Multimodal hierarchical classification of CITE-seq data delineates immune cell states across lineages and tissues

Daniel P. Caron[1], William L. Specht[1], David Chen[2], Steven B. Wells[2], Peter A. Szabo[1], Isaac J. Jensen[1], Donna L. Farber[1,3], and Peter A. Sims[2,4]*

[1]Department of Microbiology and Immunology, Columbia University Irving Medical Center, New York, NY, USA
[2]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA
[3]Department of Surgery, Columbia University Irving Medical Center, New York, NY, USA
[4]Department of Biochemistry and Molecular Biophysics, Columbia University Irving Medical Center, New York, NY, USA

*correspondence to: pas2182@cumc.columbia.edu

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) is invaluable for profiling cellular heterogeneity and dissecting transcriptional states, but transcriptomic profiles do not always delineate subsets defined by surface proteins, as in cells of the immune system. Cellular Indexing of Transcriptomes and Epitopes (CITE-seq) enables simultaneous profiling of single-cell transcriptomes and surface proteomes; however, accurate cell type annotation requires a classifier that integrates this multimodal data. Here, we describe MultiModal Classifier Hierarchy (MMoCHi), a marker-based approach for classification, reconciling gene and protein expression without reliance on reference atlases. We benchmark MMoCHi using sorted T lymphocyte subsets and annotate a cross-tissue human immune cell dataset. MMoCHi outperforms leading transcriptome-based classifiers and multimodal unsupervised clustering in its ability to identify immune cell subsets that are not readily resolved and to reveal novel subset markers. MMoCHi is designed for adaptability and can integrate CITE-seq annotation of cell types and developmental states across diverse lineages, tissues, or individuals.

INTRODUCTION

Recent advances in high-dimensional profiling of single cells, most notably by single-cell RNA sequencing (scRNA-seq), have transformed our ability to define the function and heterogeneity of cell populations in diverse biological systems[1–4]. Because many of the features that define cell types are not captured by scRNA-seq, multimodal single-cell technologies have been developed, including Cellular Indexing of Transcriptome and Epitopes by sequencing (CITE-seq)[5] and RNA expression and protein sequencing assay (REAP-seq)[6] for simultaneous profiling of surface proteomes and transcriptomes. Integrating these high-dimensional modalities to identify cell subsets, developmental states, and other cellular properties with high fidelity across disparate datasets remains a challenge.

Many analytical tools have been developed for the cellular annotation of scRNA-seq data. Unsupervised clustering is commonly used to segregate events into populations sharing similar expression profiles. This approach has proven invaluable for characterizing cellular heterogeneity[7], and has been adapted for CITE-seq datasets[8,9]. However, the number, type, and identity of clusters can be difficult to compare across studies[7,10]. Supervised classification using cell atlases enables cross-study comparisons[11,12], but requires reference datasets which are not available for all tissues and contexts[13–15]. Methods for reference-free annotation of cell types have also been developed, relying instead on knowledge-based marker definitions[16,17]. These tools allow researchers to easily adapt supervised methods to relevant cell types, but do not yet support multimodal datasets.

The development of reference-free, multimodal classifiers to analyze single-cell data is of particular importance for studying the immune system. Immune cells comprise multiple disparate lineages—each of which can be subdivided into functionally distinct, but closely related subsets

that represent various developmental and activation states[18]. Cells of the immune system localize to nearly all tissues of the body, including specialized lymphoid organs, mucosal tissues, and barrier sites[19,20], and exhibit site-specific adaptations[21]. While broad lineages, such as myeloid cells of the innate immune system and adaptive lymphocytes, are readily distinguished from one another at the mRNA level, subsets within these lineages are not. In particular, T lymphocytes can exist as naive or memory subsets stratified by migration pattern, function, tissue residence, and activation[22–24], but these subsets are inconsistently resolved by scRNA-seq[25–28]. Innate-like lymphocytes such as γδ T cells and Natural Killer (NK) cells also share functional profiles which are not readily discernible on the single-cell level[15,29–31]. Multimodal approaches, incorporating both surface markers and transcriptional expression profiles are therefore needed to accurately annotate immune cell subsets and their heterogeneous features.

Here, we developed a supervised approach for cell type annotation of CITE-seq data, designated Multi-Modal Classifier Hierarchy (MMoCHi) that incorporates surface protein and transcript features for reference-free classification. To benchmark this tool against other annotation methods, we sorted and profiled T cell subsets using CITE-seq and demonstrate improved performance by MMoCHi over alternative methods, particularly in the annotation of subsets with highly similar expression profiles. We apply MMoCHi to produce integrated annotations for a cross-tissue CITE-seq atlas of diverse immune cell populations isolated from complex tissue samples. Extracting features important for MMoCHi classification revealed highly interpretable learned representations of cell types, which we used to uncover novel markers for distinguishing transcriptionally similar T cell subsets. Together, MMoCHi readily enables multimodal cell type annotation based on marker genes and proteins and is designed for applicability to CITE-seq of any cell lineage or sample type.

RESULTS

Algorithm Overview

The MMoCHi algorithm uses a hierarchy of random forest classifiers trained on gene expression (GEX) and antibody-derived tags (ADTs) for cell type classification of CITE-seq data (Fig. 1). Prior to classification, ADT expression is batch-corrected using landmark registration, as previously applied to flow cytometry[32] and CITE-seq[33] (see Methods). We identified populations exhibiting negative (background) and positive ADT expression and applied warping functions to align their midpoints (landmarks) across batches, to effectively integrate CITE-seq expression (Fig. 1a). MMoCHi then classifies cell types based on a user-supplied hierarchy of subsets paired with marker-based definitions (Fig. 1b). At each classification node, high-confidence members of each subset are identified using manual thresholds on user-provided gene and protein markers. A random forest is then trained on a representative set of these high-confidence events and used to annotate all events—including those not labeled by high-confidence thresholding (Fig. 1b; Extended Data Fig. 1). Once trained, classifiers can be interrogated for features important for cell classification or applied to extend cell type annotation to other datasets.

Superior annotation of closely related immune cell subsets by MMoCHi

MMoCHi is designed to improve the annotation of subsets with highly similar transcriptomic profiles. We chose to test its performance using T cell subsets, which are well-defined by surface marker expression and functional readouts[22,23], but challenging to annotate by scRNA-seq alone. For example, $CD4^+$ and $CD8^+$ T cells are often imperfectly resolved, partially due to low *CD4* transcript expression[27,28,34]. Moreover, conventional human T cells are delineated into subsets based on differentiation state and migration capacity into naive T cells ($CCR7^+$

CD45RA$^+$ CD45RO$^-$) and memory T cells, which comprise central memory (T_{CM}; CCR7$^+$ CD45RA$^-$ CD45RO$^+$), effector memory (T_{EM}; CCR7$^-$ CD45RA$^-$ CD45RO$^+$), and terminally differentiated (T_{EMRA}; CCR7$^-$ CD45RA$^+$ CD45RO$^-$) subsets[22,23]. Distinguishing naive T cells from T_{CM} and T_{EM} from T_{EMRA} by transcriptome alone is not readily accomplished[1,26–28]. We sorted and performed CITE-seq on seven T cell subsets (CD4$^+$ naive, CD4$^+$ T_{CM}, CD4$^+$ T_{EM}, CD8$^+$ naive, CD8$^+$ T_{CM}, CD8$^+$ T_{EM}, and CD8$^+$ T_{EMRA}) and monocytes (Fig. 2a; Extended Data Fig. 2a; Supplementary Tables 1, 2). CCR7 staining by CITE-seq is suboptimal[5], so we used CD62L which has high concordance in human blood[22,23]. To eliminate batch effects, we performed staining with CITE-seq antibodies before sorting and labeled sorted populations with hashtag antibodies prior to pooling all samples for library preparation and sequencing. Hashtagged populations then served as known references to evaluate classification.

All sorts were of high purity (>91%; Supplementary Table 3), and subsets reflected expected ADT marker expression (Extended Data Fig. 2b). We devised and applied a MMoCHi hierarchy using the same markers used for sorting (Fig. 2b, c; Supplementary Table 4). For comparison to unsupervised approaches, we used totalVI[9] to calculate a multimodal latent space, then performed clustering and manual annotation (Fig. 2c). The high concordance between sorted cell type and MMoCHi classification compared to manual annotation is visually evident from UMAPs in Fig. 2c. Indeed, MMoCHi classification had greater than 90% agreement with sorted labels (Fig. 2c,d), and classified populations showed expected ADT expression (Extended Data Fig. 2b). MMoCHi classification was accurate when downsampling GEX and/or ADT reads, revealing robustness to lower data quality (Extended Data Fig. 3). Here, MMoCHi was insensitive to GEX coverage likely because both the reference dataset and hierarchy for classification were defined by protein expression.

To compare performance between various methods, we calculated precision, recall and F1 score for each subset, as well as overall accuracy (Fig. 2e, Supplementary Table 5). MMoCHi accurately classified all subsets, with an average F1 score of 0.93 (Fig. 2e: MMoCHi). We tested variations to validate MMoCHi's design. Performance worsened when training random forests using only GEX and/or when classifying all subsets with only one random forest (instead of a hierarchy) (Fig. 2e: MMoCHi GEX, MMoCHi Flat, and MMoCHi GEX Flat). We also trained a MMoCHi classifier directly using sorted labels, instead of high-confidence thresholding. Performance on a 20% hold-out revealed a modest improvement for some, but not all, cell types (Fig. 2e: MMoCHi Sort-ref). This effect was lost when training with only GEX (Fig. 2e: MMoCHi GEX Sort-ref). Together, these results support multimodal hierarchical classification and high-confidence thresholding for training.

We next compared MMoCHi classification to manually annotated unsupervised clusters derived from GEX, ADT expression, or the totalVI latent space (see Methods; Fig. 2e: GEX Leiden, ADT Leiden, totalVI Leiden; Extended Data Fig. 2c-g). MMoCHi outperformed all manual annotation methods, particularly GEX Leiden, with $CD4^+$ T_{CM} failing to cluster separately (Extended Data Fig. 2f). To ensure manual annotation was not hampered by lack of clustering resolution, we also over-clustered the data by re-clustering each Leiden cluster (see Methods), but this resulted in only a slight improvement (Fig. 2e: GEX Leiden OC, ADT Leiden OC, totalVI Leiden OC).

Lastly, we tested three scRNA-seq supervised classification methods: CellTypist[25], HieRFIT[34], and Garnett[16]. As there were no pretrained models at an appropriate annotation granularity, we trained new models with sorted labels and evaluated performance on a 20% hold-out. CellTypist, which annotates events individually or by majority voting across clusters, and

HieRFIT, which classifies hierarchically, outperformed GEX Leiden, but not MMoCHi. (Fig. 2e: CellTypist Sort-ref, CellTypist MV Sort-ref, HieRFIT Sort-ref, respectively). Garnett classifies hierarchically, and additionally avoids low-probability classifications. The events fully annotated by Garnett (46% of held-out events), were more accurate than other scRNA-seq classifiers, but less accurate than MMoCHi (Fig. 2e: Garnett Sort-ref). Because Garnett was developed for reference-free training using marker-based definitions[16], we also trained a model using selected marker genes for each subset (Supplementary Table 6; see Methods). Unfortunately, marker genes were inadequate, and events fully annotated by Garnett (23% of events) had low accuracy (Fig. 2e: Garnett Markers). Notably, all the supervised approaches described above effectively distinguished monocytes from T cells (Fig. 2e), consistent with previously documented high performance of these tools[16,25,34]. Our evaluation emphasizes the difficulty of segregating T cell subsets sharing highly similar transcriptomic profiles without profiling surface protein. By leveraging all available modalities, MMoCHi accurately annotated these highly similar cell types.

MMoCHi integrates classification across diverse human tissue immune cells

We next applied MMoCHi to total immune cells from lymphoid and mucosal tissue samples obtained from human organ donors[25] (Fig. 3a; Supplementary Table 1). Immune cells were enriched from eight sites across two donors and included lung (LNG), bronchial alveolar lavage (BAL), lung-associated lymph node (LLN), spleen (SPL), jejunum epithelial layer (JEL), jejunum lamina propria (JLP), bone marrow (BOM), and blood (BLD), using methods optimized for each site[25,27]. We performed CITE-seq to profile over 270 surface markers expressed by immune and non-immune cells (Supplementary Table 2). In previous analysis of single-cell transcriptomes from this dataset, we detected all lineages of immune cells, including T cells, B cells, innate lymphocytes, and myeloid cells across multiple sites[25]. However, refined immune cell

subsets and differences across tissue sites were difficult to resolve in this complex dataset and required extensive manual annotation. Here, we integrated protein and transcriptome profiling using MMoCHi to determine whether transcriptionally similar immune subsets could be identified across diverse tissues.

For visualization, we computed a UMAP of a donor-integrated totalVI latent space, revealing multiple groupings, some of which corresponded to different tissue sites (Fig. 3b; see Methods). We constructed and applied a MMoCHi hierarchy representing all expected cell types using a combination of transcript and surface protein markers (Fig. 3c; Supplementary Table 7). Training and classifying these approximately 198 thousand immune cell events into 26 subsets took less than 20 minutes (Extended Data Fig. 4, see Methods), demonstrating that MMoCHi is highly scalable. The resultant model, using 200 estimators in each random forest, was well-fit at every level of the hierarchy, as measured by the prediction accuracy for subsets in the same classification layer and across the full hierarchy (Fig. 3d; Extended Data Fig. 5; see Methods). MMoCHi classified cell types across tissue sites and consistently across donors (Fig. 3e, Extended Data Fig. 6).

