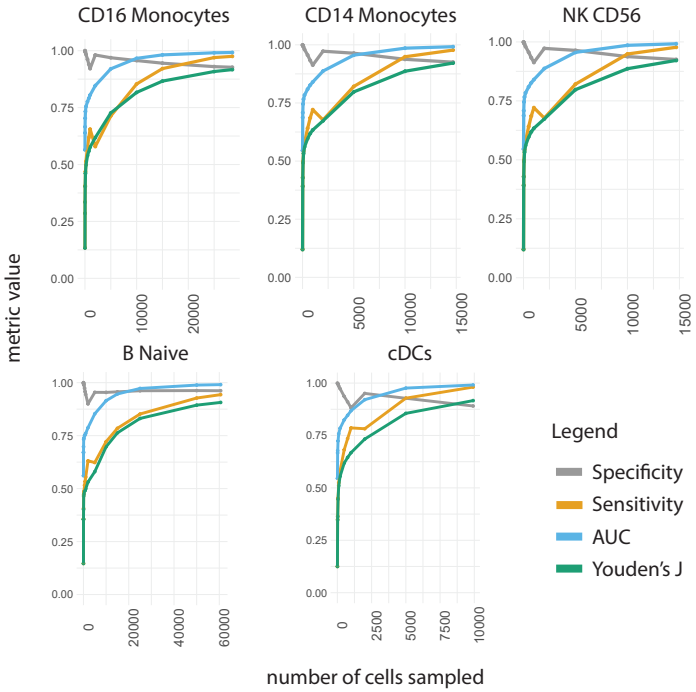
b. Training Results



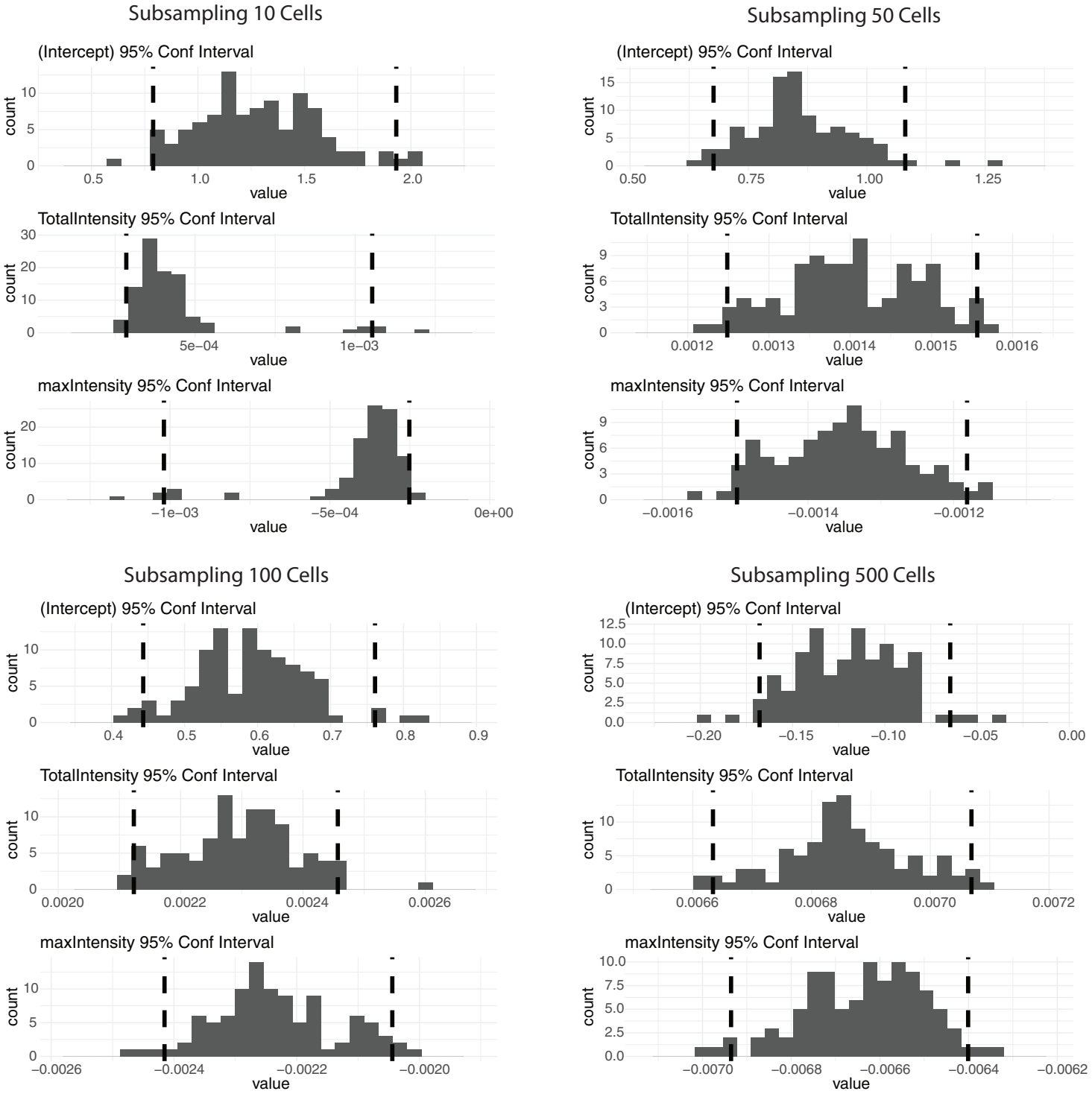
c. Validation: CD14 Monocytes



d. Validation: Various Cell Populations

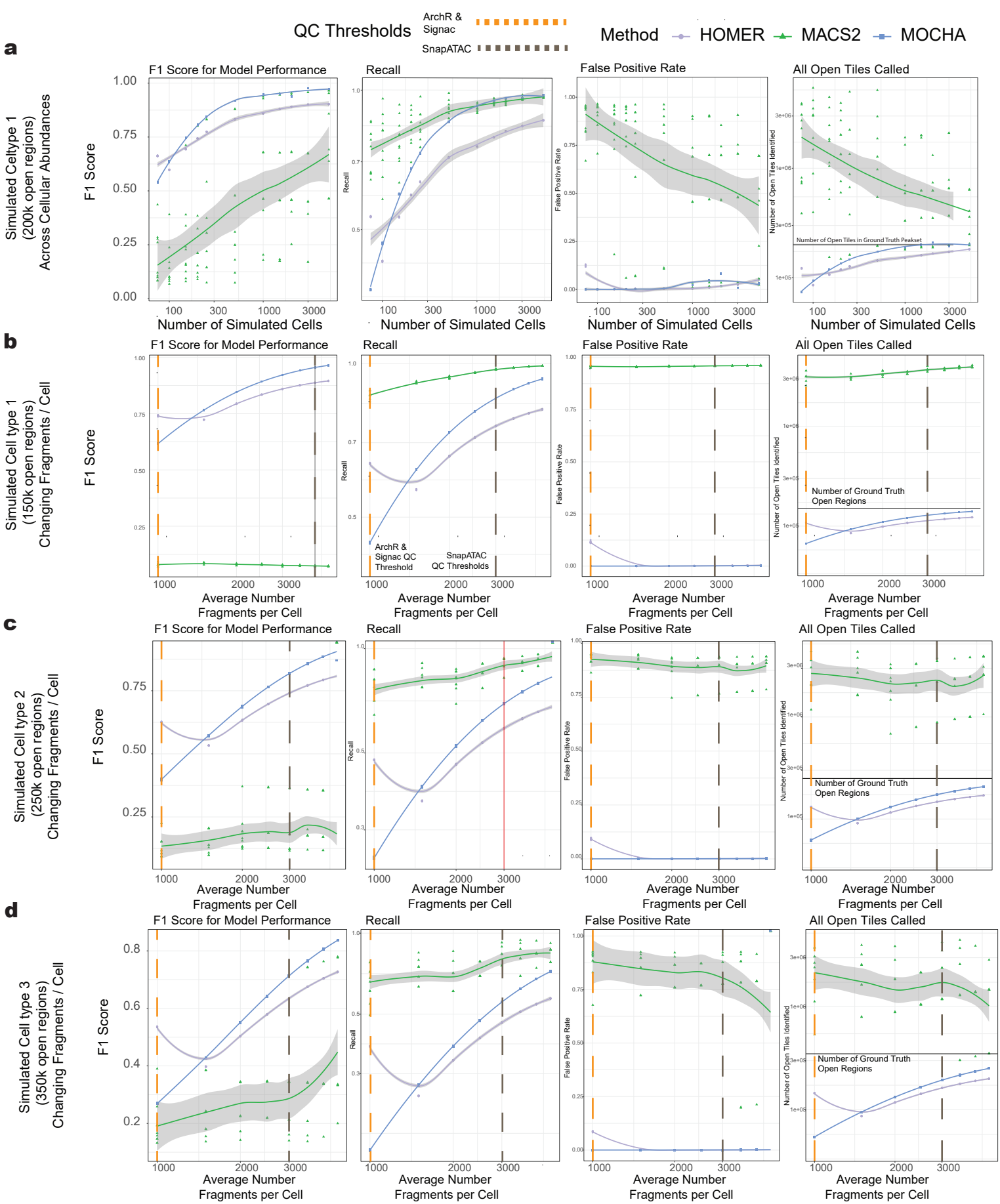


Supplementary Fig. 6. Training and validation of open accessibility models.
a, Pseudo code for training logistic regression models (LRMs) for open tiles. **b**, Training results on natural killer (NK) cells (n=179,836). Top: Coefficients of the LRMs and the corresponding threshold as a function of sampled cell count. Bottom left: Specificity, sensitivity, area under the receiver operating characteristic (ROC) curve (AUC), and Youden's J index as a function of sampled cell count. Bottom right: Histograms of the probability scores of open (blue) and closed (light purple) tiles at cell count 500, 2000, 5000, 10000, and 25000. **c**, ROC curves on the validation data of CD14 monocytes as cell count ranged from 20 to 115,000. Insert: The corresponding AUC as a function of cell count. The Loess fitting curve is plotted in blue. **d**, Validation performance on