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], Peter A. Sims[2,4]*

*Correspondence to: pas2182@cumc.columbia.edu

[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

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 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 annotation of cell types and developmental states across diverse lineages, samples, or modalities.

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], RNA Expression And Protein sequencing assay (REAP-seq)[6], and Antibody sequencing (Ab-seq)[7] 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.

Downstream of raw data processing, the first and most important basic analytical step for single-cell sequencing is classification of individual cells, often in terms of canonical subsets. The vast majority of analytical tools, from differential expression analysis to trajectory inference to statistical modeling with clinical co-variates, depend on this crucial initial step. Many analytical tools have been developed for the cellular annotation of scRNA-seq data (reviewed in Ref.[8] and Ref.[9]). Unsupervised clustering is commonly used to divide cells into groups with similar expression profiles[10,11]. This approach has proven invaluable for characterizing cellular heterogeneity[12], and has been adapted for CITE-seq datasets[13,14]. However, the number, type, and identity of clusters can be difficult to compare across studies[12,15].

On the other hand, supervised machine learning methods can address these drawbacks by incorporating reference data or marker-based definitions of cell types[16]. Examples of such tools include CellTypist[17], ImmClassifier[18], HieRFIT[19], and Garnett[20]. CellTypist uses logistic regression and stochastic gradient descent for automated annotation based on a reference atlas[17].

To overcome the many, sometimes overlapping, annotations across atlases, it includes an option called "majority voting" where users can elect to perform conventional clustering and annotate all cells in each cluster with the most dominant classification. ImmClassifier addresses the reverse problem—existing reference datasets for cell types of interest (i.e. for immune cells) may not be pre-annotated with the desired granularity or may not include all the necessary cell subsets[18]. Thus, it first annotates input data with multiple random forest classifiers trained on independent reference datasets and then generates a consensus classification using a deep neural network. To improve annotation accuracy of highly similar subsets, HieRFIT utilizes a hierarchical design, implementing multiple random forest classifiers to progressively annotate cell types[19]. Together these reference-based methods enable cross-study comparisons[16,21], but require reference atlases which are not available for all tissues and contexts[22–24]. Reference-free annotation methods address these shortcomings, instead relying on knowledge-based marker definitions[20,25]. Garnett, for example, classifies scRNA-seq data using a hierarchy of cell types defined by user-supplied markers and a regularized elastic net generalized linear model for training[20]. Overall, existing tools have successfully leveraged a variety of machine learning algorithms, both reference-based and marker-based cell type definitions, and hierarchical designs for cell type classification, all with exclusively RNA-level 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 marker-based, reference-free classification. 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

59 developmental and activation states[26]. Decades of research have delineated numerous immune cell
60 subsets across the lymphocyte and myeloid lineages, often defined by surface phenotypes that
61 shape our biological questions[26–29] and are challenging to annotate by transcriptome alone[18,30–33].
62 Thus, to benchmark this tool against other annotation methods, we sorted and profiled T cell
63 subsets using CITE-seq and demonstrate improved performance by MMoCHi over the alternative
64 methods described above, particularly in the annotation of subsets with highly similar expression
65 profiles. We applied MMoCHi to produce integrated annotations for a cross-tissue CITE-seq atlas
66 of diverse immune cell populations isolated from complex samples. Extracting important features
67 from MMoCHi's underlying random forests revealed highly interpretable learned representations
68 of cell types, which we used to identify novel markers for distinguishing transcriptionally similar
69 T cell subsets. Finally, we demonstrated the versatility of MMoCHi by extending to applications
70 beyond CITE-seq. This includes identification of transformed tumor cells from joint profiling of
71 aneuploidy and gene expression and annotation of *in situ* spatial profiling (10x Genomics Xenium)
72 by joint classification with gene expression and morphological features. Together, MMoCHi
73 readily enables multimodal cell type annotation based on marker genes and proteins, is designed
74 for applicability to CITE-seq of any cell lineage or sample type and shows promise for extension
75 to forthcoming modalities such as *in situ* spatial profiling of complex tissues.
76

RESULTS

Algorithm Overview

77 The MMoCHi algorithm uses a hierarchy of random forest classifiers trained on gene expression
78 (GEX) and antibody-derived tags (ADTs) for cell type classification of CITE-seq data (Fig. 1).
79 Prior to classification, ADT expression is batch-corrected using landmark registration, as

80 previously applied to flow cytometry[34] and CITE-seq[35] (see Methods). We identify populations
81 exhibiting negative (background) and positive ADT expression and apply warping functions to
82 align their midpoints (landmarks) across batches, to effectively integrate CITE-seq expression
83 (Fig. 1a). MMoCHi then classifies cell types based on a user-supplied hierarchy of cell subsets
84 paired with marker-based definitions (Fig. 1b). At each classification node, high-confidence
85 members of each subset are identified using manual thresholds on user-provided gene and protein
86 markers. A random forest is then trained on a representative set of these high-confidence cells and
87 used to annotate all cells—including those not labeled by high-confidence thresholding (Fig. 1b;
88 Extended Data Fig. 1). Once trained, classifiers can be interrogated for features important for cell
89 classification or applied to extend cell type annotation to other datasets.

Superior annotation of closely related immune cell subsets by MMoCHi

90 MMoCHi is designed to improve the annotation of subsets with highly similar transcriptomic
91 profiles compared to conventional manual annotation of clusters or methods that rely exclusively
92 on mRNA profiles. We chose to test its performance using T cell subsets, which are well-defined
93 by surface marker expression and functional readouts[27,28], but challenging to annotate by scRNA-
94 seq alone. For example, $CD4^+$ and $CD8^+$ T cells are often imperfectly resolved, partially due to
95 low *CD4* transcript expression[18,19,31]. Moreover, conventional human T cells are delineated into
96 subsets based on differentiation state and migration capacity into naive T cells ($CCR7^+$ $CD45RA^+$
97 $CD45RO^-$) and memory T cells, which comprise central memory (T_{CM}; $CCR7^+$ $CD45RA^-$
98 $CD45RO^+$), effector memory (T_{EM}; $CCR7^-$ $CD45RA^-$ $CD45RO^+$), and terminally differentiated
99 (T_{EMRA}; $CCR7^-$ $CD45RA^+$ $CD45RO^-$) subsets[27,28]. Distinguishing naive T cells from T_{CM} and T_{EM}
100 from T_{EMRA} by transcriptome alone is challenging[1,18,30,31]. We sorted and performed CITE-seq on
101 seven T cell subsets ($CD4^+$ naive, $CD4^+$ T_{CM}, $CD4^+$ T_{EM}, $CD8^+$ naive, $CD8^+$ T_{CM}, $CD8^+$ T_{EM}, and

102 $CD8^+$ T_{EMRA}) and monocytes (Fig. 2a; Extended Data Fig. 2a; Supplementary Tables 1, 2). CCR7
103 staining by CITE-seq is suboptimal[5], so we used CD62L which has high concordance in human
104 blood[27,28]. To eliminate batch effects, we performed CITE-seq antibody staining before sorting
105 and labeled sorted populations with hashtag antibodies prior to pooling all samples for library
106 preparation and sequencing. Hashtagged populations then served as known references to evaluate
107 classification.

108 All sorts were of high purity (>91%; Supplementary Table 3), and subsets reflected
109 expected ADT marker expression (Extended Data Fig. 2b). We devised and applied a MMoCHi
110 hierarchy with the same protein markers used for sorting (Fig. 2b, c; Supplementary Table 4). For
111 comparison to unsupervised approaches, we used totalVI[14] to calculate a multimodal latent space,
112 then performed clustering and manual annotation (Fig. 2c). The high concordance between sorted
113 cell type and MMoCHi classification compared to manual annotation is visually evident from
114 UMAPs in Fig. 2c. Indeed, MMoCHi classification had 92% agreement with sorted labels, while
115 manual annotation had only 70% agreement (Fig. 2c,d; Supplementary Table 5). MMoCHi
116 classified populations showed expected ADT expression (Extended Data Fig. 2b). MMoCHi
117 classification remained accurate when downsampling GEX and/or ADT reads, or when
118 subsampling the number of monocytes or $CD8^+$ T_{CM} used for training revealing robustness to lower
119 data quality or fewer, imbalanced, training events (Extended Data Fig. 3). Although this
120 classification uses both ADT and GEX features, MMoCHi was likely insensitive to GEX coverage
121 because both the reference dataset and hierarchy for classification were defined by protein
122 expression.

123 To compare performance between various methods, we calculated precision, recall and F1
124 score for each subset, as well as overall accuracy (Fig. 2e, Supplementary Table 5). MMoCHi

125 accurately classified all subsets, with an average F1 score of 0.93 (Fig. 2e: MMoCHi). We tested
126 variations to validate MMoCHi's design. Performance worsened when training random forests
127 using only GEX and/or when classifying all subsets with only one random forest (instead of a
128 hierarchy) (Fig. 2e: MMoCHi GEX, MMoCHi Flat, and MMoCHi GEX Flat). We noted a modest
129 decrease in performance when relying exclusively on automatic thresholds (Supplementary Fig.
130 1a,c), and thus encourage users to manually evaluate and adjust automatic thresholds (similar to
131 best-practices for flow cytometry gating[36]). For Fig. 2, we applied MMoCHi using manually
132 defined hyperparameters and the in-danger-noise algorithm enabled (see Methods), but we noted
133 similar performance with MMoCHi's built-in automatic hyperparameter optimization or in-
134 danger-noise disabled (Supplementary Fig. 1b-c). Overall, these results support MMoCHi's
135 multimodal, hierarchical, and semi-automated design.

136 We next compared MMoCHi classification to manual annotation of unsupervised clusters
137 derived from GEX, ADT expression, or the totalVI latent space (see Methods; Fig. 2e: GEX
138 Leiden, ADT Leiden, totalVI Leiden; Extended Data Fig. 2c-g). MMoCHi outperformed all
139 manual annotation methods, particularly GEX Leiden, which failed to identify $CD4^+$ T_{CM} or $CD8^+$
140 Naive T cells (Extended Data Fig. 2f). To ensure manual annotation was not hampered by lack of
141 clustering resolution, we also over-clustered the data (see Methods), but this resulted in only a
142 slight improvement (Fig. 2e: GEX Leiden OC, ADT Leiden OC, totalVI Leiden OC). Furthermore,
143 since probabilistic generative modeling on small datasets can be sensitive to feature numbers[37], we
144 also showed that manual annotation using totalVI was insensitive to the number of input genes
145 (Extended Data Fig. 4a-d).