We compared MMoCHi classification to manually annotated clusters of the multimodal totalVI latent space (Fig. 4a,b; Extended Data Fig. 7a,b). The methods were broadly concordant, with 72% of events labeled identically; however, disagreements occurred between cell types with similar expression profiles. Notably, unsupervised clustering failed to resolve any $CD8^+$ naive T cells, $CD8^+$ T_{CM}, γδ T cells, and naive B cells. We investigated discrepancies using marker expression (Fig. 4c-e, Extended Data Fig. 8). A substantial percentage (9%) of αβ T cells had conflicting $CD4^+$ or $CD8^+$ annotations, and by protein and transcript expression MMoCHi classifications were more appropriate (Fig. 4c). Cytotoxic lymphocytes are found across multiple

lineages (NK cells, ILCs, $CD8^+$ T_{EMRA}, and γδ T cells) and are difficult to resolve; however, MMoCHi correctly classified NK cells and ILCs by their lack of CD3 or TCRαβ surface expression, despite expression of *CD3E* and *TRDC* transcripts[15,29] and their co-clustering with $CD8^+$ T_{EMRA} (Fig. 4d). MMoCHi also correctly distinguished CD8 T_{EMRA} and γδ T cells, marked by expression of either CD3 and TCRαβ or CD3, TCRγδ, and variable expression of *TRDV1* and TCRVδ2, respectively (Fig. 4d). MMoCHi also improved identification of T cell memory subsets by expression of CD62L, CCR7, CD45RA, and CD45RO (Fig. 4e). Together, MMoCHi classifications improved identification of known immune cell subsets, particularly in cases of discordant mRNA and protein expression.

MMoCHi classifiers are highly interpretable and identify cell type markers

Having demonstrated MMoCHi's utility for cell type annotation, we next determined which features (transcript and surface protein expression) were useful for subset delineation. Many single-cell classifiers are trained on dimensionally reduced data[12], hampering feature-level interpretability, but random forests within MMoCHi are trained using all available protein-coding genes and surface proteins. During training, features are selected for their contribution to decreasing impurity, thus providing a natural ranking of features by their importance for classification[35,36] (see Methods).

We analyzed features important for MMoCHi classifiers at various levels, including delineating total lineages, monocytes and macrophages, lymphocyte subsets, and B cell-like populations (Fig. 5, Supplementary Table 8). To display features associated with each subset, we selected the top important features with a $\log_2$(fold-enrichment) > 2 for that subset. All selected features were within the top 1% of important features at the given level and significantly differentially expressed ($p < 0.05$; Supplementary Table 9). Important features included markers

used for high-confidence thresholding, including: *CPA3*, *MS4A2*, *TPSB2*, *PRSS57*, CD33, Podoplanin, and CD352 for Lineage, *SELL*, *S100A8*, *SEPP1*, and *FCGR3A* for Mono/Mac, *KLRF1*, *IL7R*, *JCHAIN*, *MZB1*, CD3, CD5 and TCRαβ protein for Lymphocytes, and *PLD4*, *MZB1*, and *JCHAIN* for B cell-like (Fig. 5; Supplementary Table 8). We also identified other key subset markers that were not used in high-confidence thresholding, including: *IL1R1* and *GATA2* for mast cells[37], *APOE*, *ACP5*, and *C1QC* for macrophages[38], *CD79A*, *MARCKS*, *MEF2C*, and CD32 for B cell-like[39–42], and *TYROBP*, *GZMB*, and CD123 for pDCs[43,44] (Fig. 5). Overall, we demonstrate that MMoCHi learns informative cell type representations—an important pre-requisite for novel marker identification.

MMoCHi uncovers additional gene expression markers distinguishing naive and central memory T cells

Next, we wondered if multimodal high-confidence thresholding could be used to improve transcriptome-based segregation of naive T cells and T_{CM}, which are defined by expression of specific CD45 isoforms but have very similar transcriptomes[1,23,26,28]. First, we used multimodal high-confidence thresholding to train both $CD4^+$ and $CD8^+$ Naive/T_{CM} GEX classifiers (Supplementary Table 7). Despite excluding surface proteome from training, these classifiers were highly accurate, especially for $CD8^+$ Naive/T_{CM}, and resulting subsets had expected expression of CCR7, CD62L, CD45RA, and CD45RO (Fig. 6a,b). To identify additional GEX markers, we next interrogated impurity-based important features (Supplementary Table 10). To minimize contamination of these marker sets by tissue-specific populations, we trained classifiers with the entire dataset and only with immune cells from the blood. Marker genes were selected as the top important features with a $\log_2$(fold-enrichment) > 2, and a greater than 10% change in dropout rate (Fig. 6c,d). All selected genes were within the top 1% of important features and significantly

differentially expressed ($p < 0.05$; Supplementary Table 11). In both $CD4^+$ and $CD8^+$ T_{CM}, we identified a suite of memory-associated genes, including *ITGB1*, *CCR4*, *LMNA*, and *FAM129A*[22,26]. We also identified *PRDM1*, *CCL5*, and *KLRG1*, transcripts primarily associated with T_{EM}, as markers of $CD8^+$ T_{CM}[22,26,45]. Improved prediction accuracy by classifiers trained with only the top 1000 important features confirmed that these features were useful for GEX classification (Fig. 6e-f).

Lastly, we sought to validate these findings using our scRNA-seq of sorted $CD4^+$ and $CD8^+$ naive and T_{CM} populations. Most of the markers identified for naive and T_{CM} were also differentially expressed within the sorted dataset (Fig. 6g,h; Supplementary Table 12). GEX classifiers trained on $CD4^+$ or $CD8^+$ Naive and T_{CM} effectively recapitulated the sorted labels, with F1 scores of 0.80 and 0.78 for the $CD4^+$ and $CD8^+$ classifiers, respectively (Fig. 6i,j). Overall, these findings demonstrate the capacity of MMoCHi to leverage CITE-seq to identify gene expression markers and train classifiers that can be effectively applied to scRNA-seq datasets.

DISCUSSION

The advent of multimodal single-cell technologies has enabled high-dimensional profiling of many systems, organs, diseases, and species. However, the development of analytical tools to identify cell states and their features consistently across multimodal studies is lagging behind these data acquisition technologies. Here, we present MMoCHi, a multimodal, hierarchical classification approach for cell type annotation of CITE-seq data, which integrates both gene and protein expression, does not require reference datasets, and is highly scalable to classify closely related subsets within a single lineage and diverse subsets from across lineages. Applied to immune cell CITE-seq datasets, MMoCHi outperforms current annotation algorithms, and identifies new markers for subset delineation. Together, MMoCHi provides an adaptable approach for applying marker-based annotation to multimodal datasets.

Currently available single-cell classifiers are primarily designed for scRNA-seq; however, CITE-seq can be leveraged to improve annotation[5,6]. MMoCHi robustly classifies closely related immune cell subsets with discordant transcriptome and surface proteome profiles. Here we show MMoCHi effectively distinguishes T cell subsets defined by surface expression of CD45 isoforms and homing receptors but share overlapping transcriptomes[1,23,26,28]—namely naive T cells from T_{CM} and T_{EM} from T_{EMRA}. By classifying hierarchically, MMoCHi was also able to correctly annotate diverse immune lineages without compromising its performance at segregating functionally and transcriptionally similar cell types, including cytotoxic NK cells, CD8+ T_{EMRA}, and γδ T cells[15,29–31]. MMoCHi can accurately classify cell types despite other sources of variation, as shown by integrated classification of multiple immune cell lineages across blood and 8 disparate tissue sites of two organ donors. In this way, MMoCHi can be used to identify cell types for

subsequent analysis of other features that vary across tissue sites, age, disease, sex, and other factors.

By providing a platform for using high-confidence, marker-based thresholding for training data selection, MMoCHi enables the flexible design of new classifiers based on prior knowledge without a requirement for carefully curated reference atlases. Similar to other marker-based strategies, these subset definitions are directly intelligible and can be robustly applied across studies and specific sequencing conditions to train new classifiers[16,17]. Here, we also apply MMoCHi's multimodal training data selection to improve transcriptome-based annotation of T cell subsets across datasets, demonstrating potential for MMoCHi to advance annotation of scRNA-seq data as well.

In contrast with efforts to automate cell type annotation[11,25], MMoCHi's thresholding schemas require careful marker curation and domain expertise, however; by leveraging this expertise, MMoCHi classifications may better reflect canonically defined, and biologically relevant designations. Additionally, MMoCHi is not optimized for novel subset identification, requiring the user to define all cell types for classification. Thus, MMoCHi is complemented by multimodal unsupervised exploration of the dataset to annotate or identify markers of unexpected cell types and states.

Beyond cell type classification, MMoCHi learns key features of protein and gene expression, which can be used to identify new cell type markers and derive biological insights. MMoCHi random forests do not require prior dimensionality reduction, enabling evaluation of individual surface proteins and protein-coding genes as subset markers. In classifying diverse and highly similar subsets, we show MMoCHi extracts relevant markers, relying on both features used for training data selection, and other known gene and protein markers of each subset. Using

MMoCHi, we identified improved transcriptional markers of naive T cells and T_{CM}. Some T_{CM} markers were associated with T cells in tissue[26,27,45] and T_{EM}[22,26], suggesting a role for tissue programming during T_{CM} differentiation and indicating a gradual transition between T cell memory subsets. Importantly, we did not previously identify these expression patterns of circulating T_{CM} in earlier studies of T cells in tissues and blood[26,27], likely due to the lack of CITE-seq and multimodal classification.

While we have developed MMoCHi with CITE-seq applications in mind, the algorithm is designed for easy extension to other modalities frequently paired with single-cell transcriptomes, such as profiling chromatin accessibility, T and B cell receptors, mutational landscapes, intracellular proteins, or a combination of these modalities[46–51]. We also anticipate applications to emerging technologies for multimodal, single-cell spatial profiling[52,53]. While we focused on immunology, cell type classification is a ubiquitous problem in single-cell genomics. Thus, we expect broad utility for MMoCHi in diverse biological applications, including identification of developmental states, building atlases of complex tissues and tumors, profiling model organisms, and analyzing clinical specimens.

ACKNOWLEDGEMENTS

We thank Joshua I. Gray and Rory E. Morrison-Colvin for helpful discussions and members of the Farber laboratory for help with tissue processing. This work was supported by a Seed Networks for the Human Cell Atlas grant from the Chan Zuckerberg Initiative (CZF2019-002452) and NIH grants AI128949 and AI106697 awarded to P.A.Si. and D.L.F. D.P.C. was supported by the Columbia University Graduate Training Program in Microbiology and Immunology (T32AI106711). P.A.Sz. was supported by a Canadian Institutes of Health Research (CIHR) Fellowship. Research reported here was performed in the Columbia Stem Cell Initiative Flow

Cytometry Core, the Sulzberger Columbia Genome Center, and the Columbia Single Cell Analysis Core (supported by grant P30CA013696).

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We wish to thank the donor families for their generosity and the exceptional efforts of the transplant coordinators and staff of LiveOnNY for making this study possible.

AUTHOR CONTRIBUTIONS

D.P.C. designed and performed experiments, analyzed data, made figures, and wrote the manuscript. W.L.S. performed experiments, analyzed data, made figures, and edited the manuscript. D.C., S.B.W., P.A.S., and I.J.J. prepared samples for single cell sequencing. P.A.Si. and D.L.F. designed experiments, analyzed data, wrote, and edited the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

METHODS

MMoCHi

MMoCHi is designed to standardize classification of cell subsets using multimodal (CITE-seq) data (Fig. 1). To train a new MMoCHi model, the user specifies a hierarchy of cell subsets, along with marker definitions (based on thresholds of either gene or protein expression) for each. MMoCHi iterates through each level of the hierarchy, selecting high-confidence members of each subset, training a random forest classifier on normalized gene and protein expression, and classifying cell types. Once trained, MMoCHi classifiers can be interrogated for feature importances or applied to extend cell type annotation to other datasets.

Feature selection, normalization, and batch correction

As input to MMoCHi, gene expression (GEX) was normalized to log(counts*10,000/total_counts+1), and antibody derived tag (ADT) expression was normalized to log(counts*1,000/total_counts+1). For many markers, ADT expression distributions matched the expected bimodal, trimodal, or gradients of expression observed by flow cytometry. To account for variation in antibody staining across samples, ADT expression was batch-corrected using landmark registration (Fig. 1a)[32,33]. First, landmarks (peaks) were identified in the distribution of expression for each ADT in each sample as automatically detected local maxima (*scipy.signal.find_peaks*) on kernel-density-smoothed (*scipy.stats.gaussian_kde*) ADT expression, or manually identified. Curve registration and warping were applied to align these landmarks across samples (*skfda.preprocessing.registration.landmark_elastic_registration_warping*). Once batch-corrected, thresholds delineating positive and negative populations for ADT markers could be applied across batches.