specificity, sensitivity, AUC, and Youden's J index as a function of sampled cell count for five representative cell types. cDCs: classical dendritic cells. Data in the COVID19 dataset (n=91 samples) was used for the training and validation of the LRMs. Source data are provided in **Source Data Supplementary Fig. 6**. Figures were generated using Adobe Illustrator.

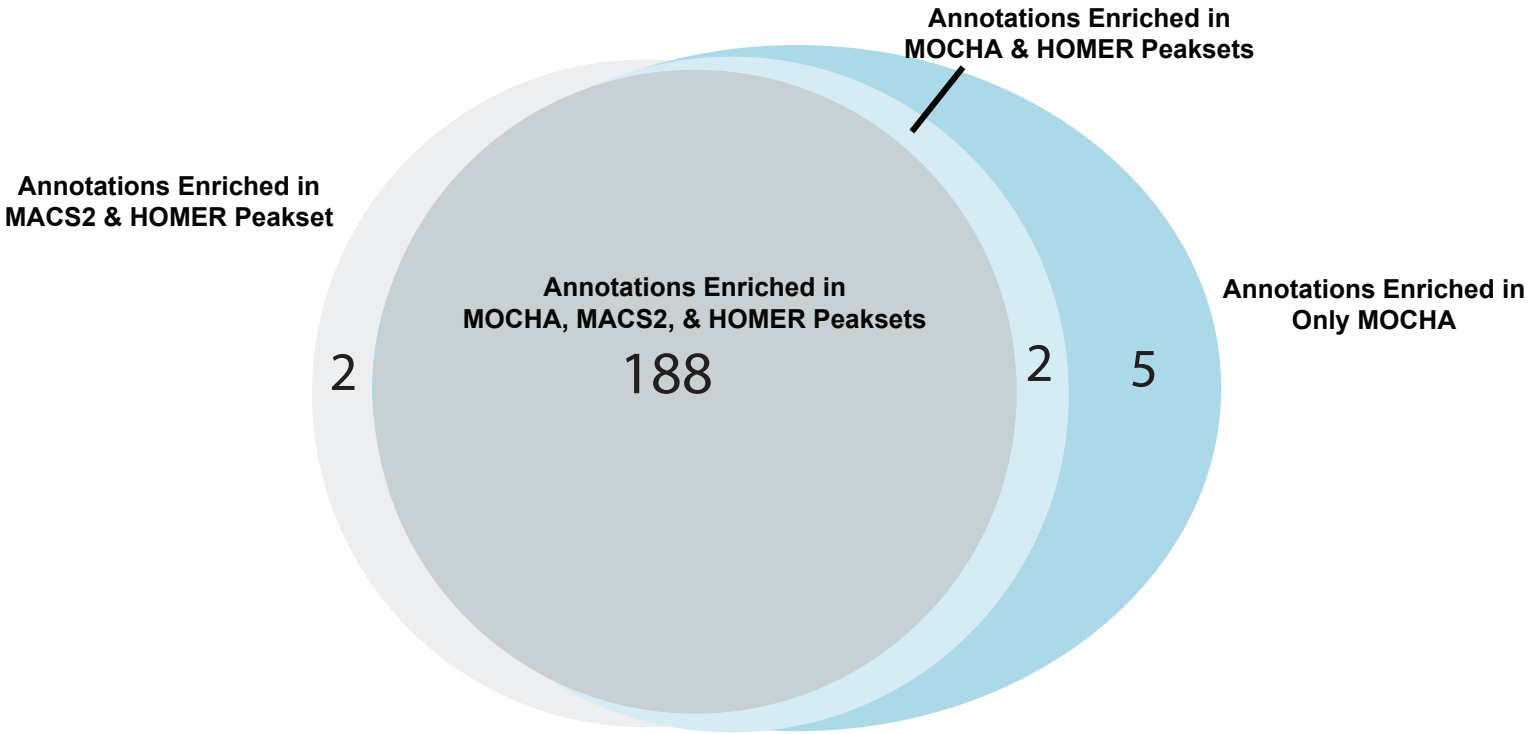


Supplementary Fig. 7. Assessing feature significance.

Histograms of the logistic regression model (LRM) coefficients for 'Intercept', $\lambda^{(1)}_{i,j,t}$ = total intensity, and $\lambda^{(2)}_{i,j,t}$ (the maximum intensity), evaluated from 100 replicates of model training on 10, 50, 100, and 500 individual cells, respectively, from the natural killer (NK) population in the full COVID19 dataset (n=91). The 95% confidence intervals (CI) for the coefficients are shown in dashed black lines to illustrate whether the coefficient is statistically significantly different from 0. Source data are provided in Source Data Supplementary Figure 7. Figures were generated using Adobe Illustrator.