146 Lastly, we tested four supervised classification methods for scRNA-seq: ImmClassifier[18],
147 CellTypist[17], HieRFIT[19], and Garnett[20]. We applied ImmClassifier, CellTypist, and HieRFIT as

148 intended, using their pre-designed reference-based training ("pre-trained models", see Methods).
149 CellTypist was applied to annotate either individual events or by majority voting across clusters
150 (MV), as advised[17]. Garnett was designed, similar to MMoCHi, for classification using marker-
151 based hierarchies,[20] and was applied using both a published hierarchy, and a custom hierarchy
152 manually selected markers ("Provided Markers" and "Manual Markers" respectively; see
153 Methods; Supplementary Table 6). These pre-trained models include classification of some non-
154 terminal or other, potentially valid, subsets (e.g. T cells, MAIT cells, T_{reg}, exhausted T cells;
155 "Alternative") and classification of populations excluded by our FACS gating (e.g. neutrophils, B
156 cell lineage, NK cells, or γδ T cells; "Erroneous"). Additionally, HieRFIT and Garnett avoid low-
157 confidence classifications, leaving some events undetermined ("Unlabeled"). For benchmarking,
158 we examined only relevantly classified events (i.e. belonging to one of the sorted subsets), but
159 notably, large portions of CellTypist and ImmClassifier classifications were alternative or
160 incorrect, and most Garnett annotations were undetermined (Fig. 2e). Focusing on relevant
161 classifications, all scRNA-seq classifiers performed on-par with GEX Leiden, significantly
162 underperforming compared to MMoCHi classification (Fig. 2e). HieRFIT Pre-trained and Garnett
163 Provided Markers were excluded from benchmarking analyses, as they could not annotate T cells
164 at an appropriate granularity.

165 All of the supervised approaches described above effectively distinguished monocytes
166 from T cells (Fig. 2e), consistent with previously documented high performance of these tools[17–
167 20]. Our evaluations emphasize the difficulty of segregating T cell subsets that share highly similar
168 transcriptomic profiles without the robust training data and surface protein profiling that MMoCHi
169 leverages. To isolate these two effects from performance of the underlying machine learning
170 algorithms employed, we assessed additional conditions. First, we allowed these tools to "cheat"

171 by supplying them with optimal, batch-free training data—training classifiers using the hashtag-
172 derived sorted subset labels and evaluating performance on a 20% hold-out ("Sort-ref"; see
173 Methods). Although this sort-reference condition is unrealistic (it requires all subsets be FACS-
174 sorted within the same dataset, obviating the need for classification), it revealed differences in
175 performance enabled by MMoCHi's high-confidence thresholding. Notably, we identified only a
176 modest improvement with idealized training data compared to standard MMoCHi classification,
177 highlighting the efficacy of high-confidence thresholding (Extended Data Fig. 4d). Using sort-
178 reference, classifiers designed for scRNA-seq performed on-par with MMoCHi GEX, but worse
179 than MMoCHi (Extended Data Fig. 4d: comparing within Sort-ref). Although Garnett Sort-ref
180 outperformed MMoCHi GEX Sort-ref, it failed to annotate 20% of the cells (Extended Data Fig.
181 4d). We next coerced these scRNA-seq classifiers to accept multimodal data ("Multi"; see
182 Methods). Two variants (HieRFIT Multi Sort-ref and Garnett Multi Sort-ref) performed on-par
183 with MMoCHi Sort-ref, suggesting MMoCHi's underlying machine learning algorithms perform
184 as well as these established tools. Importantly, however, MMoCHi trained using high-confidence
185 thresholds was the only method that could approach this ultra-high accuracy without requiring
186 access to the subset labels from FACS (i.e. "cheating"). Together, these data provide strong
187 support for MMoCHi's high-confidence thresholding and multimodal training, which together
188 enable effective annotation of these highly similar cell types.

189 As described above, MMoCHi also has a built-in batch-correction tool for CITE-seq
190 (landmark registration), which we tested on publicly available datasets using diverse antibody
191 panels, chemistries, and sequencing depth (Extended Data Fig. 5a). Landmark registration (run
192 here with no manual input; see Methods) integrated ADT expression more effectively than other
193 ADT normalization techniques (CLR[5] and dsb[38]), was robust to cell type subsampling, and resulted

194 in a modest improvement in classification performance, especially when the same high-confidence
195 thresholds were applied across batches (Extended Data Fig. 5b-g; Supplementary Table 7).

MMoCHi integrates classification across diverse human tissue immune cells

196 We next applied MMoCHi to total immune cells from lymphoid and mucosal tissue samples
197 obtained from human organ donors[17] (Fig. 3a; Supplementary Table 1). Immune cells were
198 enriched from eight sites across two donors and included lung (LNG), bronchial alveolar lavage
199 (BAL), lung-associated lymph node (LLN), spleen (SPL), jejunum epithelial layer (JEL), jejunum
200 lamina propria (JLP), bone marrow (BOM), and blood (BLD), using methods optimized for each
201 site[17,31]. We performed CITE-seq to profile over 270 surface markers (Supplementary Table 2). In
202 previous analysis of single-cell transcriptomes from this dataset, we detected all lineages of
203 immune cells, including T cells, B cells, innate lymphocytes, and myeloid cells across multiple
204 sites[17]. However, refined immune cell subsets and differences across tissue sites were difficult to
205 resolve in this complex dataset and required extensive manual annotation. Here, we integrated
206 protein and transcriptome profiling using MMoCHi to determine whether transcriptionally similar
207 immune subsets could be identified across diverse tissues.

208 For visualization, we embedded a donor-integrated totalVI latent space with UMAP,
209 revealing multiple groupings, some of which corresponded to different tissue sites (Fig. 3b; see
210 Methods). We constructed and applied a MMoCHi hierarchy representing all expected cell types
211 using a combination of transcript and surface protein markers (Fig. 3c; Supplementary Tables 8,9).
212 Training and classifying these approximately 198k immune cell events into 26 subsets took less
213 than 15 minutes (Extended Data Fig. 4, see Methods), demonstrating that MMoCHi's scalability.
214 The resultant model, using 200 estimators in each random forest, was well-fit at every level of the
215 hierarchy, as measured by the prediction accuracy for internally held-out subsets in the same

classification layer and across the full hierarchy (Fig. 3d; Extended Data Fig. 6; Supplementary Fig. 3; see Methods). MMoCHi classified cell types across tissue sites and consistently across donors (Fig. 3e, Extended Data Fig. 7).

We compared MMoCHi classification to manually annotated clusters of the multimodal totalVI latent space (Fig. 4a,b; Supplementary Fig. 4a,b). The methods were broadly concordant, with 72% of events labeled identically; however, disagreements occurred between cell types with similar expression profiles. These disagreements were not resolved by increasing the number of clusters, nor by altering the number of genes input to totalVI (Supplementary Fig. 4c-d). Notably, unsupervised clustering failed to resolve any $CD8^+$ T_{CM}, γδ T cells, naive B cells, or dendritic cells (DCs). We investigated discrepancies using marker expression (Fig. 4c-e, Supplementary Fig. 5). A substantial percentage (6%) 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[24,32] and their co-clustering with $CD8^+$ T_{EMRA} (Fig. 4a,d). MMoCHi also correctly distinguished $CD8^+$ T_{EMRA} and γδ T cells where cluster-based manual annotation failed, 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.

238 We additionally sought to evaluate MMoCHi's performance when pre-trained classifiers
239 are applied to new datasets. To this end, we applied MMoCHi classifiers trained using the FACS
240 dataset ("FACS-trained") to the αβ T cells and monocytes isolated from the PBMCs of organ
241 donors (Extended Data Fig. 8a-b; see Methods). The FACS-trained MMoCHi model outperformed
242 pre-trained or FACS-trained ImmClassifier, CellTypist, or Garnett models at replicating the
243 correct annotation for each cell type (Extended Data Fig. 8a-f, Supplementary Table 10). Although
244 we do not anticipate that this will be a primary use-case of the algorithm, these data further support
245 MMoCHi's efficacy.

MMoCHi classifiers are highly interpretable and identify cell type markers

246 Having demonstrated MMoCHi's utility for cell type annotation, we next determined which
247 features (transcript and surface protein expression) were useful for subset delineation. Many
248 single-cell classifiers are trained on dimensionally reduced data[21], hampering feature-level
249 interpretability, but random forests within MMoCHi are trained using all available protein-coding
250 genes and surface proteins. During training, features are selected for their contribution to
251 decreasing impurity, thus providing a natural ranking of features by their importance for
252 classification[39,40] (see Methods). We analyzed the important features at each level of the hierarchy,
253 which included subset markers defined by the user (for high-confidence thresholding) and revealed
254 other known subset markers, identified *de novo*, including: *IL1R1* and *GATA2* for mast cells[41],
255 *APOE*, *ACP5*, and *C1QC* for macrophages[42], *CD79A*, *MARCKS*, *MEF2C*, and CD32 for B cell-
256 like[43–46], and *TYROBP*, *GZMB*, and CD123 for pDCs[47,48] (Extended Data Fig. 9; Supplementary
257 Tables 8,11-12). This demonstrates that MMoCHi learns informative cell type representations—
258 an important pre-requisite for novel marker identification.

259 Next, we wondered if MMoCHi could be leveraged to improve transcriptome-based
260 segregation of naive T cells and T_{CM}, which are defined by specific CD45 isoform surface-
261 expression but share similar transcriptomets[1,18,28,30]. First, we trained GEX classifiers using high-
262 confidence $CD4^+$ and $CD8^+$ Naive/T_{CM} selected by multimodal thresholding (Supplementary
263 Table 8). Despite excluding ADTs from training, these classifiers were well-fit, highly accurate,
264 and resulting subsets had expected surface marker expression (Fig. 5a,b). To identify GEX markers
265 of these subsets, we next interrogated impurity-based important features (Fig. 5c,d; Supplementary
266 Tables 13-14). We applied robust filtering to highlight genes with clear differences in expression
267 and reduce the potential for tissue-specific contamination (see Methods), and identified a suite of
268 memory-associated genes, including *ITGB1*, *LMNA*, and *FAM129A*[27,30] as markers upregulated in
269 $CD4^+$ and $CD8^+$ T_{CM}. Improved prediction accuracy by classifiers trained with only the top 1000
270 important features confirmed that these features were useful for GEX classification (Fig. 5e).