Selection of training data

At each level of the hierarchy, we selected high-confidence events for each subset using manual thresholds on the user-supplied marker definitions (Fig. 1b, Supplementary Tables 4,7). ADT thresholds were drawn using established flow cytometry gating principles. GEX thresholds were drawn similar to Garnett[16], primarily capturing events with any marker expression, or occasionally capturing only the highest expressing events. A portion (20%) of these high-confidence events were held out for testing and validation, and the remaining 80% could be used for training. We then resampled this training dataset (Extended Data Fig. 1a) to remove events likely to be mistakenly labeled due to imperfect marker thresholding (noise) and to overrepresent events likely to be misclassified (in danger). To identify "noise" and "in danger" events, principal component analysis (PCA; *scanpy.pp.pca*) was run on the scaled expression of the top 5000 highly variable genes (*scanpy.pp.highly_variable_genes*) and all ADTs. The 5 nearest neighbors were calculated for each event (*sklearn.neighbors.NearestNeighbors*). Events were considered "noise" and removed if all neighbors disagreed with their high-confidence label. Events were considered "in danger" if less than half the neighbors agreed with their high-confidence label. Training events were also clustered using the Leiden algorithm[54] (*scanpy.tl.leiden*) to identify "in danger" clusters—clusters representing less than 5% of a subset's training events. All events considered "in danger" were oversampled 5 times to increase representation during training. At classification nodes where subsets were not expected to segregate by unsupervised approaches—including Lymphocytes and TCR—these selection steps were skipped. To account for class imbalance in the training dataset, events for all subsets were oversampled to equal numbers. Finally, training events were subsampled without replacement to a maximum of 20,000 events for computational performance. The selection of training data occurred separately for each batch. If a minimum of

100 events were not identified in a batch, events would be spiked into that batch from other batches (Extended Data Fig. 1b).

Training and calibration of random forest classifiers

At each level of the hierarchy, we trained a random forest (*sklearn.ensemble. RandomForestClassifier*) with 100 trees, each with a max depth of 20 to reduce overfitting. In datasets with multiple batches, 100 trees were trained separately for each batch and added to the forest. Once trained, the forest would predict the subset identity of events and provide the proportion of trees in agreement with the classification. To convert these proportions of trees in agreement to probabilities, they were transformed using an isotonic regression (*sklearn.calibration.CalibratedClassifierCV*) trained on a subset of the held-out data[55]. Classification performance was then evaluated on the remaining hold-out data. Hyperparameters were tuned in cases of poor fit. To increase fit, the maximum number of features to consider when looking for the best split during random forest training was set to 10% of total features for $CD4^+$ and $CD8^+$ memory classification for sorted T cells, as well as Lymphocyte, TCR, and CD4/8 classification for organ donor cells. The trained classifier was then used to predict subset identity for all events at that level. This process of training data selection, classifier training, and prediction were repeated until all cells had been labeled to terminal subsets.

CITE-seq profiling of FACS sorted subsets

Peripheral blood from a consenting healthy volunteer (34-year-old male; Supplementary Table 1) was obtained by venous puncture, through a protocol approved by the Columbia University IRB and complying with relevant ethical regulations for work with human participants.

Peripheral blood mononuclear cells (PBMCs) were isolated using RosetteSep Granulocyte Depletion Cocktail (StemCell Technologies), following manufacturer's protocols for density gradient centrifugation. Briefly, samples were incubated with the cocktail at 20°C for 10 minutes, diluted 1:1 with FACS buffer (DPBS 10% FBS 2mM EDTA), then layered over Ficoll-Plaque in SepMate PBMC isolation tubes (StemCell Technologies). Samples were centrifuged at 1200 x g for 10 minutes at 20°C, and PBMC layers were isolated according to instructions. Samples were washed (400 x g for 10 minutes) with FACS buffer. For further erythrocyte removal, pellets were resuspended in ACK lysis buffer (Gibco) incubated for 2 minutes at 37°C and washed with FACS buffer.

PBMCs were stained with Zombie NIR Fixable Viability dye (BioLegend) for 30 minutes. Samples were kept at 4°C in the dark. Cells were washed thrice with FACS buffer, resuspended in TrueStain FcX and TrueStain Monocyte Blocker (BioLegend), and incubated for 10 minutes. We designed a FACS-sort antibody cocktail, prioritizing antibody clones with discrete epitopes from the TotalSeq-A Universal Human Panel (BioLegend) to reduce steric hinderance during CITE-seq staining (Supplementary Table 2). Cells were incubated with the FACS-sort antibody cocktail for 30 minutes, then washed thrice with FACS buffer. Cells were incubated in TrueStain FcX (BioLegend) for 10 minutes, then stained using a custom TotalSeq-A Universal Human Panel (BioLegend) for 30 minutes, according to manufacturer instructions. Samples were washed thrice with FACS buffer. Cells were sorted using a FACS Aria III (BD Biosciences; Extended Data Fig. 2a). Seven T cell memory populations ($CD4^+$ Naive, $CD4^+$ T_{CM}, $CD4^+$ T_{EM}, $CD8^+$ Naive, $CD8^+$ T_{CM}, $CD8^+$ T_{EM}, $CD8^+$ T_{EMRA}), and monocytes were sorted into sterile, heat inactivated FBS. Sort purity was calculated as the number of events falling within a subset's gates divided by the total number of singlet events times 100, using the same gating strategy as the sort (Extended Data Fig.

2a). All sorts were high purity, with mean purity 94.4% (Supplementary Table 3). Sorted populations were washed in FACS buffer and resuspended in TrueStain FcX (BioLegend) for 10 minutes. Samples were then stained with TotalSeq-A hashtag antibodies (BioLegend). In this experiment, the hashtag-oligos (HTOs) correspond to the sorted immune cell subsets (HTO1: $CD4^+$ Naive, HTO2: $CD4^+$ T_{CM}, HTO3: $CD4^+$ T_{EM}, HTO4: $CD8^+$ Naive, HTO5: $CD8^+$ T_{CM}, HTO6: $CD8^+$ T_{EM}, HTO7: $CD8^+$ T_{EMRA}, HTO8: Monocytes). Samples were then washed thrice with FACS buffer, and pooled.

CITE-seq profiling of human tissue samples

Human tissues were obtained from deceased organ donors as previously described[25,27,56,57]. The use of tissues from organ donors is not considered human subjects research as confirmed by the Columbia University IRB because the donors are deceased. Mononuclear cells (MNCs) were isolated from tissue sites of two organ donors (D496 and D503; Supplementary Table 1), as described[25]. Approximately 1 million MNCs per tissue site were washed in FACS buffer and resuspended in TrueStain FcX (BioLegend) for 10 minutes. Samples were then stained with TotalSeq-A hashtag antibodies (BioLegend) for each tissue. For each donor, the hash-tagged MNCs from each tissue site were pooled, washed with FACS buffer, and stained with the TotalSeq-A Human Universal Cocktail panel according to the manufacturer's instructions (BioLegend).

Library preparation, sequencing, and alignment

The sorted immune cells were counted, diluted to an appropriate volume, and loaded across two lanes of a 10X Genomics Chromium instrument targeting 6,000 cells each. Samples from each organ donor were loaded across 16 lanes of a 10X Genomics Chromium instrument targeting

10,000 cells each. cDNA synthesis, amplification and sequencing libraries were generated using the Next GEM Single Cell 3' Kit v3.1 (10X Genomics) with the recommended modifications for compatibility with the TotalSeq-A cell hashing and CITE-seq reagents (BioLegend). Organ donor GEX libraries were sequenced on a NovaSeq 6000 (Illumina) with 100 cycles for reads 1 and 2. Organ donor ADT and HTO libraries were sequenced on a NextSeq 500 (Illumina) with 28 cycles for read 1 and 55 cycles for read 2. All libraries for FACS sorted subsets were sequenced on a NextSeq 500 (Illumina) with 28 cycles for read 1 and 44 cycles for read 2.

Reads were analyzed by pseudoalignment using kallisto v0.46.2 (GRCh38 with Gencode v24 annotation) and bustools v0.40.0[58–60]. CITE-seq and hashtag barcodes were demultiplexed and extracted using DropSeqPipeline8, as previously described[61]. Hashtags were demultiplexed by CLR normalization, k-means clustering, and statistical identification of singlets by fitting a negative binomial model as described[5].

Analysis and benchmarking using sorted T cells

Cells from the T cell sort were filtered to remove events with fewer than 1000 unique counts, fewer than 200 genes detected, or over 10% mitochondrial counts. A MMoCHi hierarchy was developed (Fig. 2b) and classification performed using the algorithm above. In flat classification variations (MMoCHi Flat, MMoCHi GEX Flat), high-confidence thresholding and classification were performed for all subsets in a single classification node. In GEX variations (MMoCHi GEX, MMoCHi GEX Flat, and MMoCHi GEX Sort-ref), ADT expression data was excluded from the detection of in-danger noise events and random forest training. In "Sort-ref" variants (MMoCHi Sort-ref and MMoCHi GEX Sort-ref), training events were selected using the sort labels instead of high-confidence thresholding. To mirror intra-dataset performance testing of

other reference-based tools[25,34], a portion of the dataset (20%) was held-out from training and used only for performance evaluation.

We manually annotated unsupervised clusters by average expression of gene and protein markers. Leiden clusters of gene expression (GEX Leiden) were computed on a PCA of the highly variable genes with Scanpy[62] defaults. Leiden clusters of ADT expression (ADT Leiden) were computed on a PCA of all ADTs except for isotype controls. totalVI latent space was computed on all ADTs except for isotype controls, and highly variable genes selected using top 4000 genes as defined by the Seurat v3 method[63], as recommended[9]. Leiden clustering was performed on the 10 nearest neighbors of the top 40 principal components or the entire totalVI latent space. Over-clustering was also computed where highly variable gene selection, dimensionality reduction, and Leiden clustering were repeated to sub-cluster each cluster resulting in 79 GEX clusters (GEX Leiden OC), 128 ADT clusters (ADT Leiden OC), and 57 totalVI clusters (totalVI Leiden OC). For visualization, UMAP (*scanpy.tl.umap*) embeddings of each of these feature spaces were calculated using defaults.

CellTypist[25], HieRFit[34], and Garnett[16] were trained using the sort labels as reference and a 20% hold-out for performance testing (as with MMoCHi Sort-ref). A CellTypist model was trained (*celltypist.train*) with two-pass training enabled for feature selection. This model was applied (*celltypist.annotate*) to held-out data with and without majority voting enabled (CellTypist MV Sort-ref and CellTypist Sort-ref, respectively). HieRFit and Garnett models were both trained (*HieRFIT::CreateHieR*; *garnett::train_cell_classifier*) using a hierarchy structured identically to the MMoCHi hierarchy (Fig. 2b) and applied (*HieRFIT::HieRFIT*; *garnett::classify_cells*) to held-out data with defaults enabled (HieRFIT Sort-ref and Garnett Sort-ref, respectively). An additional Garnett model was trained using high-confidence thresholding (Garnett Markers) on manually and

automatically (*garnett::top_markers*) selected transcript markers which were evaluated using built-in functions (*garnett::check_markers*) (Supplementary Table 6). Garnett models were trained without marker propagation and prediction was performed without cluster extension. Precision, recall, F1 scores and overall accuracy were calculated using sort labels as truth (*sklearn.metrics.precision_recall_fscore_support*; *sklearn.metrics.accuracy_score*). Garnett provides unknown and intermediate cell type labels, which were excluded from performance metrics calculations.

Analysis of organ donor sequencing

CITE-seq data obtained from tissue immune cells was filtered separately. Although hashtag demultiplexing removes inter-sample multiplets, this method cannot detect multiplets between cells from the same tissue site. Thus, multiplets were detected by Scrublet[64] (*scrublet.Scrublet.scrub_doublets*), using default settings, an expected doublet rate of 0.015, and applying separately to each library-tissue combination of over 100 events. Cells were then filtered to remove events with fewer than 1000 unique counts, fewer than 600 genes detected, or over 10% hemoglobin counts. We used a cluster-based method to remove events either identified as doublets by Scrublet or with high mitochondrial counts, similar to the previously described percolation method[25]. Briefly, a PCA on highly variable genes was calculated, integrated using Harmony[65] (*scanpy.external.pp.harmony_integrate*), and used for nearest neighbor calculation and Leiden clustering with a resolution of 20. Clusters with significantly higher Scrublet scores (above 0.1) or percent mitochondrial counts (above 15%) were removed. Individual events not captured by these clusters that were identified as Scrublet doublets or had over 25% mitochondrial counts were also removed.

We devised a MMoCHi hierarchy (Fig. 3d; Supplementary Table 7) and performed classification using the algorithm above. Classification fit across a varying number of random forest estimators (trees) was evaluated for prediction accuracy. Overall prediction accuracy was measured using 20% held-out events that were high-confidence thresholded in all layers of the hierarchy. Prediction accuracy was also calculated for individual classification nodes using the weighted average of precision, recall, and F1 scores for internally held-out, high-confidence events at each node. Gini impurity-based feature importances were automatically calculated during random forest training by scikit-learn. Manual annotation and UMAP calculation were performed on the totalVI latent space, as described above. Similarity matrix of classified subsets was calculated (*scanpy.pl.correlation_matrix*) on the totalVI latent space with optimal ordering enabled.

Estimation of computational performance

We measured the computational resources required to train and apply MMoCHi classifiers using a predefined hierarchy and thresholds (Fig. 3c; Supplementary Table 7). Comparisons were performed with multiprocessing enabled for random forest training, using 3rd generation Intel Xeon Scalable processors (3.5 GHz) with 32 vCPUs and 32 GiB of RAM (AWS/EC2 c6i.8xlarge instance). Tests at varying event counts were performed by randomly subsampling the dataset prior to classification.