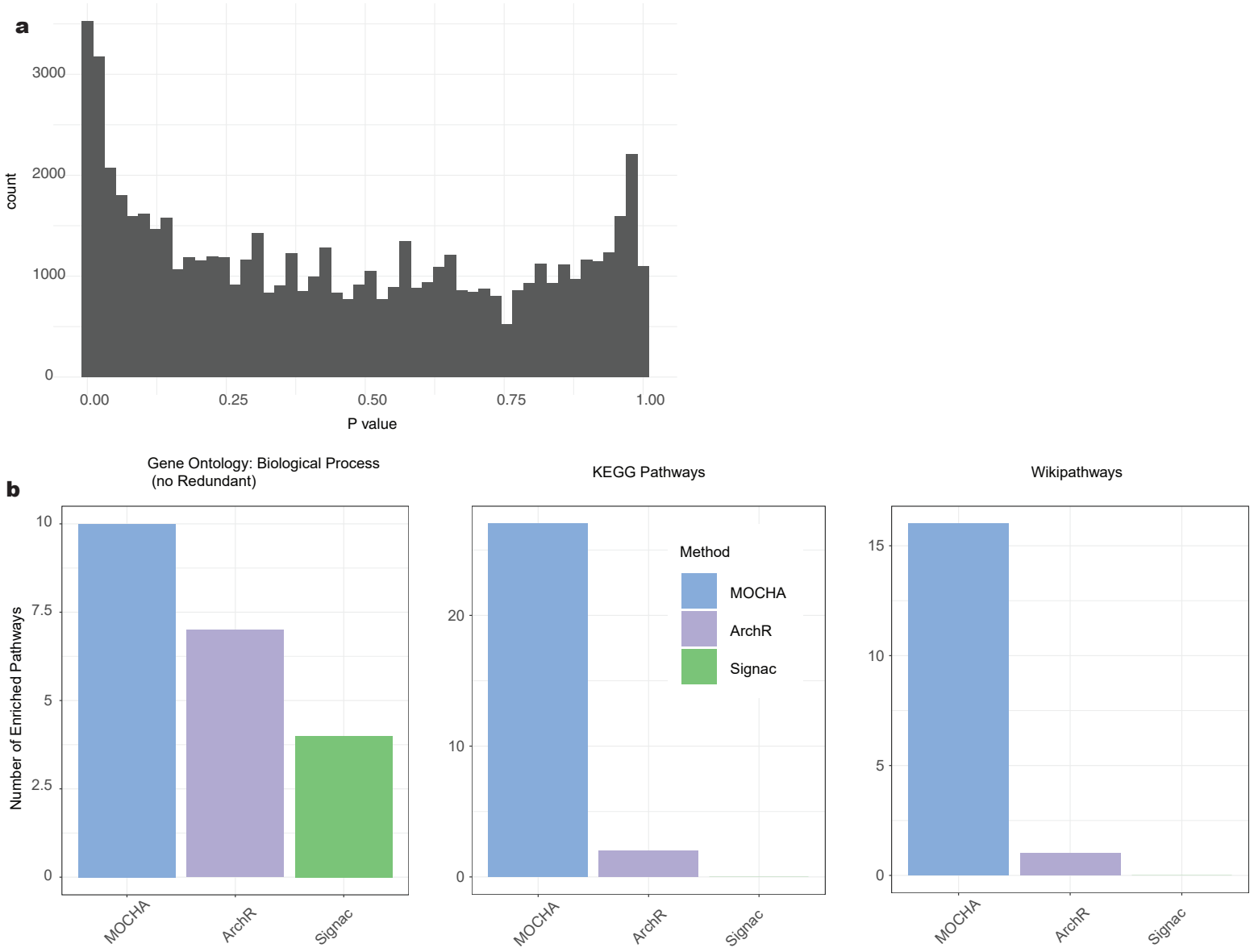


Supplementary Fig. 8. Performance on simulated scATAC-seq data.
F1 score, recall, false positive rate, and total number of open tiles obtained by MOCHA, MACS2, or HOMER on the simulated scATAC-seq data, across a range of simulated peaksets and cell counts. For all plots, MACS2 results are marked with triangles (green), HOMER with dots (red), and MOCHA with squares (blue). Trend line represents the loess curve through all points for a given method with shades denoting 95% confidence intervals. The total number of open tiles in the ground-truth simulated peakset is marked with a black horizontal line (right panel). **a**, Results from a simulated cell type with 10 peaksets (200,000 open tiles each) across a range of cell counts. **b-d**, Results from three simulated cell types with 150k (**b**), 250k (**c**), and 350k (**d**) open regions, respectively. For each simulated cell type, data was simulated with 250 cells across a wide range of sequencing depth per cell. Quality control (QC) thresholds for ArchR and Signac are marked with a dashed orange line on the left, and the QC threshold from SnapATAC's tutorial is marked with a dashed black line on the right. Source data are provided in **Source Data Supplementary Figure 8**. Figures were generated using Adobe Illustrator.

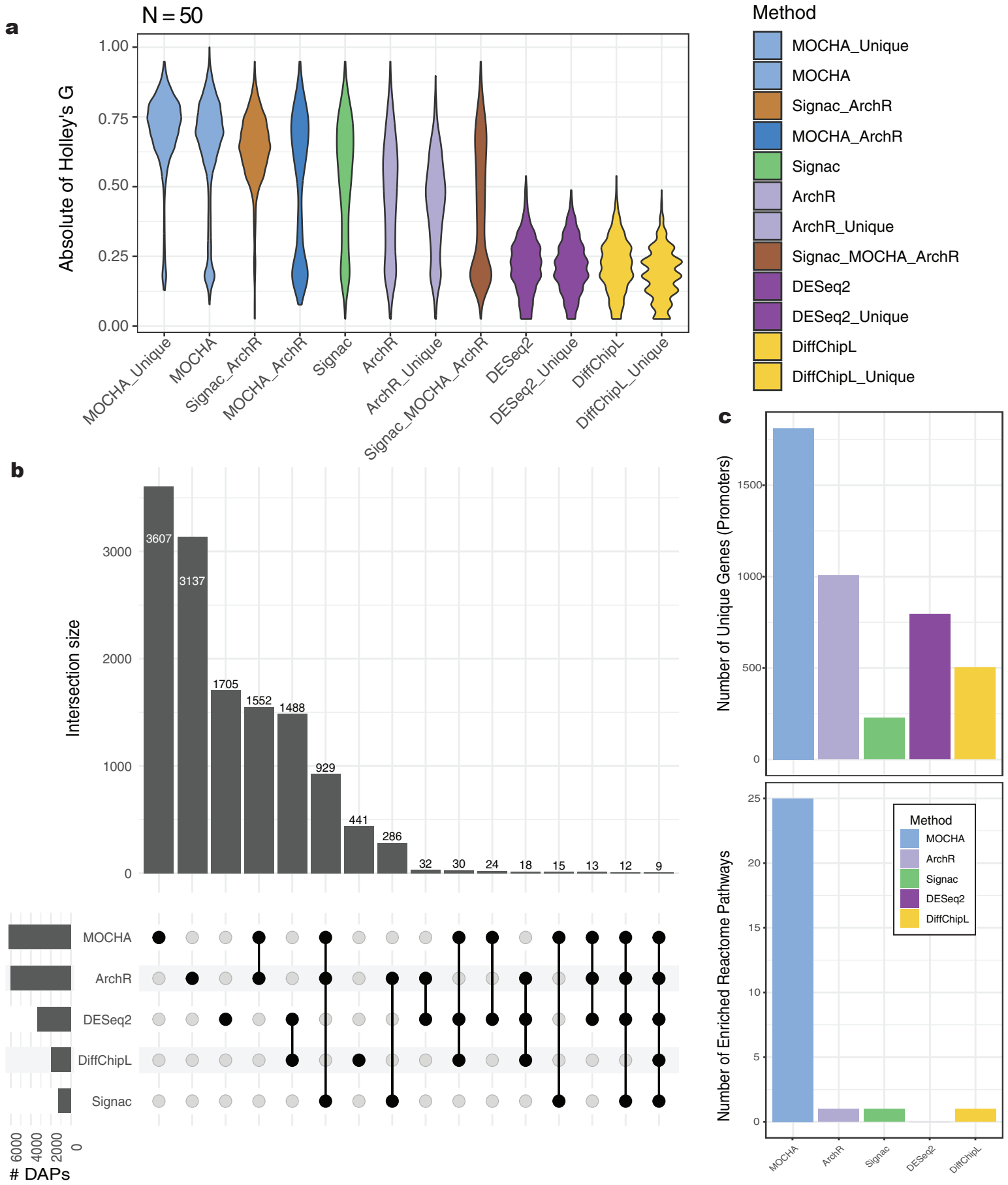


Peakset(s)	Number of Enriched Annotations
MOCHA, MACS2, & HOMER	188
MOCHA Uniquely	5
MACS2 & HOMER	2
MOCHA & HOMER	2
MACS2 Uniquely	0
HOMER Uniquely	0