271 We sought to validate these findings using our scRNA-seq of sorted $CD4^+$ and $CD8^+$ naive
272 and T_{CM} populations. Most markers identified for naive and T_{CM} were also differentially expressed
273 within the sorted dataset (Fig. 5f,g; Supplementary Table 15). Random forests trained on organ
274 donor PBMCs were well-fit, effectively recapitulating the sorted labels (Fig. 5h,i). Overall, these
275 findings demonstrate the capacity of MMoCHi to leverage CITE-seq to identify gene expression
276 markers and train classifiers that can be effectively applied to scRNA-seq datasets.

MMoCHi effectively annotates other multimodal datasets and cell types

277 Lastly, we explored whether MMoCHi could be easily extended beyond CITE-seq. First, we
278 applied MMoCHi to publicly available paired transcriptome and surface proteome of sorted T and
279 NK cells generated using the Ab-seq platform[7,49]. We obtained high concordance with cluster-
280 based annotations and expected marker expression (Fig. 6a,b; Extended Data Fig. 10a;

281 Supplementary Fig. 6a,d; Supplementary Tables 16-17). Next, we applied MMoCHi to scRNA-
282 seq of a high-grade glioma biopsy[50]. In these data, malignant cells were challenging to identify
283 based on marker genes, because they are highly heterogenous and express many of the same genes
284 as non-neoplastic cells in the brain microenvironment. However, gliomas often harbor clonal
285 aneuploidies that are detectable by averaging gene expression across whole-chromosomes[50]. Using
286 a joint feature set comprised of transcriptome and computed whole-chromosome expression (see
287 Methods), MMoCHi effectively segregated tumor cells from other non-neoplastic subsets, which
288 MMoCHi also annotated accurately (>98% agreement overall; Fig. 6c,d; Extended Data Fig. 10b;
289 Supplementary Fig. 6b,e; Supplementary Tables 16-17). Notably, chromosomal expression was
290 useful both as a marker for high-confidence thresholding and important internally for classification
291 (Extended Data Fig. 10b; Supplementary Table 17).

292 We further explored whether MMoCHi to be adapted to spatial profiling, using a human
293 lymph node dataset generated by the Xenium platform (10x Genomics), which enables multiplexed
294 profiling of 100s of RNAs *in situ* with subcellular resolution. By pairing the 377 profiled genes
295 with physical attributes of each cell extracted from the image and segmentation (see Methods),
296 MMoCHi effectively classified 17 subsets, revealing underlying spatial structure, including B cell
297 follicles and lymphatics (Fig. 6e-h; Supplementary Table 16). MMoCHi outperformed manual
298 cluster-based annotation, which failed to segregate $CD4^+/CD8^+$ T cells and various B cell-lineage
299 subsets (Fig. 6f; Extended Data Fig. 10c; Supplementary Fig. 6 c,f). As before, the added modality
300 was useful for MMoCHi classification, where physical attributes (particularly nuclear circularity
301 and DAPI intensity), ranked in the top 5% of features used (Extended Data Fig. 10c;
302 Supplementary Table 17). Overall, MMoCHi's success in annotating these datasets of varying

303 subset compositions and modalities demonstrates the algorithm's high adaptability to classification
304 of multimodal single-cell datasets.

305

306 **DISCUSSION**

307 The advent of multimodal single-cell technologies has enabled high-dimensional profiling of many
308 systems, organs, diseases, and species. However, the development of analytical tools to identify
309 cell states and their features consistently across multimodal studies is lagging behind these data
310 acquisition technologies. Here, we present MMoCHi, a multimodal, hierarchical classification
311 approach for cell type annotation of CITE-seq data, which integrates both gene and protein
312 expression, does not require reference datasets, performs landmark registration for harmonization
313 across large datasets, and accurately classifies both closely related subsets within a single lineage
314 and diverse subsets from across lineages. We devised an experimental validation approach that
315 allowed us to directly compare MMoCHi and several established annotation modalities directly to
316 FACS-based subsetting for the exact same sample of cells. This enabled an extensive
317 benchmarking study across conventional clustering-based methods, supervised machine learning
318 algorithms, and both reference-based and marker-based classifiers with a uniform and independent
319 reference annotation. Applied to immune cell CITE-seq datasets, MMoCHi outperforms current
320 annotation algorithms, and identifies new markers for subset delineation. Together, MMoCHi
321 provides an adaptable approach for applying marker-based annotation to multimodal datasets.

322 Currently available single-cell classifiers are primarily designed for scRNA-seq; however,
323 CITE-seq can be leveraged to improve annotation[5,6]. MMoCHi robustly classifies closely related
324 immune cell subsets with discordant transcriptome and surface proteome profiles. Here we show

325 that MMoCHi effectively distinguishes T cell subsets defined by surface expression of CD45
326 isoforms and homing receptors but share overlapping transcriptomes[1,18,28,30]—namely naive T cells
327 from T_{CM} and T_{EM} from T_{EMRA}. By classifying hierarchically, MMoCHi was also able to correctly
328 annotate diverse immune lineages without compromising its performance at segregating
329 functionally and transcriptionally similar cell types, including cytotoxic NK cells, CD8+ T_{EMRA},
330 and γδ T cells[24,32,33,51]. MMoCHi can accurately classify cell types despite other sources of
331 variation, as shown by integrated classification of multiple immune cell lineages across blood and
332 8 disparate tissue sites of two organ donors.

333 By providing a platform for using high-confidence, marker-based thresholding for
334 selection of training data, MMoCHi enables the flexible design of new classifiers based on prior
335 knowledge without a requirement for carefully curated reference atlases. In particular, applying
336 high-confidence thresholding to CITE-seq protein expression serves as a natural extension of the
337 vast knowledge-base developed of flow cytometry markers and gating[26,36]. Similar to other
338 marker-based strategies, these subset definitions are highly interpretable and can be robustly
339 applied across studies and specific sequencing conditions to train new classifiers[20,25]. To reduce
340 redundancy in efforts to design MMoCHi hierarchies, the hierarchies defined in this manuscript,
341 along with other community contributions, have been made available online
342 (https://mmochi.readthedocs.io). Here, we also apply MMoCHi's multimodal training data
343 selection to improve transcriptome-based annotation of T cell subsets across datasets,
344 demonstrating potential for MMoCHi to advance annotation of scRNA-seq data as well.

345 In contrast with efforts to automate cell type annotation[16,17], MMoCHi's thresholding
346 schemas require careful marker curation and domain expertise, however; by leveraging this
347 expertise, MMoCHi classifications may better reflect canonically defined, and biologically

348 relevant designations. Additionally, MMoCHi is not optimized for novel subset identification,
349 requiring the user to define all cell types within a sample for classification. Thus, MMoCHi is
350 complemented by multimodal unsupervised exploration of the dataset either before or after
351 classification to annotate or identify markers of unexpected cell types and states, or to identify
352 novel subsets of canonical cell types.

353 Beyond cell type classification, MMoCHi learns key features of protein and gene
354 expression, which can be used to identify new cell type markers and derive biological insights.
355 MMoCHi random forests do not require prior dimensionality reduction, enabling evaluation of
356 individual surface proteins and protein-coding genes as subset markers. In classifying diverse and
357 highly similar subsets, we show that MMoCHi extracts relevant markers, relying on both features
358 used for training data selection, and other known gene and protein markers of each subset. In using
359 MMoCHi to identify improved transcriptional markers of naive and central memory (T_{CM}) T cells,
360 we discovered that circulating T_{CM} actually upregulate markers that had previously been associated
361 with T cells in tissue[30,31,52] and effector memory T cells[27,30]. This finding raises the possibility that
362 early stages of tissue programming occur during T_{CM} differentiation and indicate a gradual
363 transition between T cell memory subsets. Importantly, we did not previously identify these
364 expression patterns of circulating T_{CM} in earlier studies of T cells in tissues and blood[30,31], likely
365 due to the lack of CITE-seq and multimodal classification.

366 While we have developed MMoCHi with CITE-seq applications in mind, the algorithm is
367 designed for easy extension to other modalities frequently paired with single-cell transcriptomes,
368 such as profiling chromatin accessibility, T and B cell receptors, mutational landscapes,
369 intracellular proteins, or a combination of these modalities[53–58]. We also anticipate applications to
370 emerging technologies for multimodal, single-cell spatial profiling[59,60]. We demonstrated the

potential generality of MMoCHi through extension to joint aneuploidy/gene expression profiles from a solid tumor and joint classification of *in situ* spatial profiling data using morphological and gene expression features measured using the 10x Genomics Xenium platform. While we mainly 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.Sz., 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 (e.g. 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. Similar to other hierarchical classifiers,[19,20] hierarchies progressively segregate subsets by similarity of expression profiles, which may or may not reflect a cell type's ontogeny. 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) and antibody derived tag (ADT) expression were log-normalized. 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 in multi-sample datasets was batch-corrected using a landmark registration algorithm (Fig. 1a), similar to previously described techniques for flow cytometry and CITE-seq[34,35]. Landmark registration was applied to expression of all ADTs, except for isotype controls, which were excluded from downstream analyses. 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*, using the Scipy package v1.10.1) on kernel-density-smoothed (*scipy.stats.gaussian_kde*) ADT expression, or manually identified. Curve registration and warping functions were applied to align these landmarks across samples (*skfda.preprocessing.registration.landmark_elastic_registration_warping*, using the scikit-fda package v0.8.1). For each ADT in each sample, the lowest detected peak (assumed to represent "negative" expression) and the uppermost peak (assumed to represent "positive" expression) were aligned to center on fixed values: 1 and 3, respectively. In samples with only one peak detected, the peak was aligned to the "negative" location, although this could be adjusted manually to align batches with only positive-expression for an ADT. As a batch-correction technique, landmark registration assumes that changes across samples in the degree of marker positivity on marker-positive populations is due to technical artifacts, not biological differences. This is a reliable assumption for annotation purposes, especially when samples are from similar sources or contain the same cell types at varying frequencies.[34] Once batch-corrected, thresholds delineating positive and negative populations for ADT markers could be applied across batches.