Statistical analysis

Prior to differential expression, the two groups were subsampled to the same number of events and their expression matrices were downsampled to equal total counts. Differential

expression was performed (*scanpy.tl.rank_genes_groups*) on log-normalized expression, with Wilcoxon mode and tie correction enabled.

DATA AVAILABILITY

CITE-seq data generated in this study are available in NCBI GEO with the accession code GSE229791.

CODE AVAILABILITY

Code for MMoCHi is available at: https://github.com/donnafarberlab/MMoCHi

Code for demultiplexing CITE-seq and hashtag barcodes is available at:

https://github.com/simslab/DropSeqPipeline8

REFERENCES

1. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
2. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).
3. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
4. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377–382 (2009).
5. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865–868 (2017).
6. Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* **35**, 936–939 (2017).
7. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* **20**, 273–282 (2019).
8. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
9. Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* **18**, 272–282 (2021).
10. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* **7**, 1141 (2020).
11. Pasquini, G., Rojo Arias, J. E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal* **19**, 961–969 (2021).

12. Ma, W., Su, K. & Wu, H. Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction. *Genome Biology* **22**, 264 (2021).
13. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 31 (2020).
14. Osumi-Sutherland, D. *et al.* Cell type ontologies of the Human Cell Atlas. *Nat Cell Biol* **23**, 1129–1135 (2021).
15. Pizzolato, G. *et al.* Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of human TCRVδ1 and TCRVδ2 γδ T lymphocytes. *Proceedings of the National Academy of Sciences* **116**, 11906–11915 (2019).
16. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* **16**, 983–986 (2019).
17. Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* **16**, 1007–1015 (2019).
18. Fang, P. *et al.* Immune cell subset differentiation and tissue inflammation. *J Hematol Oncol* **11**, 97 (2018).
19. Gray, J. I. & Farber, D. L. Tissue-Resident Immune Cells in Humans. *Annu. Rev. Immunol.* **40**, 195–220 (2022).
20. Szabo, P. A., Miron, M. & Farber, D. L. Location, location, location: Tissue resident memory T cells in mice and humans. *Sci Immunol* **4**, eaas9673 (2019).
21. Szabo, P. A. Axes of heterogeneity in human tissue-resident memory T cells. *Immunol Rev* 1–15 (2023) doi:10.1111/imr.13210.
22. Gattinoni, L. *et al.* A human memory T cell subset with stem cell–like properties. *Nat Med* **17**, 1290–1297 (2011).
23. Sallusto, F., Lenig, D., Förster, R., Lipp, M. & Lanzavecchia, A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature* **401**, 708–712 (1999).
24. Thome, J. J. C. *et al.* Spatial Map of Human T Cell Compartmentalization and Maintenance over Decades of Life. *Cell* **159**, 814–828 (2014).
25. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
26. Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nature Communications* **10**, 4706 (2019).
27. Poon, M. M. L. *et al.* Tissue adaptation and clonal segregation of human memory T cells in barrier sites. *Nat Immunol* **24**, 309–319 (2023).
28. Liu, X. *et al.* Knowledge-based classification of fine-grained immune cell types in single-cell RNA-Seq data. *Briefings in Bioinformatics* **22**, bbab039 (2021).
29. Lanier, L. L., Chang, C., Spits, H. & Phillips, J. H. Expression of cytoplasmic CD3 epsilon proteins in activated human adult natural killer (NK) cells and CD3 gamma, delta, epsilon complexes in fetal NK cells. Implications for the relationship of NK and T lymphocytes. *J Immunol* **149**, 1876–1880 (1992).
30. Pont, F. *et al.* The gene expression profile of phosphoantigen-specific human γδ T lymphocytes is a blend of αβ T-cell and NK-cell signatures. *European Journal of Immunology* **42**, 228–240 (2012).
31. Tosolini, M. *et al.* Assessment of tumor-infiltrating TCRVγ9Vδ2 γδ lymphocyte abundance by deconvolution of human cancers microarrays. *Oncoimmunology* **6**, e1284723 (2017).
32. Hahne, F. *et al.* Per-channel basis normalization methods for flow cytometry data. *Cytometry A* **77**, 121–131 (2010).
33. Zheng, Y., Jun, S.-H., Tian, Y., Florian, M. & Gottardo, R. Robust Normalization and Integration of Single-cell Protein Expression across CITE-seq Datasets. 2022.04.29.489989 Preprint at https://doi.org/10.1101/2022.04.29.489989 (2022).

34. Kaymaz, Y. *et al.* HieRFIT: a hierarchical cell type classification tool for projections from complex single-cell atlas datasets. *Bioinformatics* **37**, 4431–4436 (2021).
35. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
36. Chen, X., Wang, M. & Zhang, H. The use of classification trees for bioinformatics. *WIREs Data Mining and Knowledge Discovery* **1**, 55–63 (2011).
37. Baba, Y. *et al.* GATA2 Is a Critical Transactivator for the Human IL1RL1/ST2 Promoter in Mast Cells/Basophils. *Journal of Biological Chemistry* **287**, 32689–32696 (2012).
38. Evren, E. *et al.* Distinct developmental pathways from blood monocytes generate human lung macrophage diversity. *Immunity* **54**, 259-275.e7 (2021).
39. Karnell, J. L. *et al.* CD19 and CD32b Differentially Regulate Human B Cell Responsiveness. *J Immunol* **192**, 1480–1490 (2014).
40. Debnath, I., Roundy, K. M., Pioli, P. D., Weis, J. J. & Weis, J. H. Bone marrow-induced Mef2c deficiency delays B-cell development and alters the expression of key B-cell regulatory proteins. *International Immunology* **25**, 99–115 (2013).
41. Mason, D. Y. *et al.* CD79a: A Novel Marker for B-Cell Neoplasms in Routinely Processed Tissue Samples. *Blood* **86**, 1453–1459 (1995).
42. Xu, C. *et al.* MARCKS regulates tonic and chronic active B cell receptor signaling. *Leukemia* **33**, 710–729 (2019).
43. Rissoan, M.-C. *et al.* Subtractive hybridization reveals the expression of immunoglobulinlike transcript 7, Eph-B1, granzyme B, and 3 novel transcripts in human plasmacytoid dendritic cells. *Blood* **100**, 3295–3303 (2002).
44. Sjölin, H. *et al.* DAP12 Signaling Regulates Plasmacytoid Dendritic Cell Homeostasis and Down-Modulates Their Function during Viral Infection1. *The Journal of Immunology* **177**, 2908–2916 (2006).
45. Rutishauser, R. L. *et al.* Transcriptional Repressor Blimp-1 Promotes CD8+ T Cell Terminal Differentiation and Represses the Acquisition of Central Memory T Cell Properties. *Immunity* **31**, 296–308 (2009).
46. Redmond, D., Poran, A. & Elemento, O. Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome Medicine* **8**, 80 (2016).
47. Mimitou, E. P. *et al.* Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat Methods* **16**, 409–412 (2019).
48. Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* **10**, e63632 (2021).
49. Olsen, T. R. *et al.* Scalable co-sequencing of RNA and DNA from individual nuclei. 2023.02.09.527940 Preprint at https://doi.org/10.1101/2023.02.09.527940 (2023).
50. Andor, N. *et al.* Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution. *NAR Genom Bioinform* **2**, lqaa016 (2020).
51. Reimegård, J. *et al.* A combined approach for single-cell mRNA and intracellular protein expression analysis. *Commun Biol* **4**, 1–11 (2021).
52. Takei, Y. *et al.* Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590**, 344–350 (2021).
53. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol* **40**, 1794–1806 (2022).
54. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).
55. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. in *Proceedings of the 22nd International Conference on Machine Learning* 625–632 (ACM Press, 2005). doi:10.1145/1102351.1102430.

56. Carpenter, D. J. *et al.* Human immunology studies using organ donors: impact of clinical variations on immune parameters in tissues and circulation. *Am J Transplant* **18**, 74–88 (2018).
57. Kumar, B. V. *et al.* Human Tissue-Resident Memory T Cells Are Defined by Core Transcriptional and Functional Signatures in Lymphoid and Mucosal Sites. *Cell Reports* **20**, 2921–2934 (2017).
58. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).
59. Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* **39**, 813–818 (2021).
60. Melsted, P., Ntranos, V. & Pachter, L. The barcode, UMI, set format and BUStools. *Bioinformatics* **35**, 4472–4473 (2019).
61. Yuan, J. & Sims, P. A. An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. *Sci Rep* **6**, 33883 (2016).
62. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
63. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
64. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* **8**, 281-291.e9 (2019).
65. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).

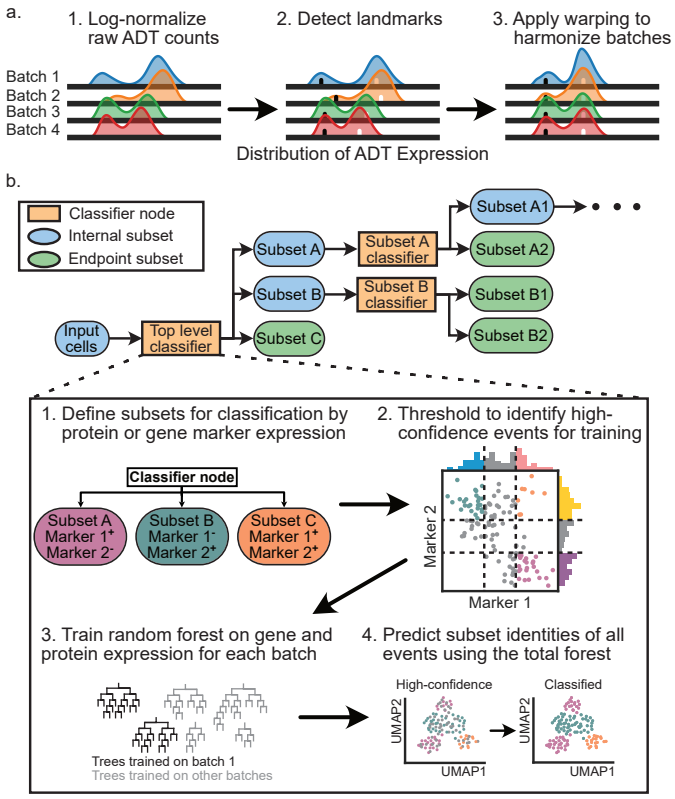


Figure 1 **Schematic showing general workflow of MMoCHi.** a. Batch correction of Antibody Derived Tag (ADT) expression was performed by landmark registration on log-normalized counts (1). After detection of positive (white ticks) and negative (black ticks) peaks in each batch (2), a warping function was applied to align these landmarks across batches (3). b. MMoCHi hierarchy demonstrating the classification workflow. User supplies a hierarchy of cell subsets, and marker definitions for each subset (1). Thresholding is performed to select high-confidence events for each subset (2). A portion of these high-confidence events are used to train a random forest (3; see Extended Data Fig 1). Finally, the trained random forest is used to predict subset identities of all events from the parent subset (4). This process is repeated for each classifier node within the hierarchy until all cells are classified to endpoint subsets.