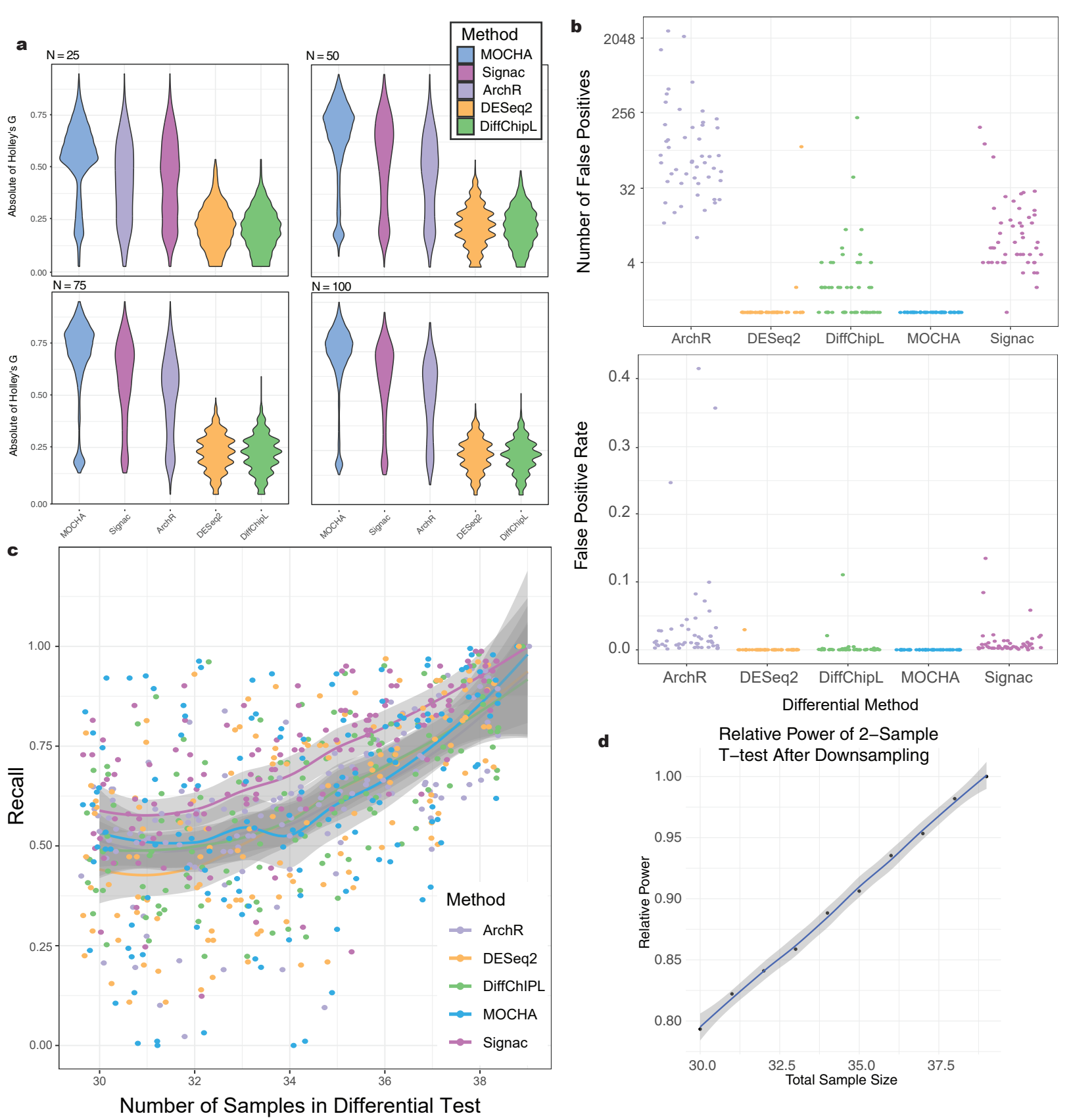
Supplementary Fig. 9. Venn diagram of enriched annotations on peakset from linkage disequilibrium score regression (LDSC). The LSDC analysis summarized across three genome-wide association studies (GWAS) studies (Immune Aging, Auto-immunity, and Abnormal Immune System studies) using the open tiles from the 9 cell types across their respective datasets from Figure 2 (COVID19X, Hematopoiesis, Healthy Donors). Source data are provided in **Source Data Supplementary Figure 9**. Figures were generated using Adobe Illustrator.



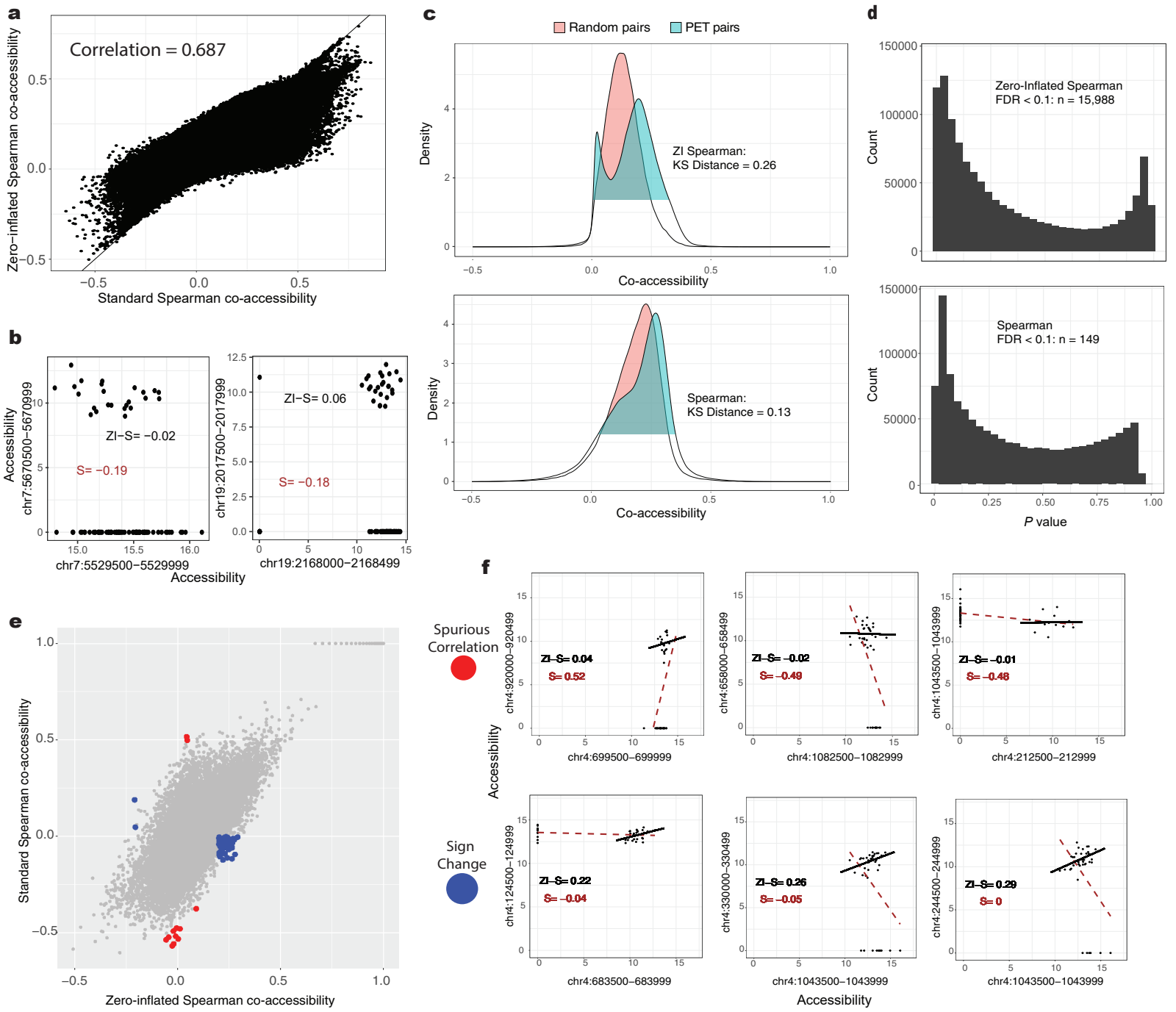
Supplementary Fig. 10. Additional information on differential accessibility analysis.
a, Histogram of P values evaluated by MOCHA on filtered open tiles. Only P ≤ 0.95 values were used for estimating false discovery rate (FDR). **b**, Pathway enrichment analysis on genes having differential accessibility tiles (DATs) in their promoter regions using the Gene Ontology (left), KEGG (middle), or Wikipathway (right) database. The DATs were identified by MOCHA, ArchR, or Signac. Source data are provided in **Source Data Fig. 3-S5**. Figures were generated using Adobe Illustrator.