Selection of training data

429 At each level of the hierarchy, we selected high-confidence events for each subset using a
430 combination of automatically-derived and manually-adjusted thresholds on the user-supplied
431 marker definitions (Fig. 1b, Supplementary Tables 4,7,8,16). ADT markers were fitted with 1- or
432 2- component Gaussian mixture models (*sklearn.mixture.GaussianMixture*, using the Sklearn
433 package v1.3.0), and thresholds were automatically defined as 1 to 4 standard deviations above or
434 below the means of each Gaussian. ADT thresholds were then manually refined using established
435 flow cytometry gating principles (e.g. Ref.[36]). GEX thresholds were drawn similar to Garnett[20],
436 primarily capturing events with any marker expression, or occasionally (in genes with high
437 background expression) capturing only the highest expressing events. A portion (20%) of these
438 high-confidence events were held out for testing and validation, and the remaining 80% could be
439 used for training. We then resampled this training dataset (Extended Data Fig. 1a) to remove events
440 likely to be mistakenly labeled due to imperfect marker thresholding (noise) and to overrepresent
441 events likely to be misclassified (in danger). To identify "noise" and "in danger" events, principal
442 component analysis (PCA; *scanpy.pp.pca*, using the Scanpy package v1.9.3) was run on the scaled
443 expression of the top 5000 highly variable genes (*scanpy.pp.highly_variable_genes*) and all ADTs.
444 The 5 nearest neighbors were calculated for each event (*sklearn.neighbors.NearestNeighbors*).
445 Events were considered "noise" and removed if all neighbors disagreed with their high-confidence
446 label. Events were considered "in danger" if less than half the neighbors agreed with their high-
447 confidence label. Training events were also clustered using the Leiden algorithm[61]
448 (*scanpy.tl.leiden*) to identify "in danger" clusters—clusters representing less than 5% of a subset's
449 training events. All events considered "in danger" were oversampled 5 times to increase
450 representation during training. At classification nodes where subsets were not expected to
451 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 high-confidence 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 a default of 100 trees, each with a max depth of 25 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[62]. Classification performance was then evaluated on the remaining hold-out data.

Hyperparameters were tuned in cases of poor fit, either automatically or manually. Hyperparameter tuning on the underlying random forest classifiers focused on optimizing values for the number of trees in the forest (*n_estimators*), the maximum depth of each tree (*max_depth*), and the maximum number of features to consider when looking for the best split during random forest training (*max_features*). For each hyperparameter tested, fit was determined by calculating the balanced accuracy score (*sklearn.metrics.balanced_accuracy_score*) between predicted and high-confidence labels on a subset of the held-out data (distinct from the subset used for performance evaluation). For automatic optimization, *n_estimators* were tested at 50, 100

474 (MMoCHi default), 200, 400, 800, and 1200; *max_depth* of 10, 25 (MMoCHi default), or
475 unlimited; and *max_features* of $\log_2$(total), square-root of total (MMoCHi default), 5% of total, or
476 10% of total features. Optimal hyperparameter values were selected as either the maximum score,
477 or as the lowest value before this maximum score where score did not improve by at least 0.004.
478 To increase fit, the *max_features* hyperparameter was manually set to 10% of total features for
479 $CD4^+$ and $CD8^+$ memory classification for sorted T cells, as well as Lymphocyte, TCR, and CD4/8
480 classification for organ donor cells.

481 The trained classifier was then used to predict subset identity for all events at that level.
482 This process of training data selection, classifier training, and prediction were repeated until all
483 cells had been labeled to terminal subsets.

CITE-seq profiling of FACS sorted subsets

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

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

494 resuspended in ACK lysis buffer (Gibco) incubated for 2 minutes at 37°C and washed with FACS
495 buffer.

496 PBMCs were stained with Zombie NIR Fixable Viability dye (BioLegend) for 30 minutes.
497 Samples were kept at 4°C in the dark. Cells were washed thrice with FACS buffer, resuspended in
498 TrueStain FcX and TrueStain Monocyte Blocker (BioLegend), and incubated for 10 minutes. We
499 designed a FACS-sort antibody cocktail, prioritizing antibody clones with discrete epitopes from
500 the TotalSeq-A Universal Human Panel (BioLegend) to reduce steric hinderance during CITE-seq
501 staining (Supplementary Table 2). Cells were incubated with the FACS-sort antibody cocktail for
502 30 minutes, then washed thrice with FACS buffer. Cells were incubated in TrueStain FcX
503 (BioLegend) for 10 minutes, then stained using a custom TotalSeq-A Universal Human Panel
504 (BioLegend) for 30 minutes, according to manufacturer instructions. Samples were washed thrice
505 with FACS buffer. Cells were sorted using a FACS Aria II (BD Biosciences; Extended Data Fig.
506 2a). Seven T cell memory populations ($CD4^+$ Naive, $CD4^+$ T_{CM}, $CD4^+$ T_{EM}, $CD8^+$ Naive, $CD8^+$
507 T_{CM}, $CD8^+$ T_{EM}, $CD8^+$ T_{EMRA}), and monocytes were sorted into sterile, heat inactivated FBS. Sort
508 purity was calculated as the number of events falling within a subset's gates divided by the total
509 number of singlet events times 100, using the same gating strategy as the sort (Extended Data Fig.
510 2a). All sorts were high purity, with mean purity 94.4% (Supplementary Table 3). Sorted
511 populations were washed in FACS buffer and resuspended in TrueStain FcX (BioLegend) for 10
512 minutes. Samples were then stained with TotalSeq-A hashtag antibodies (BioLegend). In this
513 experiment, the hashtag-oligos (HTOs) correspond to the sorted immune cell subsets (HTO1:
514 $CD4^+$ Naive, HTO2: $CD4^+$ T_{CM}, HTO3: $CD4^+$ T_{EM}, HTO4: $CD8^+$ Naive, HTO5: $CD8^+$ T_{CM},
515 HTO6: $CD8^+$ T_{EM}, HTO7: $CD8^+$ T_{EMRA}, HTO8: Monocytes). Samples were then washed thrice
516 with FACS buffer, and pooled.

CITE-seq profiling of human tissue samples

517 Human tissues were obtained from deceased organ donors as previously described[17,31,63,64]. The
518 use of tissues from organ donors is not considered human subjects research as confirmed by the
519 Columbia University IRB because the donors are deceased. Mononuclear cells (MNCs) were
520 isolated from tissue sites of two organ donors (D496 and D503; Supplementary Table 1), as
521 described[17]. Approximately 1 million MNCs per tissue site were washed in FACS buffer and
522 resuspended in TrueStain FcX (BioLegend) for 10 minutes. Samples were then stained with
523 TotalSeq-A hashtag antibodies (BioLegend) for each tissue. For each donor, the hash-tagged
524 MNCs from each tissue site were pooled, washed with FACS buffer, and stained with the
525 TotalSeq-A Human Universal Cocktail panel according to the manufacturer's instructions
526 (BioLegend).

Library preparation, sequencing, and alignment

527 The sorted immune cells were counted, diluted to an appropriate volume, and loaded across two
528 lanes of a 10X Genomics Chromium instrument targeting 6,000 cells each. Samples from each
529 organ donor were loaded across 16 lanes of a 10X Genomics Chromium instrument targeting
530 10,000 cells each. cDNA synthesis, amplification and sequencing libraries were generated using
531 the Next GEM Single Cell 3' Kit v3.1 (10X Genomics) with the recommended modifications for
532 compatibility with the TotalSeq-A cell hashing and CITE-seq reagents (BioLegend). Organ donor
533 GEX libraries were sequenced on a NovaSeq 6000 (Illumina) with 100 cycles for reads 1 and 2.
534 Organ donor ADT and HTO libraries were sequenced on a NextSeq 500 (Illumina) with 28 cycles
535 for read 1 and 55 cycles for read 2. All libraries for FACS sorted subsets were sequenced on a
536 NextSeq 500 (Illumina) with 28 cycles for read 1 and 44 cycles for read 2.

537 Reads were analyzed by pseudoalignment using kallisto v0.46.2 (GRCh38 with Gencode
538 v24 annotation) and bustools v0.40.0[65–67]. CITE-seq and hashtag barcodes were demultiplexed and
539 extracted using DropSeqPipeline8, as previously described[68]. Hashtags were demultiplexed by
540 CLR normalization, k-means clustering, and statistical identification of singlets by fitting a
541 negative binomial model as described[5]. For both datasets, the GEX matrices were normalized to
542 log(counts*10,000/total_counts+1), and ADT expression matrix were normalized to
543 log(counts*1,000/total_counts+1).

Analysis and benchmarking using sorted T cells

544 Cells from the T cell sort were filtered to remove events with fewer than 1000 unique counts, fewer
545 than 200 genes detected, or over 10% mitochondrial counts. A MMoCHi hierarchy was developed
546 (Fig. 2b) and classification performed using the algorithm above. In flat classification variations
547 (MMoCHi Flat, MMoCHi GEX Flat), high-confidence thresholding and classification were
548 performed for all subsets in a single classification node. In GEX variations (e.g. MMoCHi GEX,
549 MMoCHi GEX Flat, etc.), ADT expression data was excluded from the detection of in-danger
550 noise events and random forest training. In "Sort-ref" variants (e.g. MMoCHi Sort-ref, MMoCHi
551 GEX Sort-ref, etc.), training events were selected using the sort labels instead of high-confidence
552 thresholding. For benchmarking tests, to mirror intra-dataset performance testing of other
553 reference-based tools[17,19] and avoid leaky pre-processing bias[69] a portion of the dataset (20%) was
554 set aside to only be used for performance evaluation (completely held-out from the delineation of
555 high-confidence thresholds, training data selection, random forest training, random forest
556 calibration, and any manual or automatic hyperparameter optimization). To test MMoCHi's
557 robustness, performance was also tested by subsampling training event numbers and
558 downsampling reads. Random subsampling was performed on the number of monocyte or $CD8^+$

559 T_{CM} events included in the training dataset, down to 5% of the total of monocyte/CD8$^+$ T_{CM} events
560 profiled. Aligned reads for GEX and ADT libraries were downsampled prior to generation of raw-
561 count expression matrices. Downsampling was performed down to 1% of the total aligned reads,
562 and the effect of this downsampling on total counts is displayed.