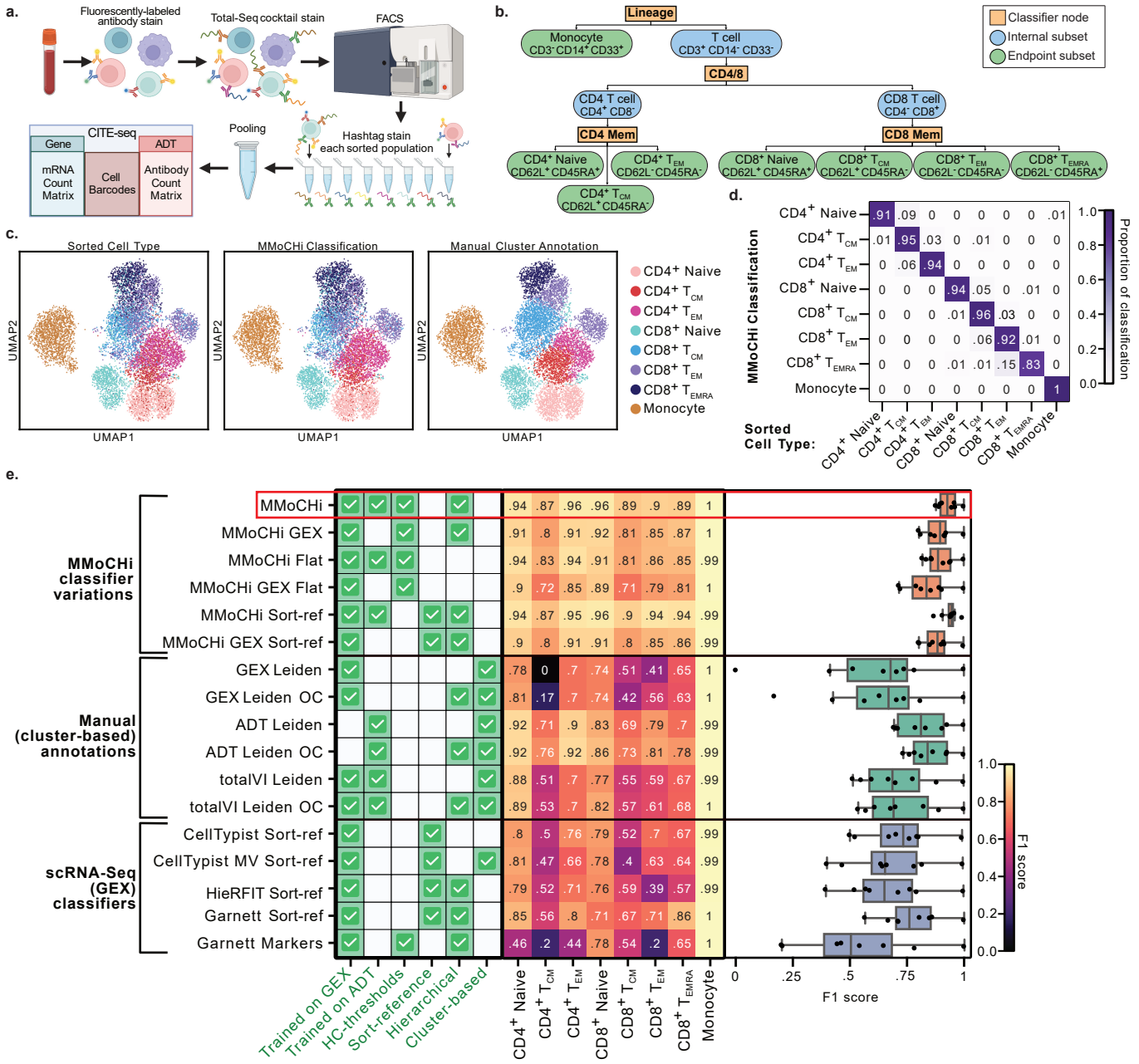


Figure 2 **MMoCHi classification of predefined T cell subsets outperforms other annotation methods.** a. PBMCs were stained with fluorescently-labeled antibody cocktail and a cocktail of oligo-tagged antibodies for sequencing. Seven T cell subsets (CD4+ Naive, CD4+ T_{CM}, CD4+ T_{EM}, CD8+ Naive, CD8+ T_{CM}, CD8+ T_{EM}, CD8+ T_{EMRA}) and monocytes were sorted, hashtagged, pooled, and sequenced. b. MMoCHi hierarchy defining subsets using the same markers used for sorting. c. UMAPs of totalVI latent space colored by sorted cell type (identified by hashtag oligo; HTO), MMoCHi classification, and manual cluster annotations. d. Row-normalized heatmap comparing MMoCHi classification to sorted cell type. Color represents proportion of cells in each MMoCHi classification from each sorted subset. e. Performance comparison using F1 scores, calculated for each cell subset using HTO-derived sorted cell type labels as truth. F1 scores for each method were aggregated in box and whisker plots. Features of each method are labeled using green checks for context: “Trained on GEX/ADT”—whether gene expression (GEX) or antibody derived tag (ADT) expression data were used for model training. "HC-thresholds"—whether high-confidence thresholding was used for training data selection. "Sort-reference"—whether sorted cell type labels were used for training, and accuracy was measured on a 20% hold-out dataset. "Hierarchical"—whether annotation was performed on multiple levels. “Cluster-based”—whether annotations were applied to unsupervised clusters. "OC"—whether over-clustering was performed. MV, Majority voting; T_{CM}, central memory T cell; T_{EM}, effector memory T cell; T_{EMRA}, terminally differentiated effector memory T cell. Schematic in (a) created with BioRender.com

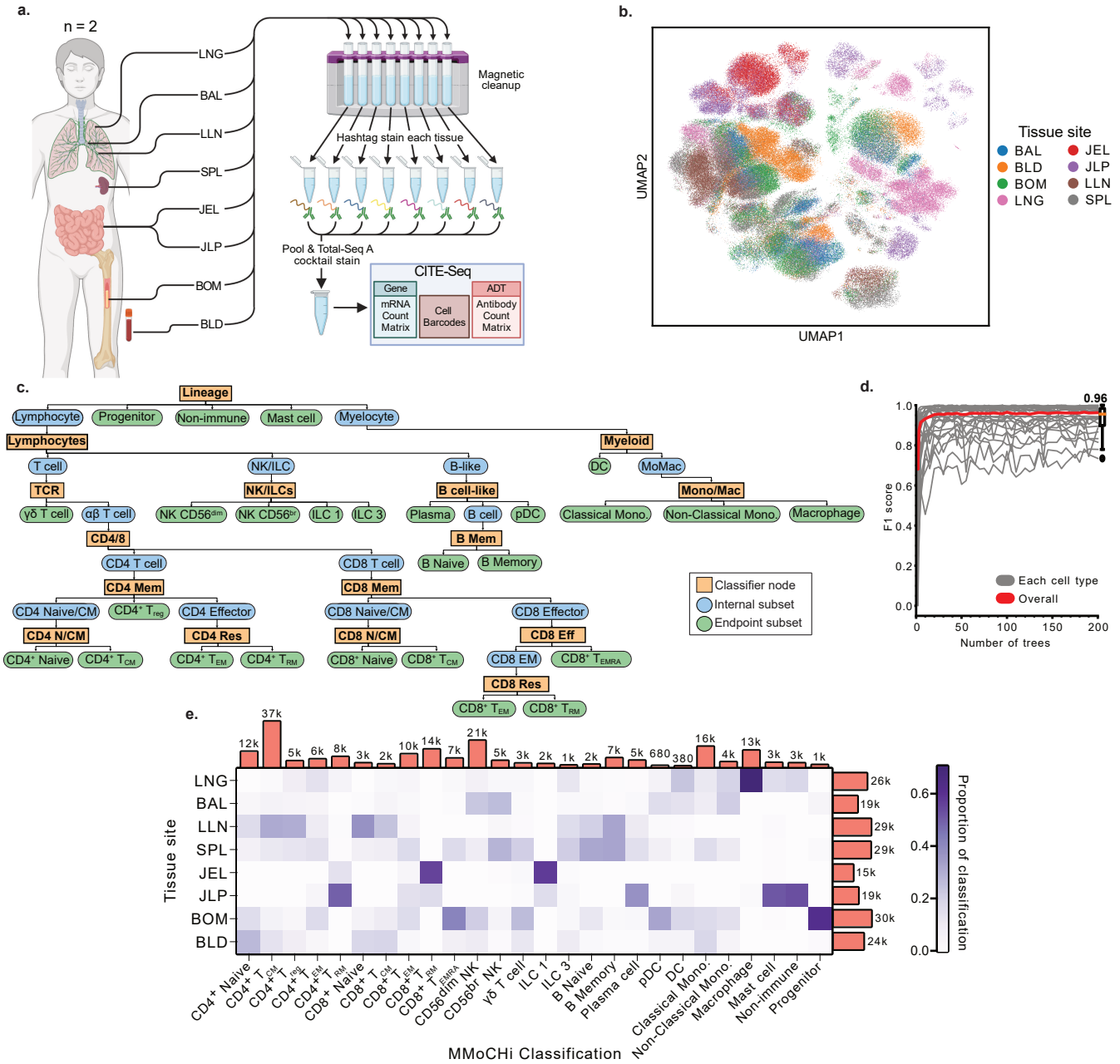


Figure 3 **MMoCHi classification applied to human immune cells from blood and diverse tissue sites.** a. Single cell suspensions of immune cells were isolated from blood and indicated tissue sites of two organ donors. CD45+ immune cells selected for by magnetic enrichment, hashtagged by tissue site, pooled, and sequenced. b. UMAP of donor-integrated totalVI latent space, colored by tissue site. c. MMoCHi hierarchy used for classification (See Supplementary Table 7 for full specification). d. Performance curves showing the F1 score for each endpoint subset when training random forests with various numbers of trees (estimators). F1 scores were calculated on held-out data using high-confidence thresholded events as truth. The red line indicates the median F1 score across all predicted cell types. e. Column-normalized heatmap depicting the distribution of classified cell types across tissue sites. The number of total events in each classification or tissue site are displayed.

LNG, lung; BAL, bronchoalveolar lavage; LLN, lung lymph node; SPL, spleen; JEL, jejunum epithelial layer; JLP, jejunum lamina propria; BOM, bone marrow; BLD, blood; T_{CM}, central memory T cell; Treg, regulatory T cell; T_{EM}, effector memory T cell; T_{RM}, resident memory T cell; T_{EMRA}, terminally differentiated effector memory T cell; NK, natural killer cell; ILC, innate lymphoid cell; DC, dendritic cell; pDC, plasmacytoid dendritic cell; Mono, monocyte. Schematic created with BioRender.com

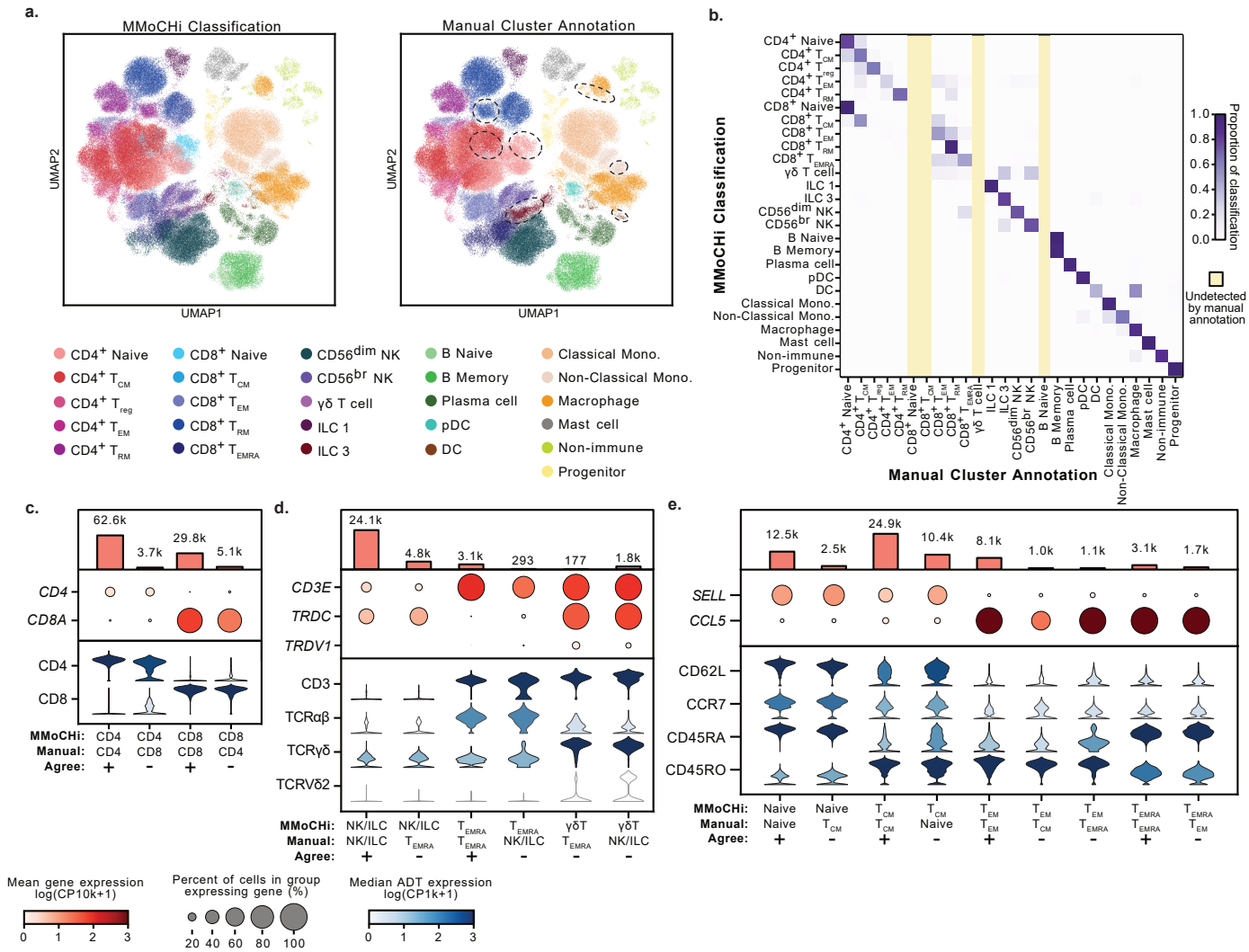


Figure 4 **MMoCHi classification of human immune cells outperforms manual annotation by clustering.** a. UMAPs of immune cells from Fig. 3, colored by MMoCHi classification and manual cluster annotations. Dashed ellipses highlight areas of major disagreement between MMoCHi classification and manual annotation. b. Row-normalized heatmap comparing MMoCHi classification to manual annotation. Color represents proportion of cells in each MMoCHi classification that were manually annotated as each subset. Yellow columns indicate subsets that were not detected by manual annotation. c-e. Plots depicting expression of selected cell type markers, on cells grouped by their MMoCHi classification and manual annotation. Dot plots display gene expression (GEX). Dot size represents the percent of cells in the group expressing a gene, and dots are colored by the mean log-normalized GEX counts per ten thousand. Violin plots display the distribution of antibody derived tag (ADT) expression for each population. Violins are colored by the median log-normalized ADT counts per thousand. The number of events in each grouping are displayed above the dot plots. Events are denoted by a "+" where MMoCHi and manual annotation agree, and a "-" where they disagree.
T_{CM}, central memory T cell; T_{reg}, regulatory T cell; T_{EM}, effector memory T cell; T_{RM}, resident memory T cell; T_{EMRA}, terminally differentiated effector memory T cell; ILC, innate lymphoid cell; NK, natural killer cell; pDC, plasmacytoid dendritic cell; DC, dendritic cell; Mono, monocyte