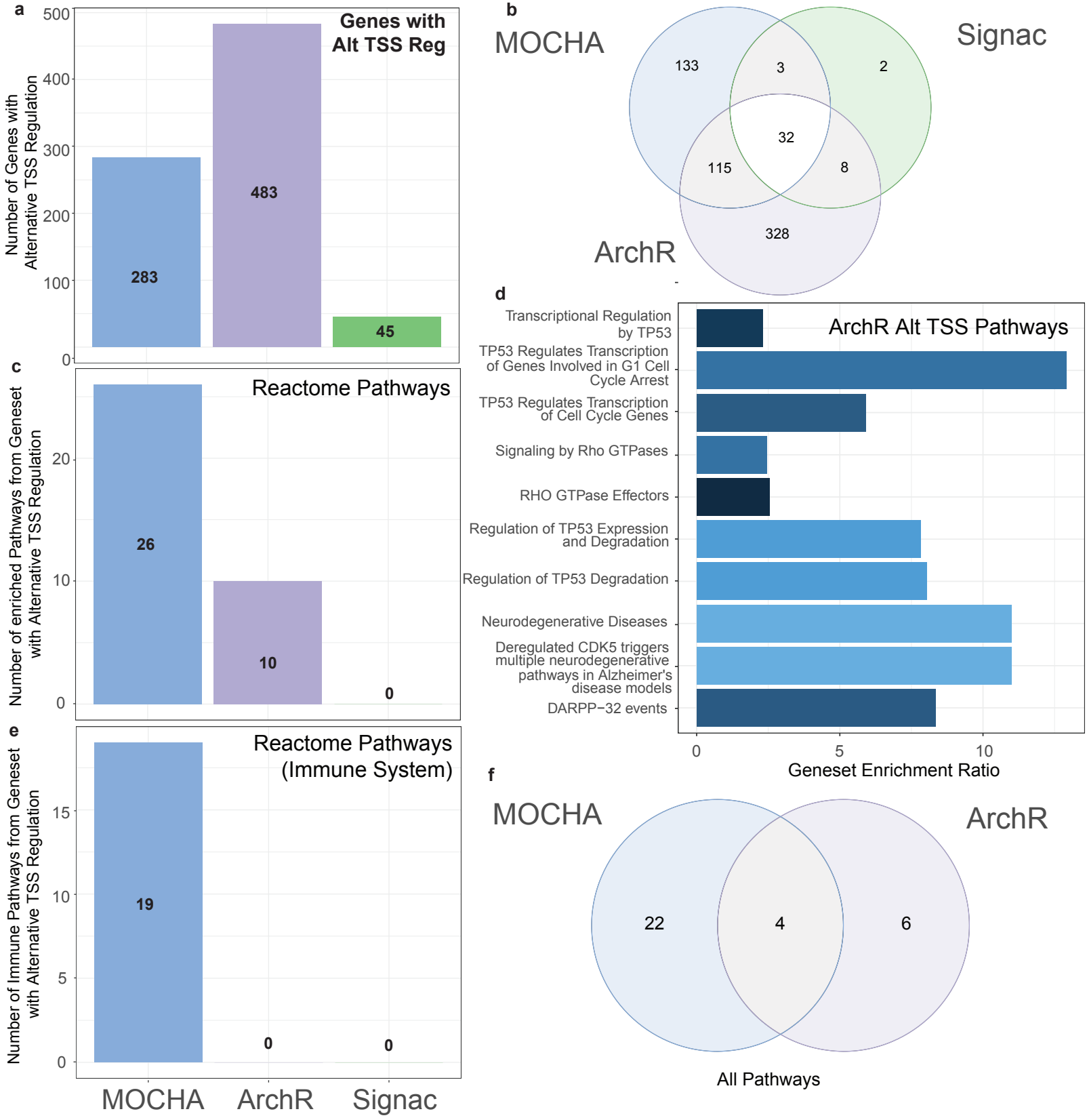
Supplementary Fig. 11. Benchmarking against additional methods for differential accessibility analysis.
a, Violin plot of Holley's |G| from 1000 bootstrapped samples, each containing 50 randomly selected differential accessible tiles (DATs) from each category. The DATs were evaluated based on CD16 monocytes between COVID+ samples during early infection (n=17) and COVID- samples (n=22) in the COVID19X dataset (n=39). Methods tested include MOCHA, Signac, ArchR, DESeq2, DiffChipL, their pairwise intersections, and their uniquelyidentified tiles. Any subset having less than 100 DATs was excluded from the comparison. **b**, Upset plot indicating the overlap of DATs across methods. The total count is displayed in each bar plot. **c**, Bar plots showing the number of unique genes with differential promoters (top) and the corresponding number of enriched Reactome pathways (bottom) as obtained by each method. Source data are provided in **Source Data Supplementary Figure 11**. Figures were generated using Adobe Illustrator.