563 We manually annotated unsupervised clusters by average expression of gene and protein
564 markers. Leiden clusters of gene expression (GEX Leiden) were computed on a PCA of the highly
565 variable genes with Scanpy[70] defaults. Leiden clusters of ADT expression (ADT Leiden) were
566 computed on a PCA of all ADTs except for isotype controls. totalVI latent space was computed
567 on all ADTs except for isotype controls, and highly variable genes selected using top 4000 genes
568 as defined by the Seurat v3 method[71], as recommended[14] (*scvi.model.TOTALVI.train* and
569 *scvi.model.TOTALVI.get_latent_representation* using scvi-tools v1.0.2). To ensure results of
570 totalVI were not sensitive to the number of features used, we also computed totalVI latent spaces
571 using the top 1000, 2000, and 8000 highly variable genes. Leiden clustering was performed on the
572 10 nearest neighbors of the top 40 principal components or the entire totalVI latent space. Over-
573 clustering was also computed where highly variable gene selection, dimensionality reduction, and
574 Leiden clustering were repeated to sub-cluster each cluster (excluding clusters of fewer than 40
575 events) resulting in 72 GEX clusters (GEX Leiden OC), 128 ADT clusters (ADT Leiden OC), and
576 116 totalVI clusters (totalVI Leiden OC). For visualization, UMAP (*scanpy.tl.umap*) embeddings
577 of each of these feature spaces were calculated using defaults.

578 ImmClassifier[18], CellTypist[17], and HieRFIT[19] were applied as recommended, using pre-
579 trained models downloaded from online. ImmClassifier was run using *docker* with the built-in
580 model, trained by inputting a $\log_2$(counts*10,000/total_counts+1) transformed GEX matrix, as
581 recommended. CellTypist v1.6.2 was run by applying the "Healthy_COVID19_PBMC" model

582 (trained on PBMCs from healthy individuals and individuals with COVID-19[72]) with and without
583 majority voting enabled (*celltypist.annotate*; CellTypist MV and CellTypist, respectively). A
584 HieRFIT v0.1.0 model trained on 10x PBMC data was applied (*HieRFIT::HieRFIT*) with defaults
585 settings.

586 Garnett[20] models were trained using high-confidence thresholding with the provided online
587 hierarchy (Provided Markers) and on manually and automatically (*garnett::top_markers*) selected
588 transcript markers which were evaluated using built-in functions (*garnett::check_markers*;
589 Manual Markers; Supplementary Table 6). Garnett v0.2.20 models were trained
590 (*garnett::train_cell_classifier*) with marker propagation disabled and prediction
591 (*garnett::classify_cells)* was performed with cluster extension disabled.

592 Additionally, CellTypist, HieRFIT, and Garnett were trained using the HTO-derived sort
593 labels as a reference dataset and a evaluated using a 20% hold-out ("Sort-ref"). A CellTypist model
594 was trained (*celltypist.train*) with two-pass training enabled for feature selection. This model was
595 applied to held-out data with and without majority voting enabled (CellTypist MV Sort-ref and
596 CellTypist Sort-ref, respectively). HieRFIT and Garnett models were both trained
597 (*HieRFIT::CreateHieR*; *garnett::train_cell_classifier*) using a hierarchy structured identically to
598 the MMoCHi hierarchy (Fig. 2b) and applied to held-out data (HieRFIT Sort-ref and Garnett Sort-
599 ref, respectively). The three models were also retrained using the combined multimodal matrix of
600 GEX and ADT expression ("Multi"). To coerce these tools to accept the multimodal matrix, the
601 *check_expression* parameter in CellTypist functions and the *db* parameter in Garnett functions
602 were disabled.

603 Precision, recall, F1 scores and overall accuracy were calculated using sort labels as truth
604 (*sklearn.metrics.precision_recall_fscore_support*; *sklearn.metrics.accuracy_score*). Garnett and

605 HieRFIT can provide unknown and intermediate cell type labels, which were quantified and
606 excluded from performance metrics calculations. Pre-trained classifier models could also provide
607 erroneous (e.g. not included in the FACS sort) or alternative (i.e. potentially valid annotations but
608 not using the subsetting-axis of interest) cell type labels, which were also quantified, then excluded
609 from performance calculations (Supplementary Table).

Analysis of ADT batch-correction with published CITE-seq data

610 To test landmark registration against other published CITE-seq normalization techniques, we used
611 four public datasets of CITE-seq performed on healthy PBMCs available from 10x Genomics
612 (Extended Data Fig. 5a). Each dataset was treated as a separate batch. For each batch,
613 log(counts*1,000/total_counts+1) and landmark registration were performed as described above.
614 For these data, no manual modifications were made to the built-in automatic peak detection. CLR
615 and dsb normalizations were performed using the Muon package[73] v0.1.5 (*muon.prot.pp.clr* with
616 *axis* set to 1; *muon.prot.pp.dsb*). The dsb algorithm uses empty droplets to identify protein-specific
617 noise originating from unbound antibodies[38]. Empty droplets were defined as events that contained
618 between 2 and 3 $\log_{10}$(ADT-counts), or (for 5k_pbmc_3p) between 1.5 and 2.8 $\log_{10}$(ADT-counts)
619 in the provided unfiltered ("raw") matrix and did not appear in the filtered matrix. Isotype controls
620 (IgG1, IgG2a and IgG2b) were then used to remove cell-specific noise, as described[38]. Methods
621 were evaluated using UMAPs and Leiden clustering computed as described above, using
622 normalized expression of the 13 proteins profiled in all batches (CD3, CD4, CD8a, CD14, CD16,
623 CD19, CD25, CD45RA, CD45RO, CD56, CD127, PD-1, and TIGIT). Nearest-neighbors were
624 computed using "spearmanr" as the metric, as ADT expression was not scaled (to best-preserve
625 each integration method). Batch integration was quantified with adjusted rand index (ARI)
626 (*sklearn.metrics.adjusted_rand_score*). To test whether landmark registration was robust to

627 changing cell proportions, landmark registration was also performed after removing 95% of the
628 events with positive expression of CD14 (defined as above 2 log(CP1k+1)). The effect of landmark
629 registration on MMoCHi classification was tested by applying MMoCHi to classify subsets across
630 batches to log(CP1k+1)-transformed or landmark registered protein expression, with thresholds
631 manually defined for all batches together or each batch separately (Supplementary Table 7).

Analysis of organ donor sequencing

632 CITE-seq data obtained from tissue immune cells was filtered separately. Although hashtag
633 demultiplexing removes inter-sample multiplets, this method cannot detect multiplets between
634 cells from the same tissue site. Thus, multiplets were detected by Scrublet[74]
635 (*scrublet.Scrublet.scrub_doublets*), using default settings, an expected doublet rate of 0.015, and
636 applying separately to each library-tissue combination of over 100 events. Cells were then filtered
637 to remove events with fewer than 1000 unique counts, fewer than 600 genes detected, or over 10%
638 hemoglobin counts. We used a cluster-based method to remove events either identified as doublets
639 by Scrublet or with high mitochondrial counts, similar to the previously described percolation
640 method[17]. Briefly, a PCA on highly variable genes was calculated, integrated using Harmony[75]
641 (*scanpy.external.pp.harmony_integrate*), and used for nearest neighbor calculation and Leiden
642 clustering with a resolution of 20. Clusters with significantly higher Scrublet scores (above 0.1) or
643 percent mitochondrial counts (above 15%) were removed. Individual events not captured by these
644 clusters that were identified as Scrublet doublets or had over 25% mitochondrial counts were also
645 removed.

646 We devised a MMoCHi hierarchy (Fig. 3d; Supplementary Table 8) and performed
647 classification using the algorithm above. Due to insufficient numbers, T_{SCM} were not included in
648 the hierarchy (Supplementary Table 9). Classification fit across a varying number of random forest

649 estimators (trees) was evaluated for prediction accuracy. Overall prediction accuracy was
650 measured using internally held-out events (events that were high-confidence thresholded but
651 excluded from the training of random forest classifiers). Prediction accuracy was also calculated
652 for individual classification nodes using the weighted average of precision, recall, and F1 scores
653 for each child subset using internally held-out, high-confidence events. Gini impurity-based
654 feature importances were automatically calculated during random forest training by scikit-learn.
655 Manual annotation and UMAP calculation were performed on the totalVI latent space, as described
656 above. The clustered dendrogram of classified subsets in each donor was calculated
657 (*scanpy.tl.dendrogram*) on the totalVI latent space with optimal ordering enabled.

658 Pre-trained classifiers were applied to organ donor PBMC events that were classified as
659 one of: $CD4^+$ Naive, $CD4^+$ T_{CM}, $CD4^+$ T_{EM}, $CD8^+$ Naive, $CD8^+$ T_{CM}, $CD8^+$ T_{EM}, $CD8^+$ T_{EMRA}, or
660 classical monocytes. MMoCHi or Garnett models trained on the sorted FACS dataset (FACS-
661 trained MMoCHi, FACS-trained MMoCHi GEX, and FACS-trained Garnett) or pre-trained
662 ImmClassifier or CellTypist models were applied as described above. For ImmClassifier,
663 CellTypist, and Garnett, Alternative, Unknown, or Erroneous classifications were quantified as
664 described above.

665 To display of features associated with each subset (Extended Data Fig. 9), the top
666 important features with a $\log_2$(fold-enrichment) > 2 for that subset. For identification of
667 transcriptomic markers for Naive T cells and T_{CM} (Fig. 5), MMoCHi models were retrained at the
668 $CD4^+$ and $CD8^+$ Naive/T_{CM} level, given the same multimodal high-confidence thresholding, but
669 trained on only the GEX matrix. To ensure these GEX markers were not learned through tissue-
670 specific contamination, additional models were trained using only T cells from organ-donor blood,
671 and features displayed had to be in the top 1% of important features in both models. To improve

robustness, they were additionally filtered to only those with a greater than 10% change in dropout rate.