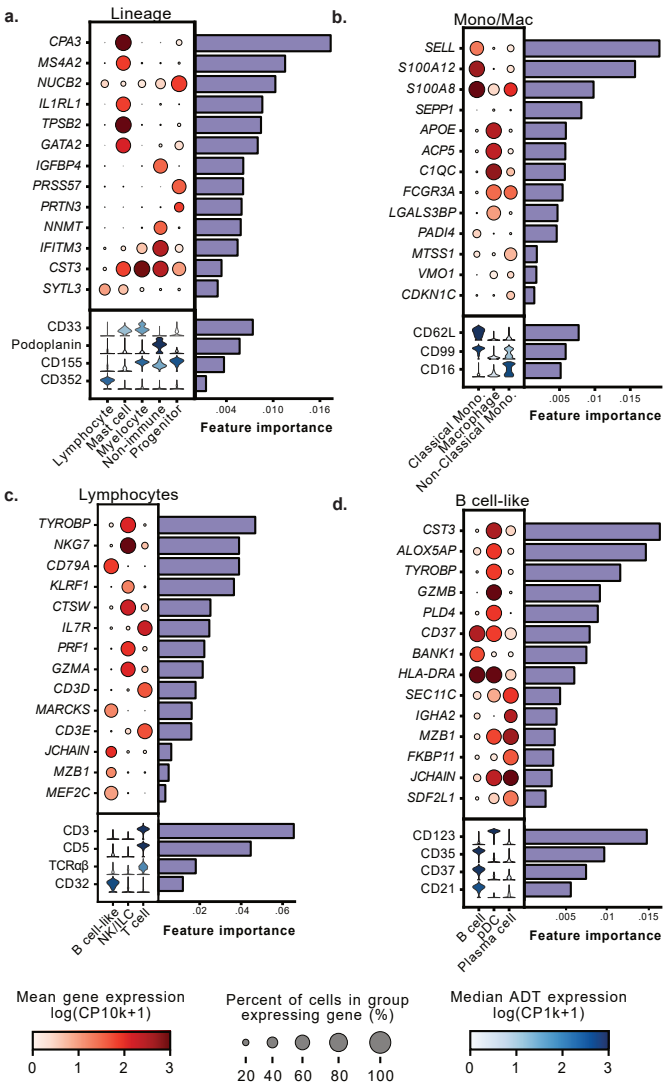


Figure 5 **Interpretation of MMoCHi random forests using feature importances reveal immune cell lineage markers.** a-d. Expression of important features for classification for the Lineage (a), Mono/Mac (b), Lymphocytes (c), and B cell-like (d) classification nodes. The top important features associated with each subset ($\log_2$(fold change) > 2) are displayed to include representation of features specific to each subset. All features displayed are within the top 1% of important features. Dot plots display gene expression (GEX), where dot size represents the percent of cells in the group expressing a gene, and dots are colored by the mean log-normalized GEX counts per ten thousand. Violin plots display the distribution of antibody derived tag (ADT) expression and are colored by the median log-normalized ADT counts per thousand. The impurity-based importance of each feature displayed is shown in a bar chart to the right. All features displayed were also significantly differentially expressed ($p < 0.05$). Statistical significance was calculated using a two-sided Wilcoxon with tie correction, followed by a Benjamini–Hochberg adjustment for multiple comparisons.
Mono, monocyte; NK, natural killer cell; ILC, innate lymphoid cell; pDC, plasmacytoid dendritic cell

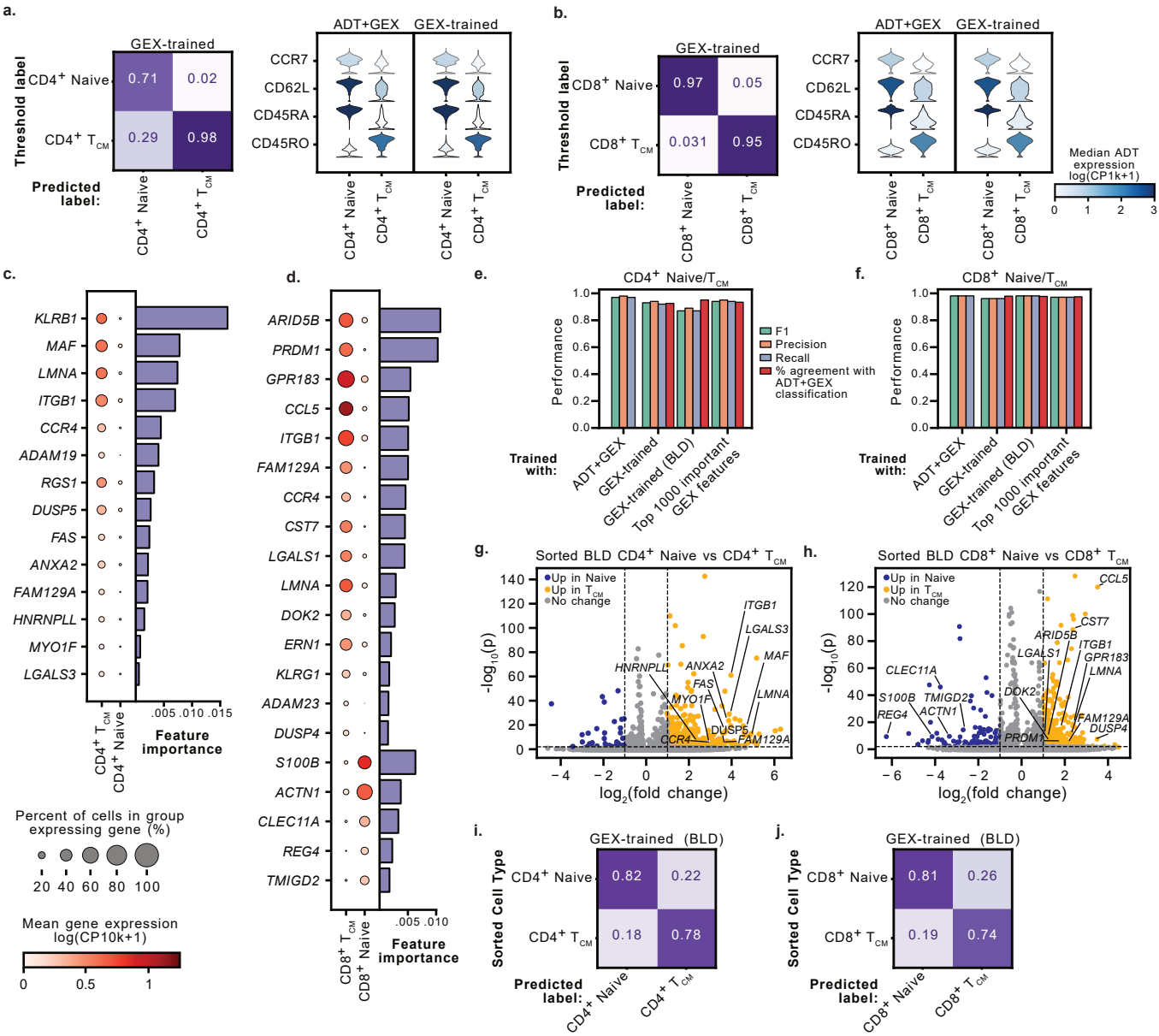
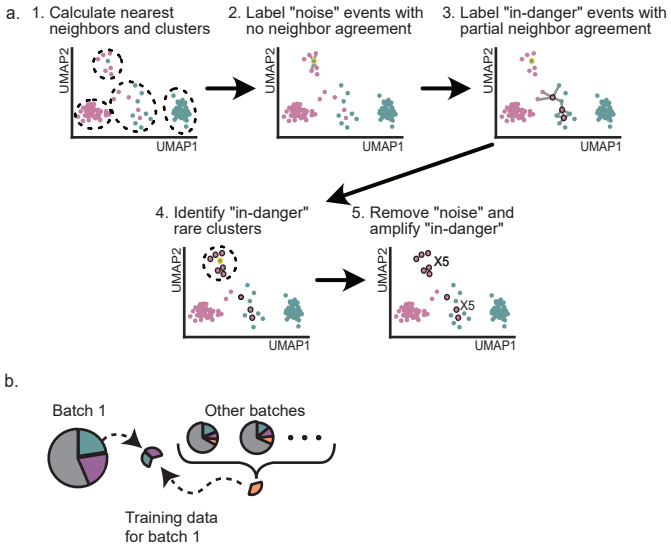
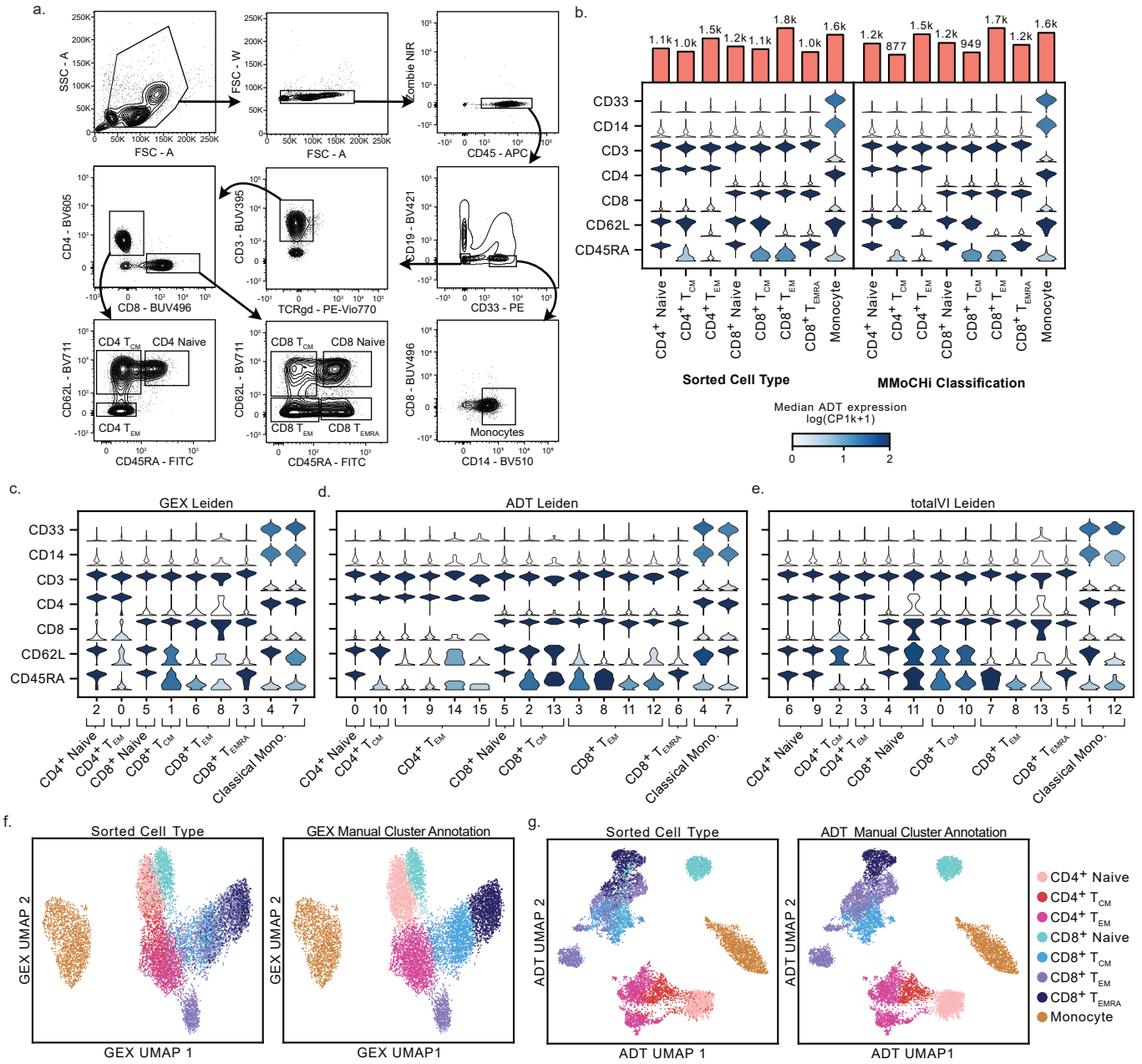