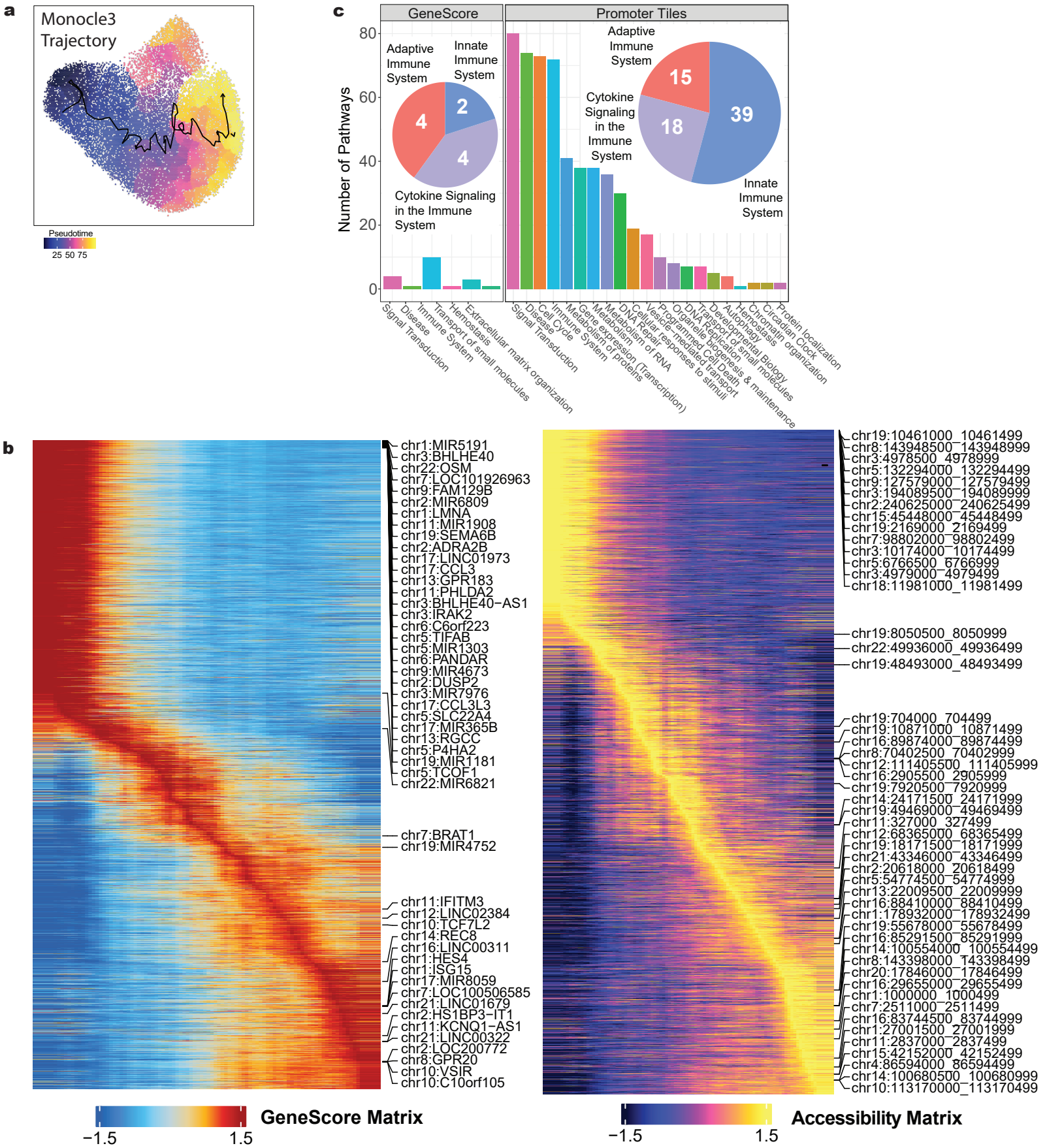
Supplementary Fig. 12. Additional benchmarking on methods for differential accessibility analysis.
a, Violin plots of Holley's |G| from 1000 bootstrapped samples, each evaluated with 25, 50, 75, or 100 randomly selected differential accessible tiles (DATs) from each category. The DATs were evaluated based on CD16 monocytes between COVID+ samples during early infection (n=17) and COVID- samples (n=22) in the COVID19X dataset (n=39). **b**, Number of false positives (top panel) and false positive rate (FPR, bottom panel) from 50 permutation tests in which COVID+ and COVID- labels were randomized. **c**, Recall results from downsampling analysis on tiles in Chromosome 4. Samples were randomly selected (without replacement) for differential accessibility analysis. Recall was calculated as the ratio of the number of DATs in the downsampled sample set to the original number of DATs with n=39 samples. Downsampling was conducted 15 times for each sample size. For all panels, methods tested include MOCHA, Signac, ArchR, DESeq2, and DiffChipL. **d**, The relative power of detecting an effect size of d=0.5 under an =0.05 in a 2-sample t-test as sample size reduces from n=39 to n=30. The statistical power at each n was divided by the power observed at n=39 to obtain the relative statistical power. **b-d**, The solid lines show the Loess smoothing curves and the shaded regions the corresponding 95% confidence intervals. Source data are provided in **Source Data Supplementary Figure 12**. Figures were generated using Adobe Illustrator.



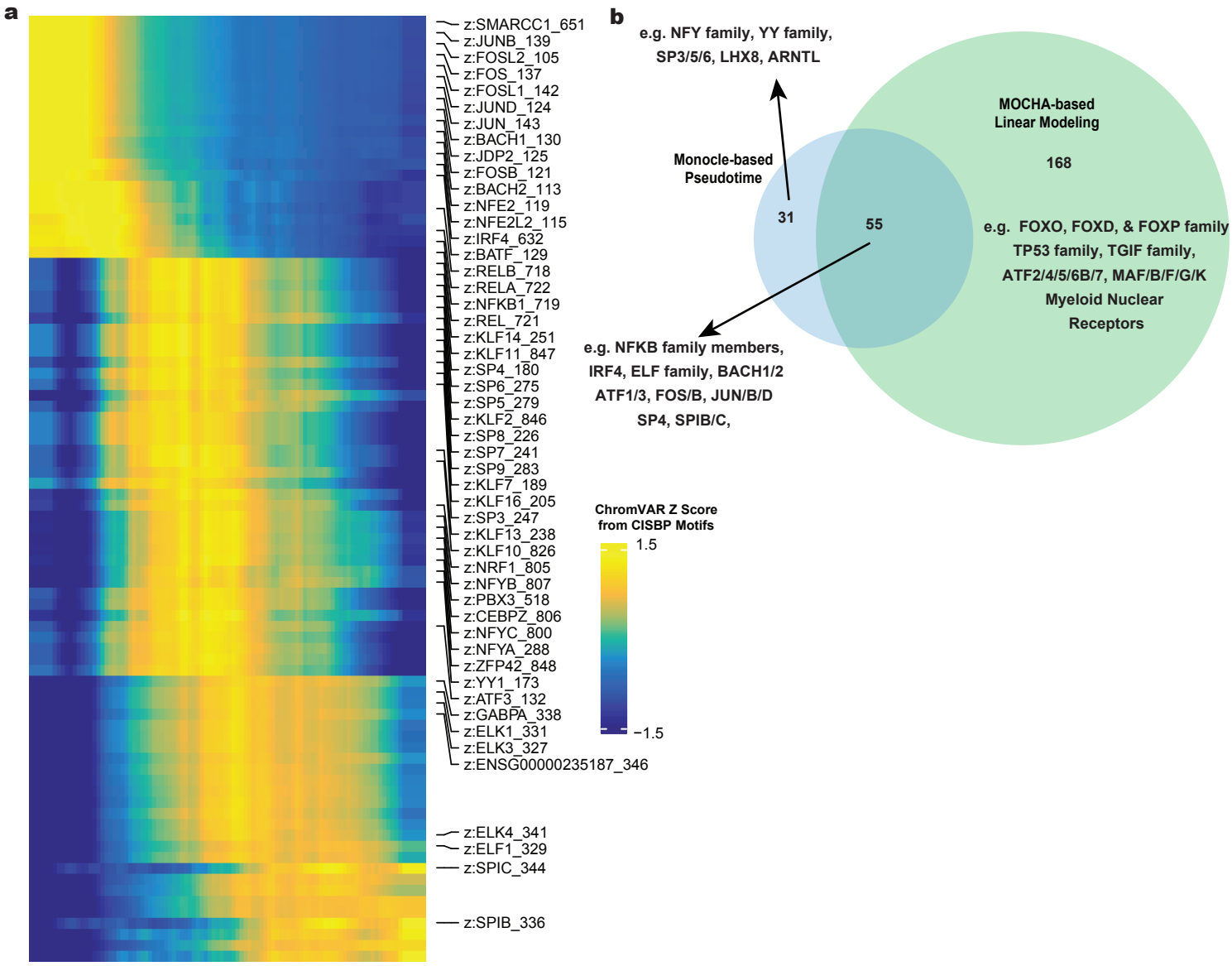
Supplementary Fig. 13. Comparing standard Spearman and zero-infated Spearman in evaluating co-accessibility. **a**, A direct comparison between standard Spearman (S) and zero-inflated Spearman (ZI-S) in evaluating inter-cell type co-accessibility between tile pairs linked to promoter-enhancer interactions in naive CD8 and CD4 T cells54. Co-accessibilities between individual promoter-enhancer tile (PET) pairs (n=1.2 million) were conducted across 17 cell types in the full COVID19 dataset (n=91 samples). **b**, Examples of divergence between the zero-inflated Spearman and the standard Spearman co-accessibilities on two PET pairs. Each dot represents one cell type and sample combination (n=1547 data points). **c**, Distributions of co-accessibilities between PET pairs and randomly selected tile pairs (n=100k) as evaluated by ZI-Spearman (top) or Spearman (bottom). The Kolmogorov-Smirnov (KS) distance was used to quantify the separation between the distributions of PET pairs and random pairs. **d**, Histogram of empirical P values of PET pairs that were calculated based on the percentile of random pairs, using either ZI-Spearman (top) or Spearman (bottom). False discovery rate (FDR) estimation was conducted for each P value set. The number of PET pairs with FDR < 0.1 is shown. **e**, Comparison of inter-sample co-accessibility within CD16 monocytes in the COVID19X dataset (n=39 samples) as evaluated by standard Spearman or ZI-Spearman. All possible pairs of tiles within the first million base pairs of Chromosome 4 were evaluated for an illustrative purpose. **f**, Examples of spurious correlations (top) and sign changes (bottom) generated by s tandard Spearman correlations on ZI data, as compared to results from the ZI-Spearman. Source data are provided in **Source Data Supplementary Fig. 13-1** and **13-2**. Figures were generated using Adobe Illustrator.