Application of MMoCHi to other modalities

To apply MMoCHi to other modalities beyond CITE-seq, we used three public datasets: Ab-seq of sorted T and NK cells[49], scRNA-seq of a glioma biopsy[50], and Xenium of a human lymph node obtained from 10x Genomics. For the Ab-seq dataset, only resting cells were used. The gene expression matrix was normalized to log(counts*1,000/total_counts+1), and the protein expression matrix was normalized to log(counts*100/total_counts+1). Clustering was performed on a totalVI latent space, calculated as described above. Manual annotations of these clusters were highly congruent with published annotations.[49] In the glioma dataset, the gene expression matrix was normalized to log(counts*10,000/total_counts+1). The ratio of the average expression of genes on chromosome 7 to chromosome 10 was used as an additional marker, effectively segregating malignant cells, as described[50]. Manual annotations and UMAP coordinates were obtained from the published clustering analysis. In the Xenium dataset, the gene expression matrix was normalized to log(counts*100/total_counts+1). Gene-expression probes marked as "negative control" or "blank" were removed. The dataset was filtered to remove any events with fewer than 10 total counts and 5 unique genes expressed. Morphological features were calculated using the nucleus and cell boundaries, as determined by the Xenium Onboard Analysis pipeline v1.5.0. Nuclear or cellular area was calculated in μm^2 by converting boundary vertices to Polygon objects using the Shapely package v2.0.2 (*shapely.geometry.Polygon; shapely.area*). Nuclear or cellular roundness, or circularity, was calculated as the ratio of a shape's area to the area of the minimum bounding circle of that shape (*shapely.minimum_bounding_circle*). Nuclear DAPI intensity was calculated by obtaining average pixel-intensity of the provided 2D maximum intensity projection

694 (“MIP") image. To obtain the pixels overlapping with the nucleus, we used the Rasterio package
695 v1.3.9 (*rasterio.features.rasterize*) using a transformation (*affine.Affine.scale*) of 0.2125µm per
696 pixel, as defined in image metadata. These features were supplied to MMoCHi without additional
697 scaling. Manual annotation of clusters was performed on the provided graph-based clusters and
698 displayed on the provided graph-based UMAP projection. MMoCHi was applied to all three
699 datasets using hierarchies of known marker genes, proteins, or chromosome-level expression
700 patterns (Supplementary Table 16). For figure generation, Xenium cell boundaries were plotted on
701 top of the provided 2D autofocus projection image of DAPI staining using the same scaling as
702 above. Cell boundaries were shrunk slightly for display to reduce drawing overlap (using the
703 *shapely.geometry.Polygon.buffer* method, with *join_style* “mitre”). Feature importances were
704 obtained as described above.

Estimation of computational performance

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

Statistical analysis

711 Statistical comparison of classification performance was calculated using a Friedman rank-sum
712 test (*stats::friedman.test* using the stats library v4.3.1), matched by cell-subset, on F1 scores,
713 followed by multiple comparisons using a paired two-sided Wilcoxon signed-rank test

714 (*stats::pairwise.wilcox.test*) with FDR correction. Significance groups for each method are
715 displayed by lettering, where methods not sharing a letter are significantly different ($p < 0.05$).

716 Prior to differential expression, the two groups were subsampled to the same number of
717 events and their expression matrices were downsampled to equal total counts. Differential
718 expression was performed (*scanpy.tl.rank_genes_groups*) on log-normalized expression, with
719 Wilcoxon mode, tie correction, and FDR correction enabled. Genes or proteins with $|\log_2(\text{fold}$
720 $\text{change})| > 2$ and $p < 0.05$ were identified as significantly differentially expressed.

DATA AVAILABILITY

721 CITE-seq data generated in this study are available in NCBI GEO with the accession code
722 GSE229791. Public datasets used for analyses were downloaded online and are available from:
723 10x Genomics (https://www.10xgenomics.com/resources/datasets), Open Science Framework
724 (https://doi.org/10.17605/OSF.IO/EDCTN), or NCBI GEO with the accession code GSE116621.

CODE AVAILABILITY

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

726 Code for demultiplexing CITE-seq and hashtag barcodes is available at:
727 https://github.com/simslab/DropSeqPipeline8

MAIN FIGURES

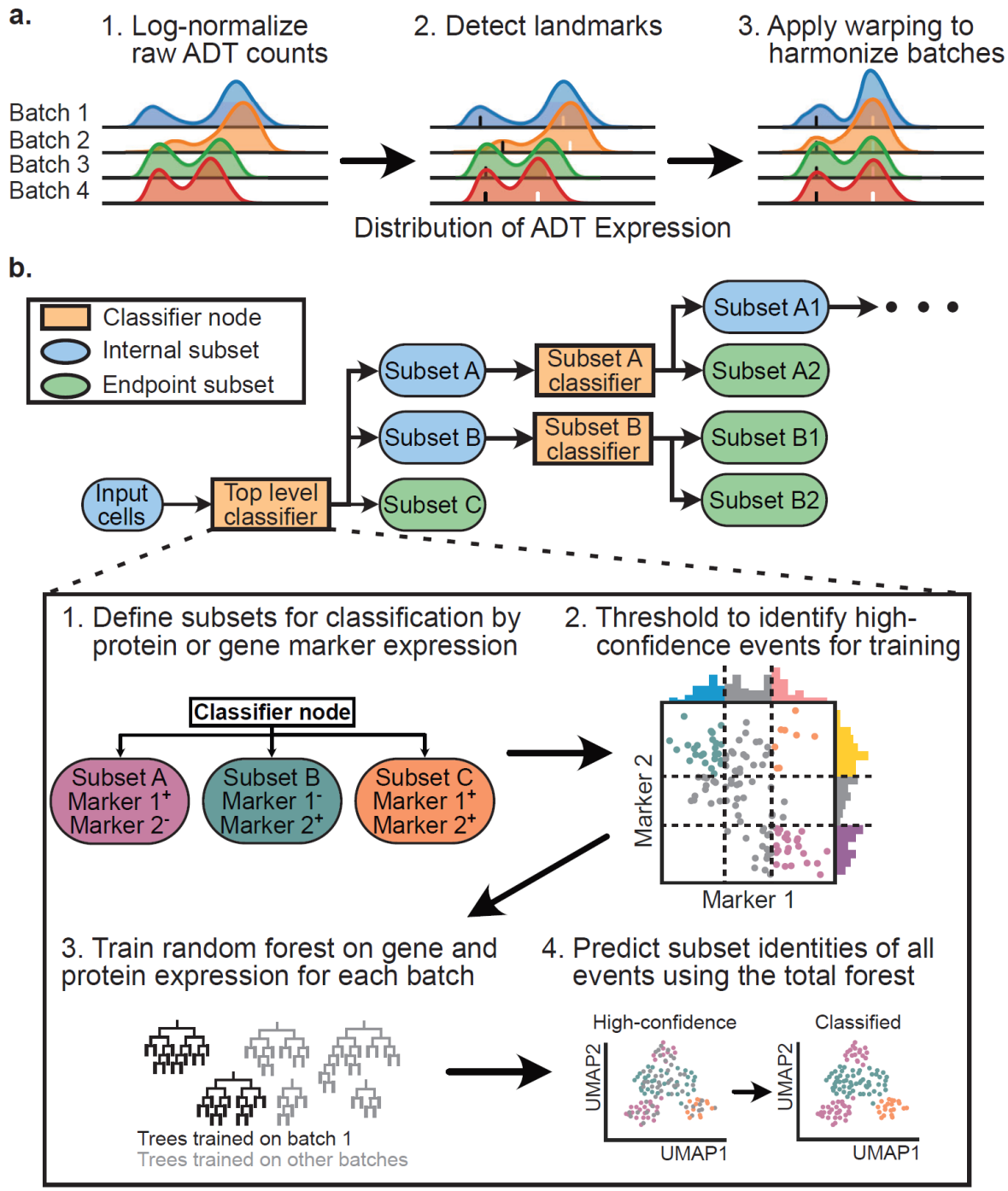


Figure 1 **Schematic showing general workflow of MMoCHi.** a. Batch correction of Antibody Derived Tag (ADT) expression was performed by landmark registration on log-normalized counts (1). After detection of positive (white ticks) and negative (black ticks) peaks in each batch (2), a warping function 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.