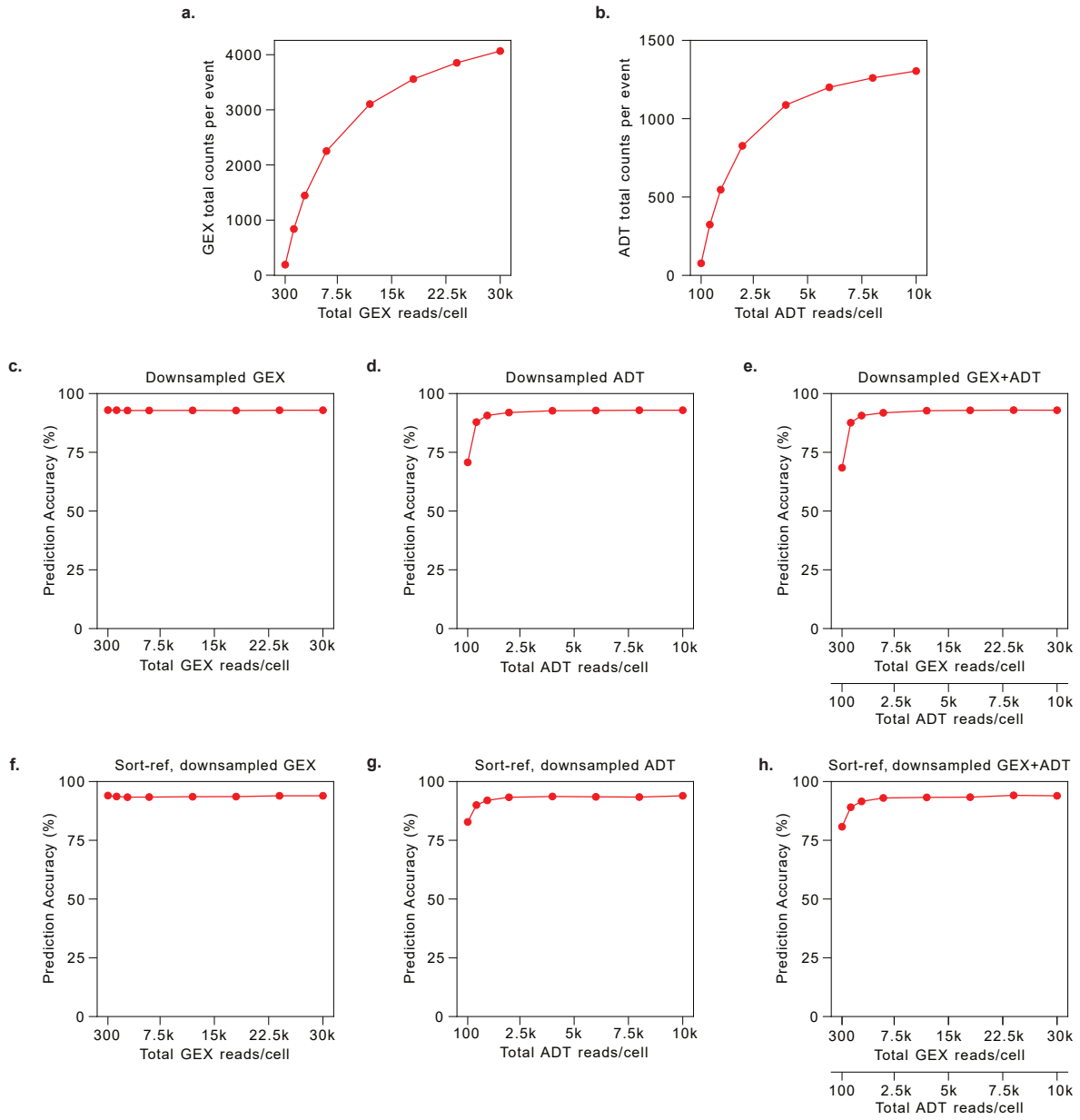
Figure 6 **MMoCHi reveals additional transcript markers distinguishing naive and central memory T cells.** a-b. Performance of gene expression (GEX)-only classification of Naive/T_{CM} $CD4^+$ (a) and $CD8^+$ (b) T cells. Column-normalized confusion matrices (left) comparing classification to held-out, high-confidence threshold labels. Violin plots (right) display the distribution of antibody derived tag (ADT) expression for each population as classified by MMoCHi trained on multimodal data (ADT+GEX) or GEX-only. Violins are colored by the median log-normalized ADT counts per thousand. c-d. Dot plots displaying expression of important features for GEX-only classification of Naive/T_{CM} $CD4^+$ (c) or $CD8^+$ (d) T cells. Dot size represents the percent of cells in the group expressing a gene, and dots are colored by the mean log-normalized GEX counts per ten thousand. The top important features associated with a single subset (log_2(fold change) > 2 and greater than 10% change in dropout rate) are displayed. All features displayed were within the top 1000 important features when training with GEX-only on all tissues, and with blood (BLD) only. The importance of each feature is shown in a bar chart. All displayed features were significantly differentially expressed ($p < 0.05$). e-f. Bar plots comparing performance of Naive/T_{CM} $CD4^+$ (e) or $CD8^+$ (f) T cell classifiers trained using multimodal data (ADT+GEX), GEX only, GEX with only BLD cells, or only the top 1000 important GEX features. g-h. Volcano plots displaying differential gene expression between sorted blood Naive and T_{CM} $CD4^+$ (g) or $CD8^+$ (h) T cells. Points represent genes and are colored by differential expression ($|log_2$(fold change)$| > 1$ and $P < 0.05$). Important features for classification that were differentially expressed are highlighted. f-j. Column-normalized heatmaps displaying performance of GEX only classifiers trained on organ donor BLD applied to sorted Naive/T_{CM} $CD4^+$ (i) and $CD8^+$ (h) T cells. Statistical significance was calculated using a two-sided Wilcoxon with tie correction, followed by a Benjamini–Hochberg adjustment for multiple comparisons.



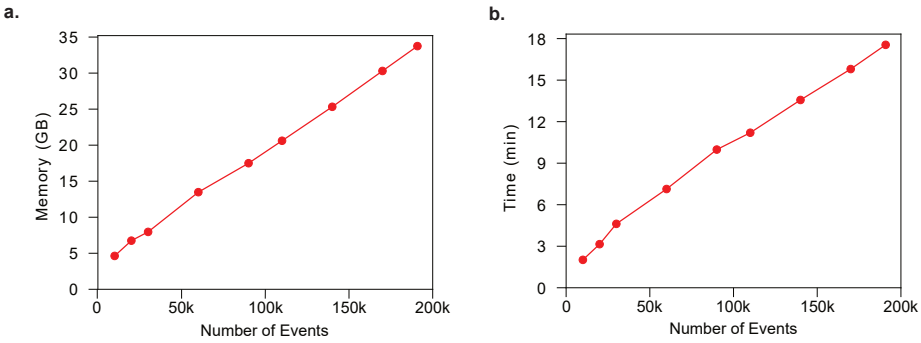
Extended Data Figure 1 **Selection and cleanup of training data.** a. Prior to training, high-confidence events were resampled to remove potentially mislabeled events ("noise") and amplify events potentially challenging to classify ("in danger"). Nearest neighbors are calculated on high-confidence events (1). Events with no neighbors in agreement with their high confidence label are identified as "noise" (2). Events with only some neighbors in agreement or in poorly represented clusters are identified as "in danger" (3,4). Events labeled "noise" are removed from the training data. Events labeled "in danger" are duplicated 5 times in the training dataset (5). b. Training data is selected from a random sample of cleaned-up high-confidence calls with rare subsets oversampled so all subsets are equal in the training data. In the case of insufficient training events in individual batches, high-confidence events are spiked into the training data from other batches.



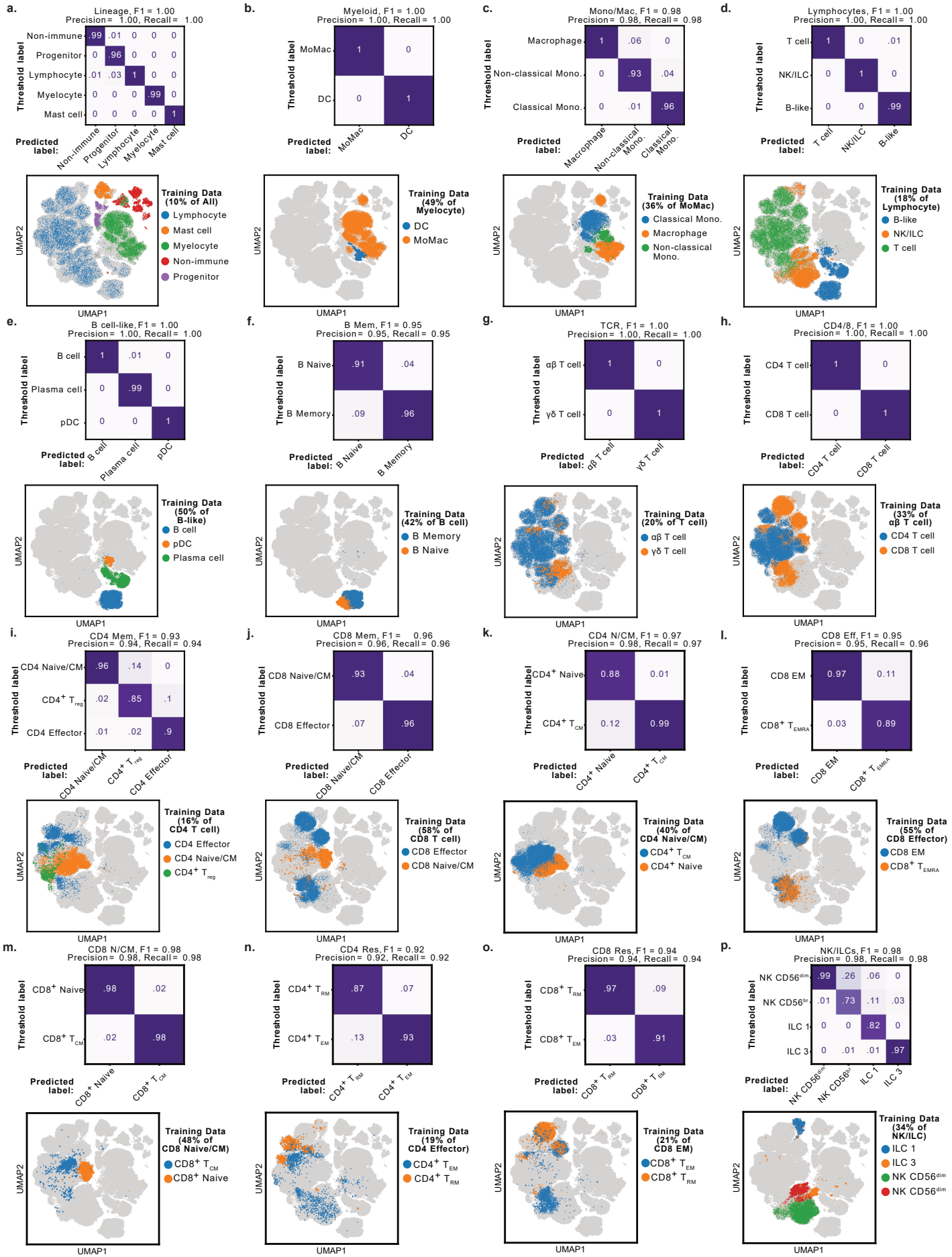
Extended Data Figure 2 **Generation and analysis of MMoCHi validation dataset.** a. Gating strategy for fluorescence-activated cell sorting (FACS) of T cell memory subsets and monocytes. b. Violin plots displaying the distribution of antibody derived tag (ADT) expression for select markers for each sorted population, and each MMoCHi classification. Violins are colored by the median log-normalized ADT counts per thousand. The number of events in each grouping are displayed above the violin plots. c-e. Violin plots displaying distribution of ADT expression for select markers for Leiden clusters of gene expression (GEX; c), ADT expression (d), or totalVI latent space (e). Clusters are grouped by their manual annotation. f-g. UMAPs of principal component analysis used for GEX annotation (f), or ADT annotation (g), colored by sorted cell type (identified by hashtag oligo; HTO) or manual annotation.
T_{CM}, central memory T cell; T_{EM}, effector memory T cell; T_{EMRA}, terminally differentiated effector memory T cell