Supplementary Fig. 14. Detection of alternative transcription starting site regulation across methods.
a, Bar plot showing the number of alternatively regulated genes (ARGs) in CD16 monocytes during early COVID19 infection as compared to uninfected controls. The ARGs were identified by MOCHA, ArchR, or Signac on the COVID19X dataset (n=39). The total number of ARGs is labeled within the barplot. **b**, Venn Diagram showing the overlap between ARGs as identified by the three methods. The total number of ARGs for each subset is marked. **c**, Bar plot and count indicating the enriched Reactome pathways, based on ARGs from each method. **d**, Barplot of geneset enrichment ratio for Reactome pathways, based on ARGs by ArchR. **e**, Bar plot and count showing the enriched immune pathways in the Reactome datasets, based on ARGs for each method. **f**, Venn Diagram showing the overlap of all enriched Reactome pathways identified by MOCHA and ArchR from panel **c**. Each subset is labeled with the corresponding number of pathways. Source data are provided in **Source Data Supplementary Figure 14**. Figures were generated using Adobe Illustrator.

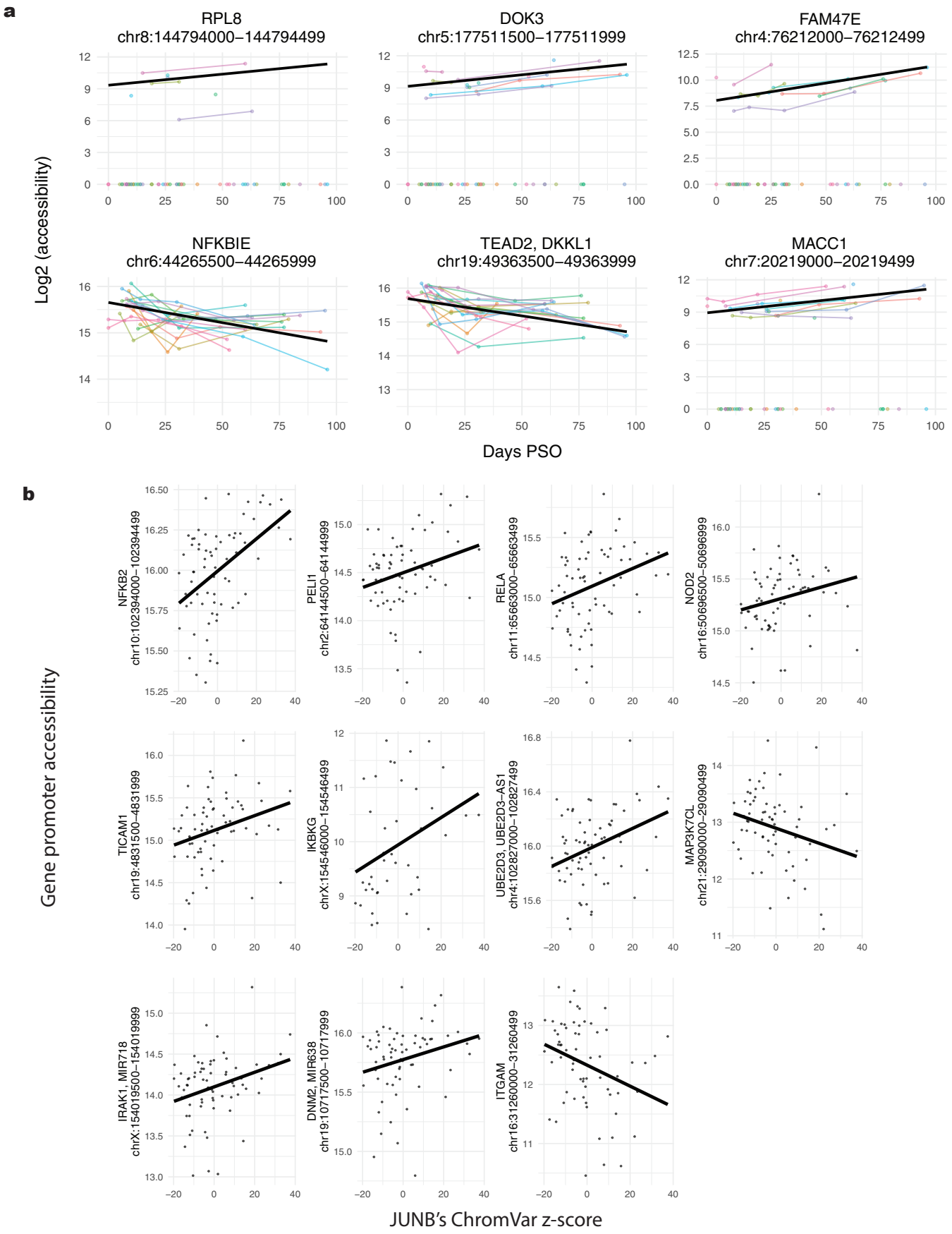


Supplementary Fig. 15. Benchmarking pseudotime trajectory analysis between MOCHA's tiles and ArchR's genescores. **a**, Monocle3 trajectory constructed from CD16 monocytes belonging to samples in the order of early infection (1–15 days PSO, n=21), late infection (16–30 days PSO, n=13), recovery (>30 days PSO, n=35), and uninfected (n=22) in the COVID19 dataset. The trajectory is overlaid on the corresponding single-cell UMAP. **b**, Pseudotime heatmaps of ArchR's genescores (left) and MOCHA's accessible tiles (right) that were generated using ArchR standard settings. The top 50 genes or tiles, respectively, are labeled. **c**, Significant Reactome pathways (FDR < 0.05) enriched with genes having highly variable genescores or promoter accessibility changes along the pseudotime trajectory. The variability threshold was set using ArchR's standard threshold (varCutOff = 0.9). The pathways were aggregated into upper-level pathway annotations using Reactome's database hierarchy. The barplot shows the number of pathways in each category. The pie chart breaks down the immune system pathways by Reactome's next level categories. PSO: post symptom onset; FDR: false discovery rate. Source data are provided in **Source Data Supplementary Fig. 15-16**. Figures were generated using Adobe Illustrator.