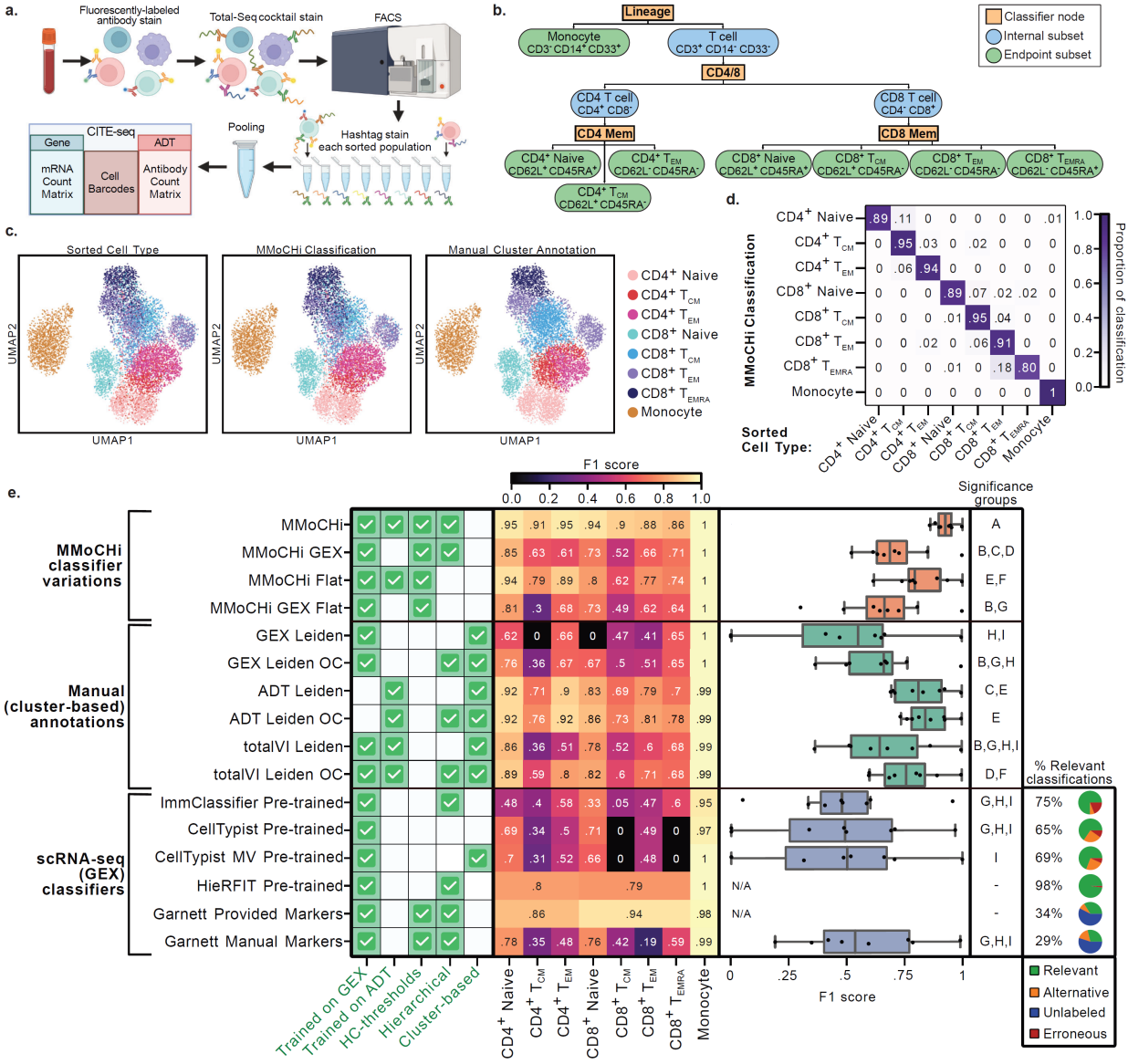


Figure 2 **MMoCHi classification of predefined T cell subsets outperforms other annotation methods.** a. PBMCs were stained with fluorescently-labeled antibodies and a cocktail of oligo-tagged antibodies for sequencing. Seven T cell subsets and monocytes were sorted, hashtagged, pooled, and sequenced. 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, or 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 (right). Features of each method are labeled for context (left) using green checks: “Trained on GEX/ADT”—whether gene expression (GEX) or antibody derived tag (ADT) expression data were used during model training. "HC-thresholds"—whether high-confidence thresholds were used for training data selection. "Hierarchical"—whether annotation was performed on multiple levels. “Cluster-based”—whether

annotations were applied to unsupervised clusters. For scRNA-seq classifiers, the percent of classifications that were relevant to this analysis, were alternative (but potentially valid) classifications, were unlabeled, or were entirely erroneous are shown using pie charts (far right). Only relevant classifications were included for calculation of F1 scores. Statistical significance was calculated using a Friedman rank-sum test, matched by subset, followed by multiple comparisons testing using paired Wilcoxon signed-rank tests and FDR correction. Significance groups for each method are shown by lettering, where methods not sharing a letter are significantly different ($p < 0.05$). Methods that could not label T cell memory subsets were excluded from significance testing and the merged efficacy of $CD4^+$ or $CD8^+$ T cell classifications were displayed instead. 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

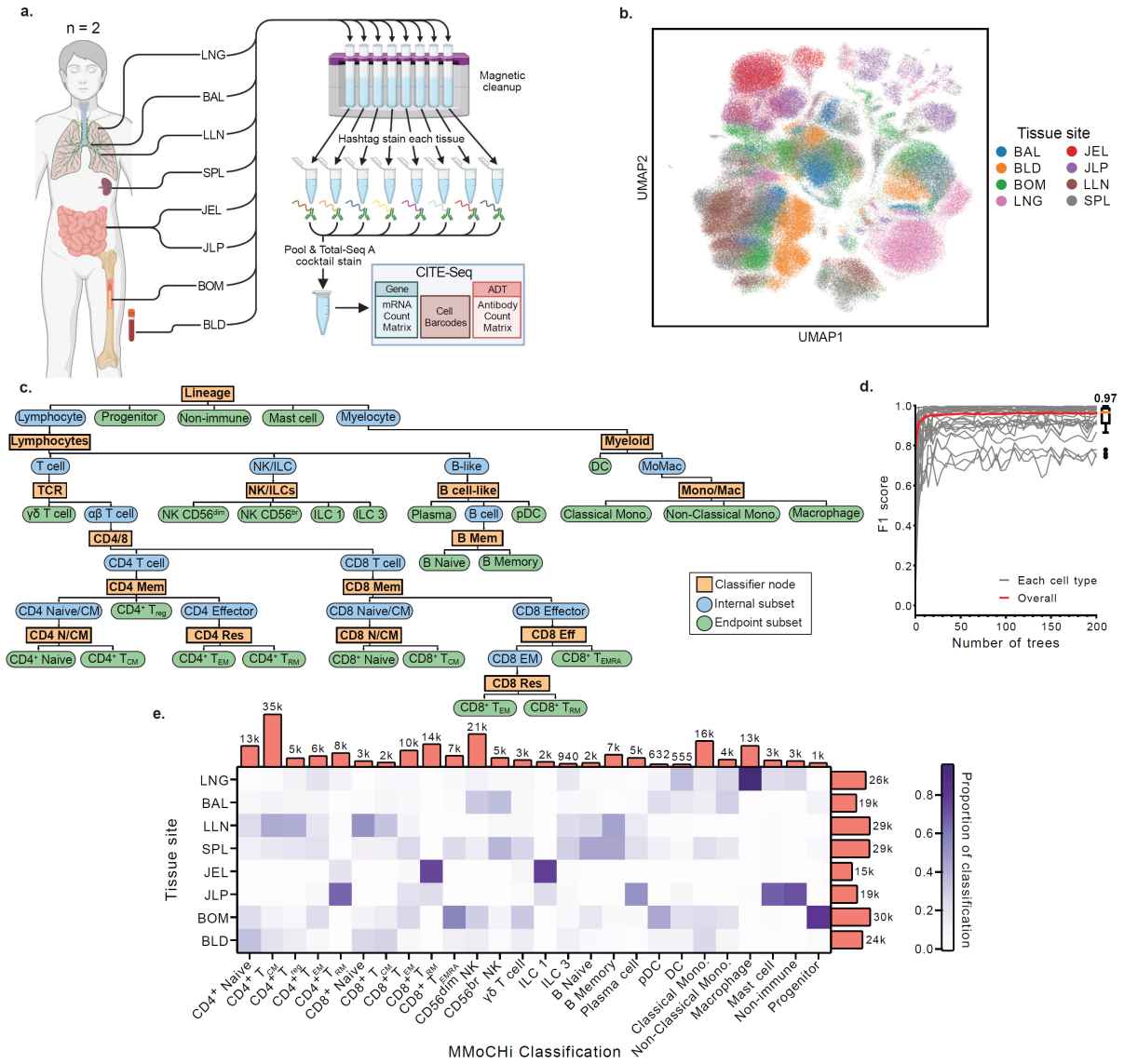


Figure 3 **MMoCHi classification applied to human immune cells from blood and diverse tissue sites.** a. Single cell suspensions of immune cells were isolated from blood and the indicated tissue sites of two organ donors. CD45^{+} immune cells selected for by magnetic enrichment, hashtagged by tissue 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 8 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; Mem, Memory; N/CM, Naive/Central Memory; Res, Residency; Eff, Effector. Schematic in (a) created with BioRender.com

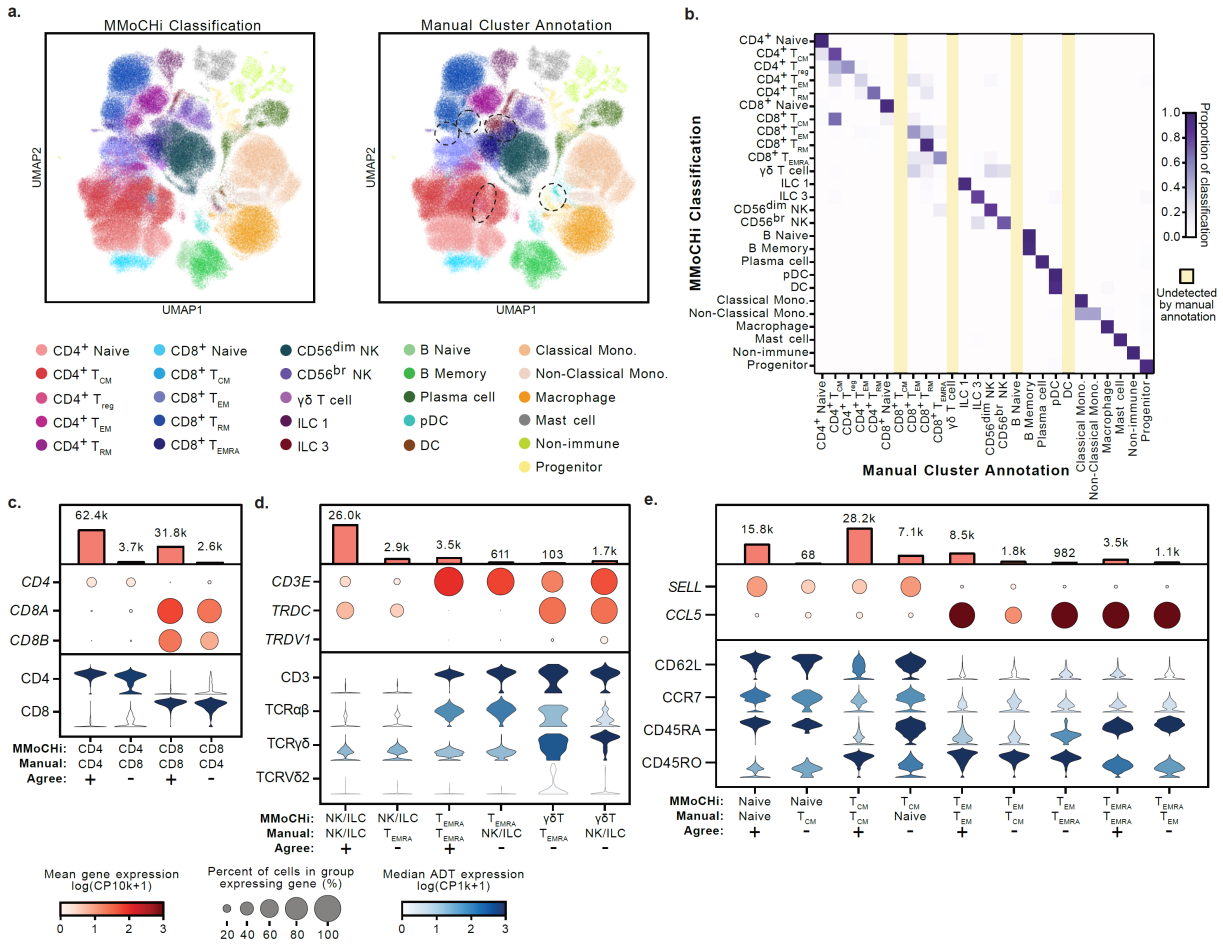


Figure 4 **MMoCHi classification of human immune cells outperforms manual annotation by clustering.** a. UMAPs of organ donor cells, colored by MMoCHi classification and manual cluster annotations. Dashed ellipses highlight areas of major disagreement between MMoCHi classification and manual annotation. b. Row-normalized heatmap comparing MMoCHi classification to manual annotation. Color represents proportion of cells in each MMoCHi classified population 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