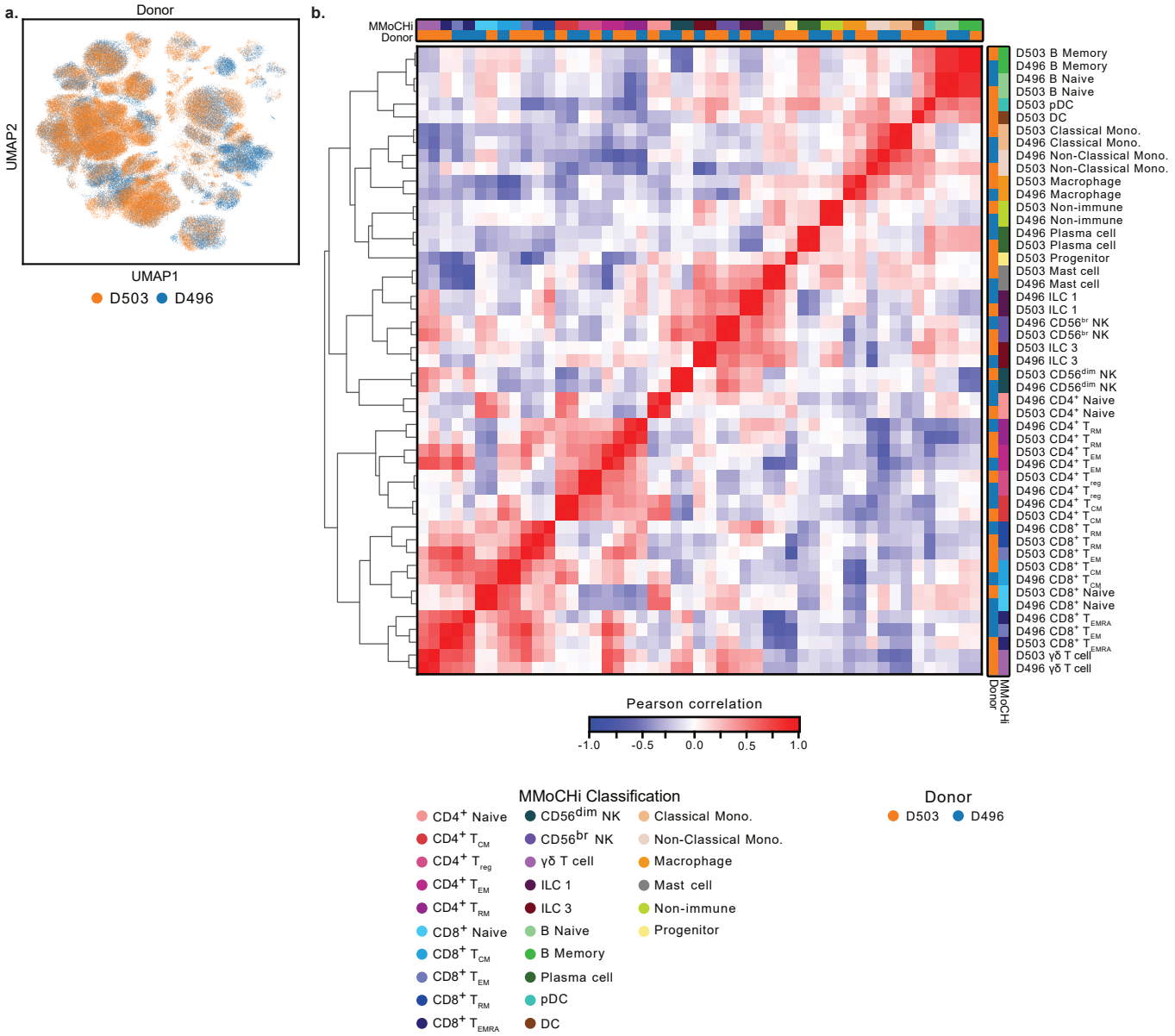
Extended Data Figure 3 **MMoCHi classification is robust to downsampled reads.** a-b. Line plots depicting the effect of downsampling reads of gene expression (GEX; a) or antibody derived tag (ADT) expression b) on total counts per event. c-h. Classification accuracy of sorted T cell memory subsets and monocytes, with MMoCHi classifiers trained using high-confidence thresholding on downsampled events (c-e) or trained using the hashtag oligo (HTO)-derived sort labels as reference, and tested on 20% held-out data (f-g). Classifiers were exposed to downsampled GEX and normal ADT expression (c,f), normal GEX and downsampled ADT expression (d,g), or downsampled GEX and ADT expression (e,h) for both training and prediction.



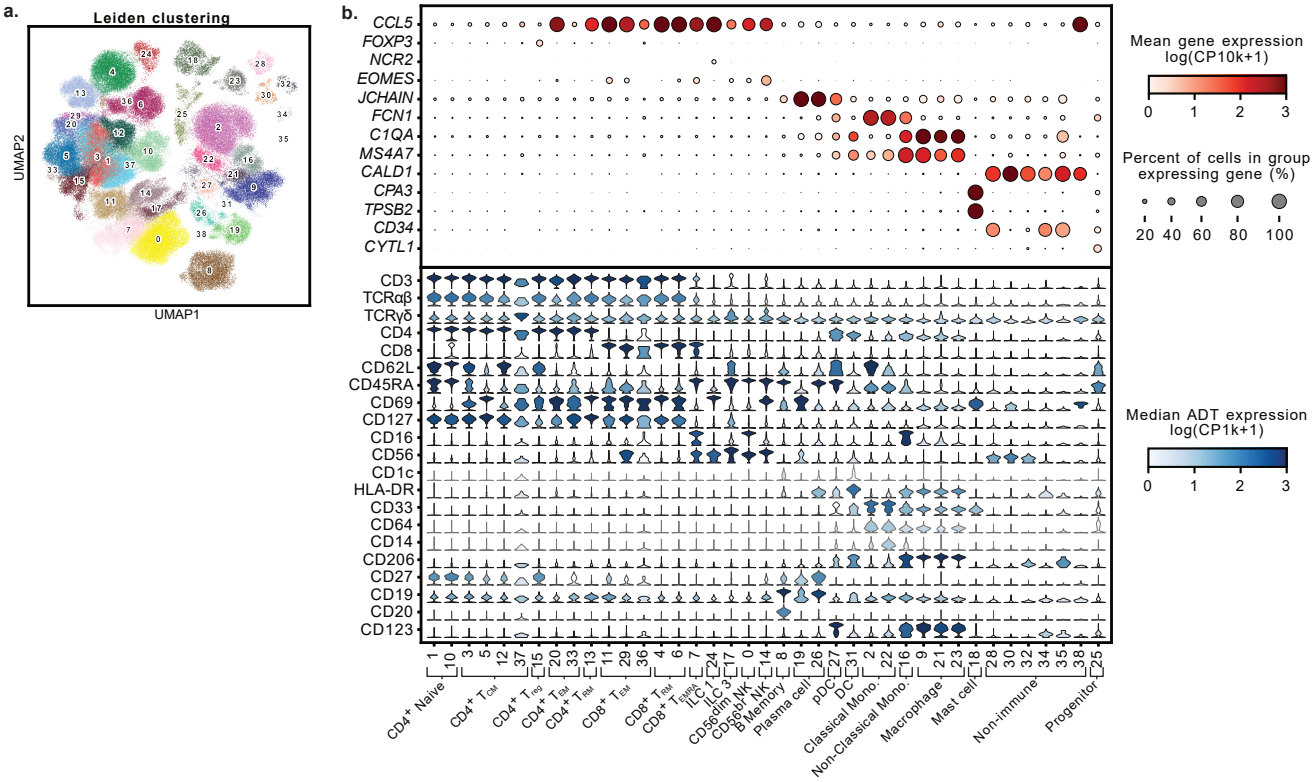
Extended Data Figure 4 **Time and memory performance of MMoCHi classification.** a-b. Memory usage (a) and run time (b) to classify 26 subsets using CITE-seq of two human organ donors across multiple tissues. Tests were performed on predefined hierarchy and thresholds with multiprocessing enabled for random forest training on a computer with 3rd Generation Intel Xeon Scalable processors (3.5 GHz), with 32 vCPUs and 32 GiB of RAM. Tests at different event counts were performed with random subsampling prior to classification.



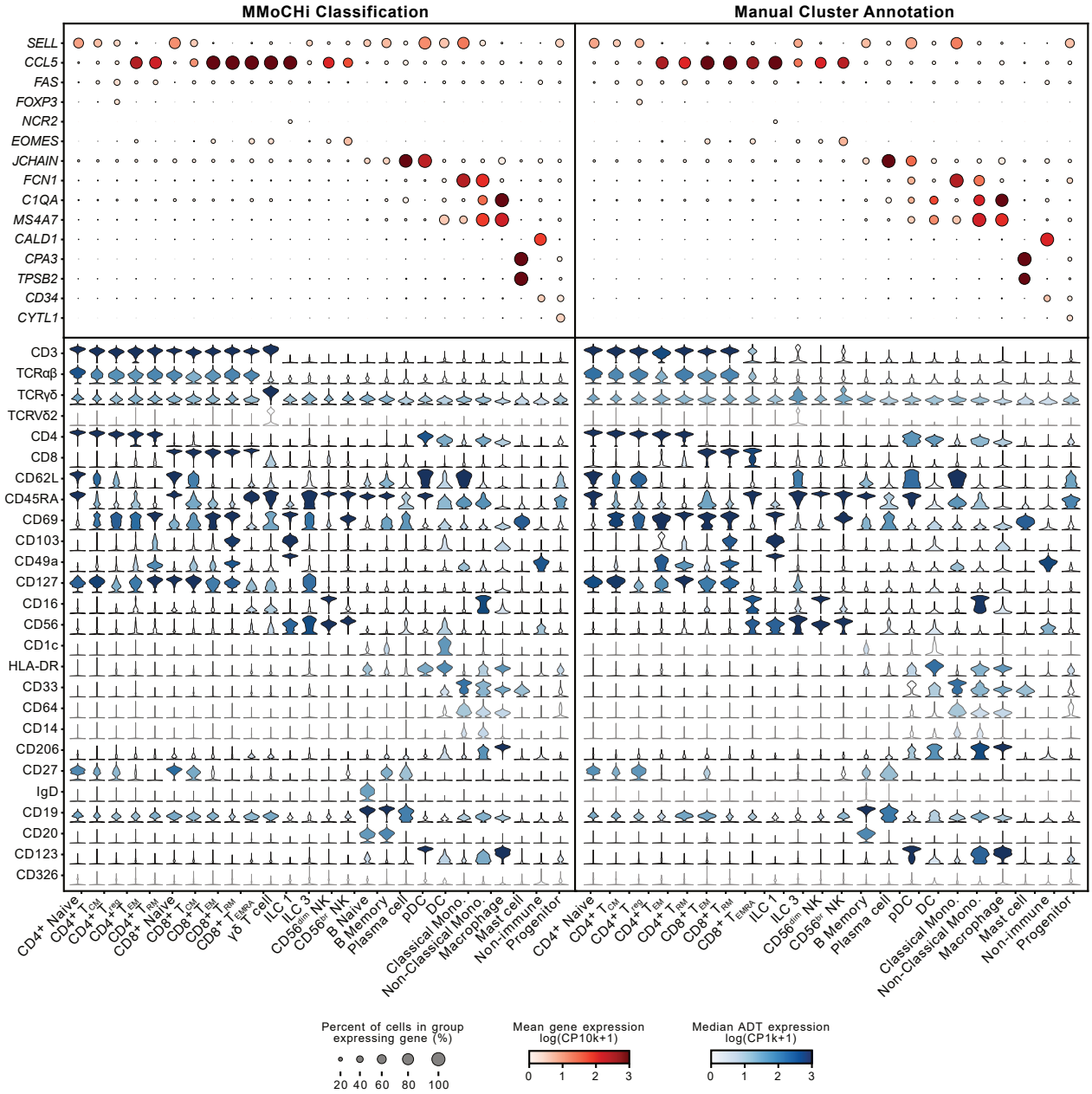
Extended Data Figure 5 **Random forests perform well at each level of MMoCHi hierarchy.** a-p. Training data and performance of each classification node (Fig. 3D). Column-normalized confusion matrices depicting performance by comparing the classification to held-out high-confidence threshold labels at each node. UMAPs of totalVI latent space, where events used for training are colored by their high-confidence threshold label. Dot size is proportional to the percent of the training dataset they represented after resampling of training data (see Extended Data Fig. 1).
T_{CM}, central memory T cell; T_{reg}, regulatory T cell; T_{EM}, effector memory T cell; T_{RM}, resident memory T cell; T_{EMRA}, terminally differentiated effector memory T cell; ILC, innate lymphoid cell; NK, natural killer cell; pDC, plasmacytoid dendritic cell; DC, dendritic cell; Mono, monocyte



Extended Data Figure 6 **MMoCHi classifies cell types across donor.** a. UMAP of donor-integrated totalVI latent space, colored by donor. b. Similarity matrix of each classified cell type for each donor with over 100 events, calculated using Pearson correlation on the totalVI latent space. Outer bars (top, right) colored by MMoCHi classification (outer) and donor (inner). Matrix is colored by Pearson correlation.
LNG, lung; BAL, bronchoalveolar lavage; LLN, lung lymph node; SPL, spleen; JEL, jejunum epithelial layer; JLP, jejunum lamina propria; BOM, bone marrow; BLD, blood; T_{CM}, central memory T cell; T_{reg}, regulatory T cell; T_{EM}, effector memory T cell; T_{RM}, resident memory T cell; T_{EMRA}, terminally differentiated effector memory T cell, NK, natural killer cell; ILC, innate lymphoid cell; DC, dendritic cell; pDC, plasmacytoid dendritic cell; Mono, monocyte.



Extended Data Figure 7 **Manual annotation of Leiden clusters by marker gene and protein expression.** a. UMAP of totalVI latent space displaying Leiden clusters. b. Expression of select markers for each cluster, grouped by manually annotated cell type. Dot plots display gene expression (GEX). Dot size represents the percent of cells in the group expressing a gene, and dots are colored by the mean log-normalized GEX counts per ten thousand. Violin plots display the distribution of antibody derived tag (ADT) expression for each sorted population, and each MMoCHi classification. Violins are colored by the median log-normalized ADT counts per thousand.



Extended Data Figure 8 **Marker gene and protein expression on manually annotated and MMoCHi classified populations.** Expression of select cell type markers on cells grouped by either their MMoCHi classification (left) or their manual cluster annotation (right). Dot plots displaying gene expression (GEX), Dot size represents the percent of cells in the group expressing a gene, and dots are colored by the mean log-normalized GEX counts per ten thousand. Violin plots display the distribution of antibody derived tag (ADT) expression for each sorted population, and each MMoCHi classification. Violins are colored by the median log-normalized ADT counts per thousand.

T_{CM}, central memory T cell; T_{reg}, regulatory T cell; T_{EM}, effector memory T cell; T_{RM}, resident memory T cell; T_{EMRA}, terminally differentiated effector memory T cell, NK, natural killer cell; ILC, innate lymphoid cell; pDC, plasmacytoid dendritic cell