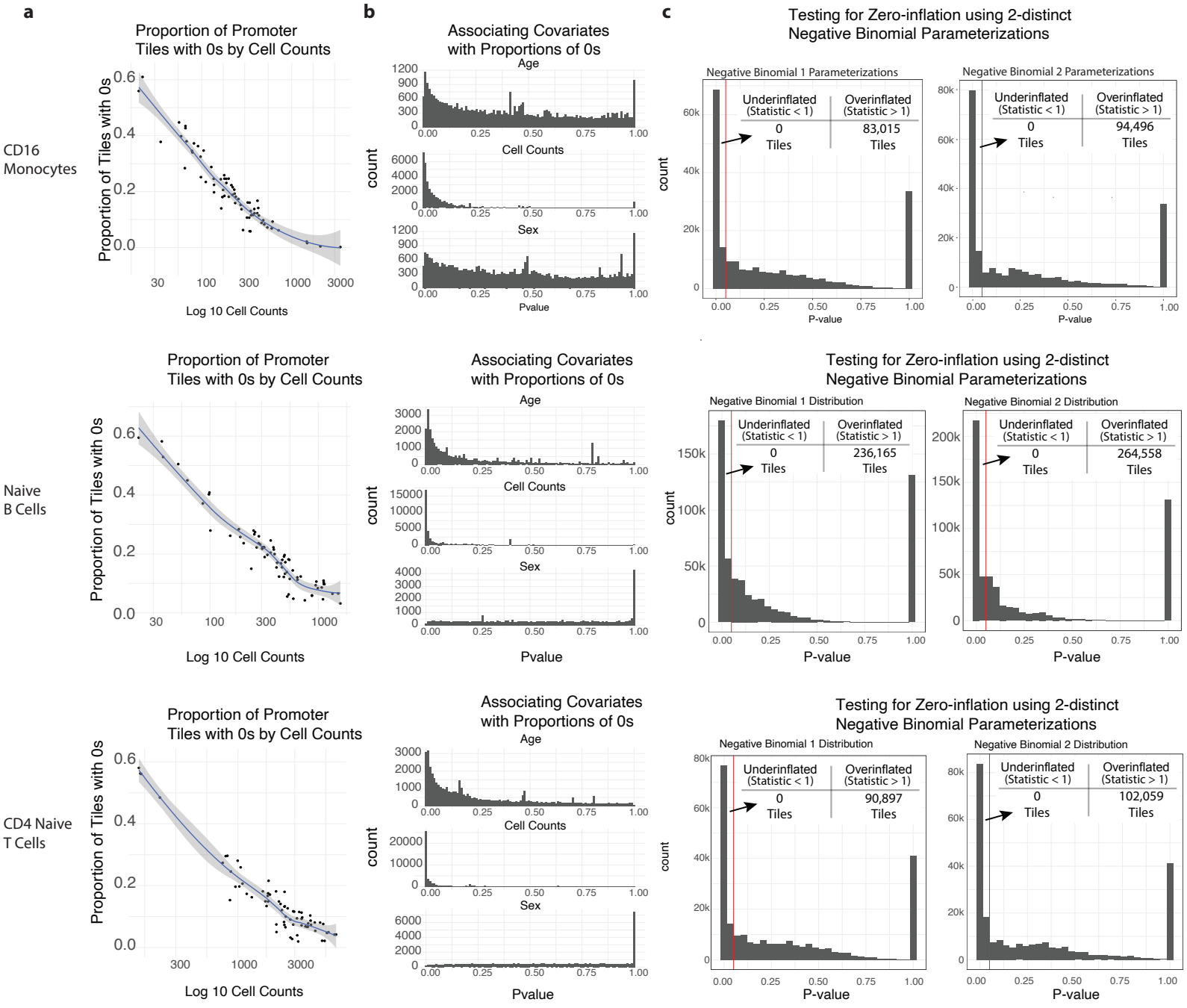
Supplementary Fig. 16. Benchmarking motif usage between single-cell level pseudotime trajectory analysis and pseudo-bulk level real time longitudinal modeling.

a, Pseudotime heatmap of ChromVAR z-scores along the trajectory constructed on CD16 monocytes in the COVID19 dataset (n=91 samples). ChromVAR z-scores were evaluated based on the CISBP database. The top 50 motifs are labeled. Uninfected samples were excluded from the analysis. **b**, Venn diagram comparing motifs showing significant ChromVAR z-score changes either at single-cell level along the pseudotime trajectory (ArchR standard threshold) or at pseudo-bulk level in real time (days PSO) as modeled by generalized linear mixed models (FDR < 0.1). Motif examples for each subsets are provided. FDR: false discovery rate. PSO: post symptom onset. Source data are provided in **Source Data Supplementary Fig. 15-16**. Figures were generated using Adobe Illustrator.



Supplementary Fig. 17. Examples illustrating longitudinal shifts in gene promoter accessibility and motif-promoter associations during COVID-19 recovery.

a, The top 6 genes showing significant promoter accessibility changes (FDR < 0.1) based on ZI-GLMM. Data of individual participants are shown in thin colored lines. **b**, Scatter plots illustrating examples of significant associations (P < 0.05) between JUNB's ChromVAR z-score and significantly changing (FDR < 0.1) promoter accessibility of genes within the TLR4 Reactome pathway. **a-b**, Based on the COVID19L dataset (n=69 samples). The thick black line shows the population trend from ZI-GLMM. PSO: post symptom onset; FDR: false discovery rate; ZI-GLMM: zero-inflated generalized linear mixed effects model; TLR4: toll-like receptor 4. Source data are provided in **Source Data Supplementary Fig. 17**. Figures were generated using Adobe Illustrator.



Supplementary Fig. 18. Patterns of zero-inflation.
a, Scatter plots showing the association between the proportion of promoter tiles being zero (empty tiles) in a given sample (y-axis) and the number of cells (in Log10) in the sample. **b**, P-value distribution for the association between a given covariate and the proportion of zeroes across samples. Zero-inflated linear mixed effect model was applied to analyze the proportion of zeroes as a function of age (top panel), cell count (middle panel), and sex (bottom panel). The histograms summarize the p-values across models for all promoter tiles. Both analyses were based on CD16 monocytes (n =49,679 promoter tiles), naive B Cells (n=53,628 promoter tiles), and CD4 naive T Cells (n=85,016 promoter tiles) in the full COVID19 dataset (n=91). **c**, P-value distribution from testing the presence of zero-inflation against negative binomial (NB) distributions, assuming the variance grows either linearly (left panel) or quadratically (right panel) with the mean. The analysis was performed on open tiles in CD16 monocytes (n =215,649 tiles), naive B cells (n=245,341 tiles), and CD4 naive T cells (n=576,418 tiles) of the COVID19X dataset (n=39). Source data are provided in **Source Data Supplementary Figure 18**. Figures were generated using Adobe Illustrator.

Supplementary Table 1. Contrasting the functionalities between PALMO and MOCHA.

Analysis	PALMO	MOCHA
Variance	Y	N
Outlier	Y	N
Multi-omic	Y	N
Identify	Y	N
scATAC-seq	Y	Y
Longitudinal	Y	Y
Identify open	N	Y
Zero-inflated	N	Y
Identify tile	N	Y
Zero-inflated	N	Y
Identifying	N	Y
Zero-inflated	N	Y
Inferring	N	Y

Supplementary Table 2. Literature references for all Type II alternatively regulated genes, which are indicated as either altered in COVID-19 or identified as potential therapeutic targets.

Gene	Full Name	Comments	References
ATP1A1	ATPase Na+/K+ transporting subunit alpha 1	This protein binds to SARS-CoV-2 viral RNA and has been proposed as a COVID-19 therapeutic target. Previously it is shown to be crucial for coronavirus (CoV) infection and has been proposed as a therapeutic target against CoV infection.	1. Nature Microbiology 6, 339-353 (2021). 2. J. Virol. 89(8), 4434-48 (2015).
UBAP2L	Ubiquitin-associated protein 2 like	This protein competes with SARS-CoV-2 N protein for binding with NTF2 domain of G3BP. It also interacts with BMI1 and regulates the activity of hematopoietic stem cells. It may also activate NF-ŒʃB/RELA (p65).	1. Nat. Commun. 12, 6761 (2021). 2. Blood 124(15), 2362-2369 (2014). 3. Cell Death Discov. 8, 123 (2022).
YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta	This protein is upregulated upon SARS-CoV-2 infection. It is also be proposed as a key player underlying neurodegenerative disease after SARS-CoV-2 infection.	1. Nature 583, 469-472 (2020). Data in Supplementary Table 1 & 2. 2. Med. Hypotheses. 144, 110212 (2020).
CAPN1	Calpain 1	This protein belongs to the calpain protein family which plays multiple, important roles in SARS-CoV-2 infection. It has been proposed as a therapeutic target for both acute and long COVID.	1. Inflammopharmacology 30(5), 1479-1491 (2022). 2. Clin Sci (Lond) 136 (20), 1439-1447 (2022).
ARHGAP9	Rho GTPse activating protein 9	Host factor important to allowing viral replication, Putatively downregulated in response to N-protein from COVID19, linked to mild outcome	1. PNAS 118 (39), e2104759118 (2021). 2. Front. Genet. 12, 763995 (2021).
SOCS3	Suppressor of Cytokine Signaling 3	Supressor of cytokine signaling that who's knockout is embryonically lethal, and who's agonists have been suggested as a prophylatic and/or therapeutic agent for COVID19	1. Frontiers in Immunology 11 (October): 582102. 2. Annals of the Rheumatic Diseases 80 (Suppl 1): 881–881.