878
879
880
881
882
883
884
885
886
887
888
889
890
891
892

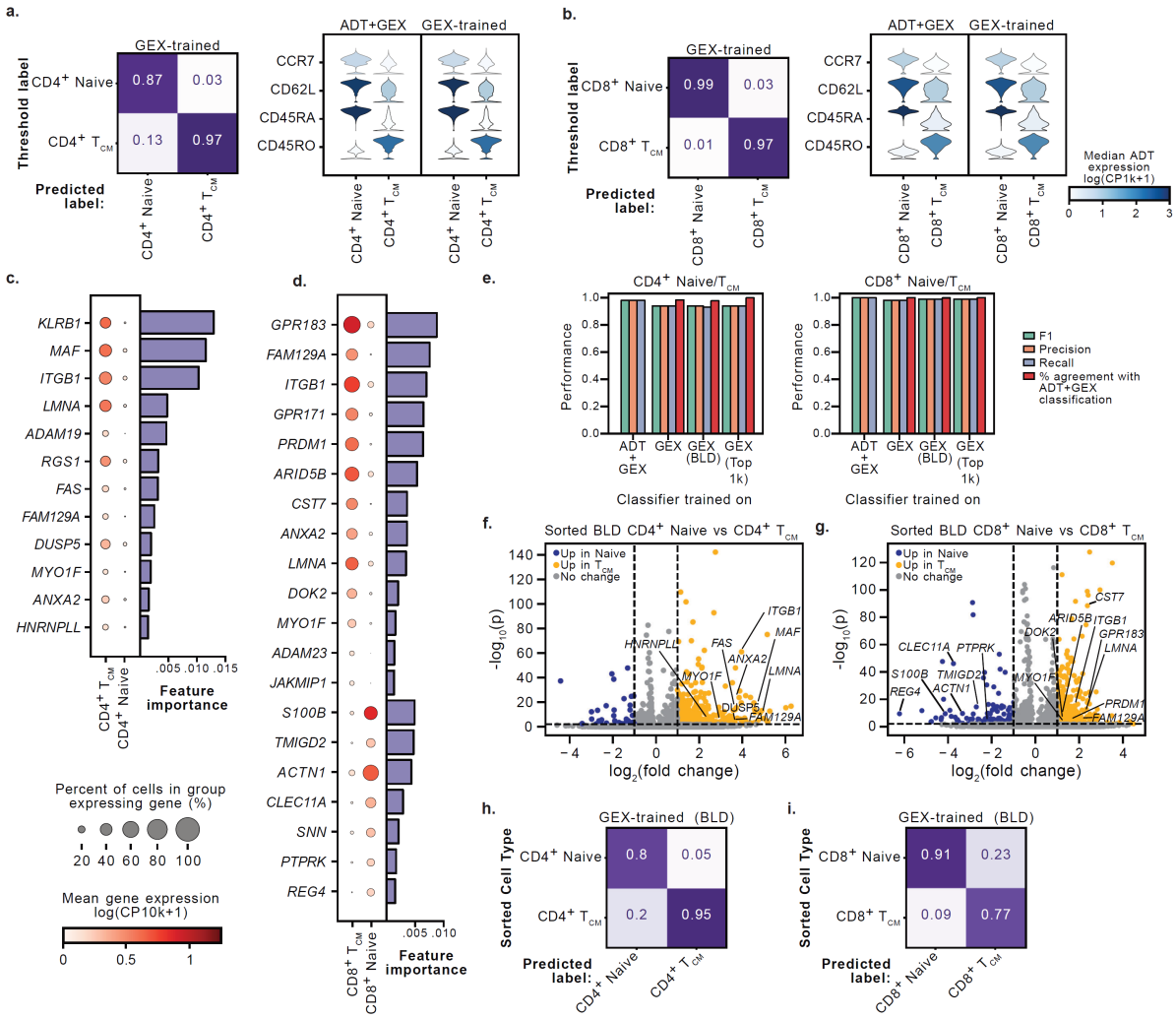


Figure 5 **MMoCHi reveals additional transcript markers distinguishing naive and central memory T cells.** a-b. Performance of gene expression (GEX)-only classification of Naive/T_{CM} $CD4^+$ (a) and $CD8^+$ (b) T cells. Column-normalized confusion matrices (left) comparing classification to held-out, 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 (right). All displayed features were significantly differentially expressed ($p < 0.05$). e. Bar plots comparing performance of Naive/T_{CM} $CD4^+$ and $CD8^+$ T cell classifiers trained using multimodal data (ADT+GEX), GEX only, GEX with only BLD cells, or only the top 1000 important GEX features. f-g. Volcano plots displaying differential gene expression between sorted blood Naive and T_{CM} $CD4^+$ (f) or $CD8^+$ (g) T cells. Points represent genes and are colored by differential

expression ($|\log_2(\text{fold change})| > 1$ and $P < 0.05$). Important features for classification that were differentially expressed are highlighted. h-i. Column-normalized heatmaps displaying performance of GEX only classifiers trained on organ donor BLD applied to sorted Naive/T_{CM} $CD4^+$ (h) and $CD8^+$ (i) T cells. Statistical significance was calculated using a two-sided Wilcoxon with tie correction, followed by a Benjamini–Hochberg adjustment for multiple comparisons. T_{CM}, central memory T cell.

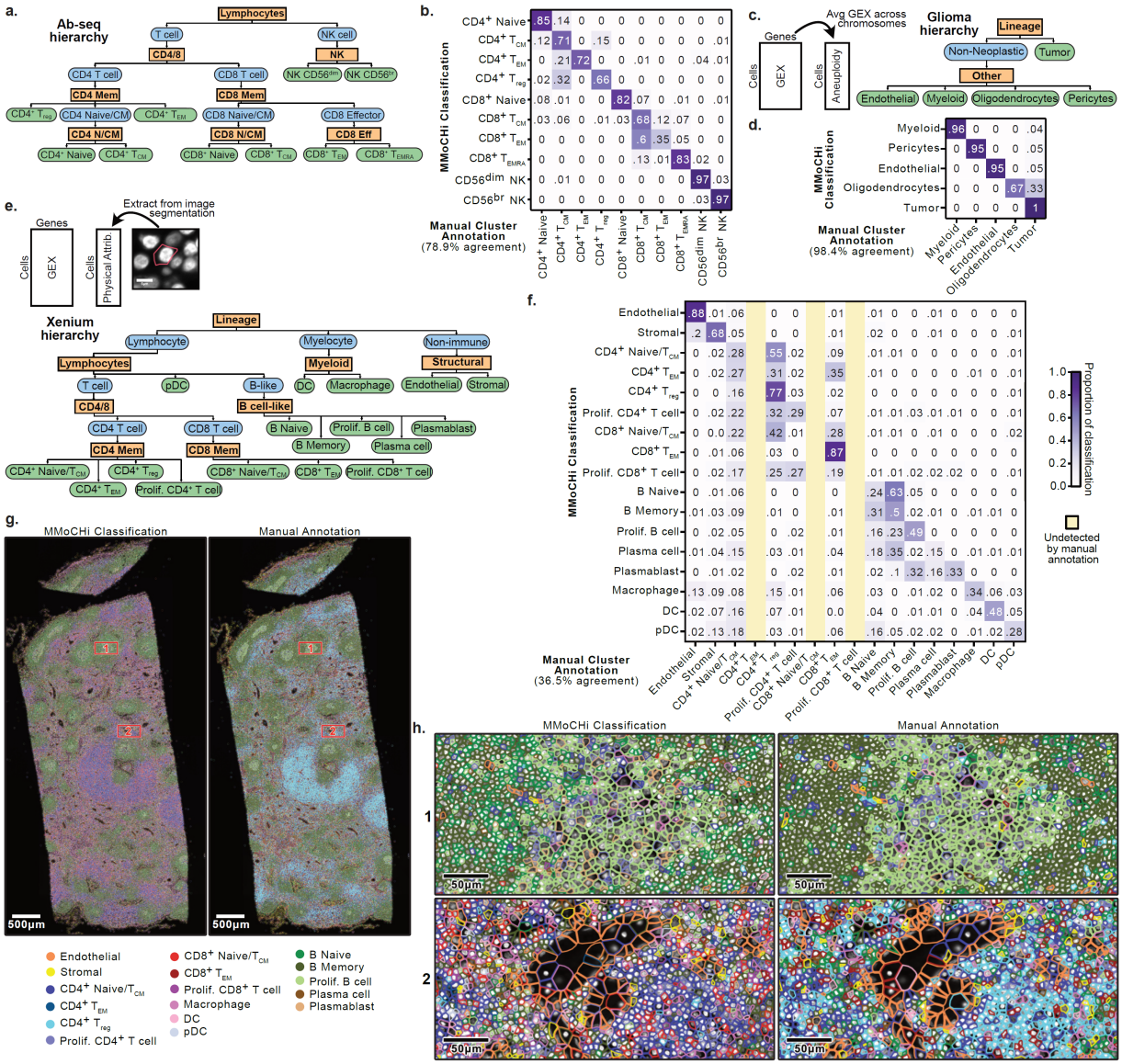
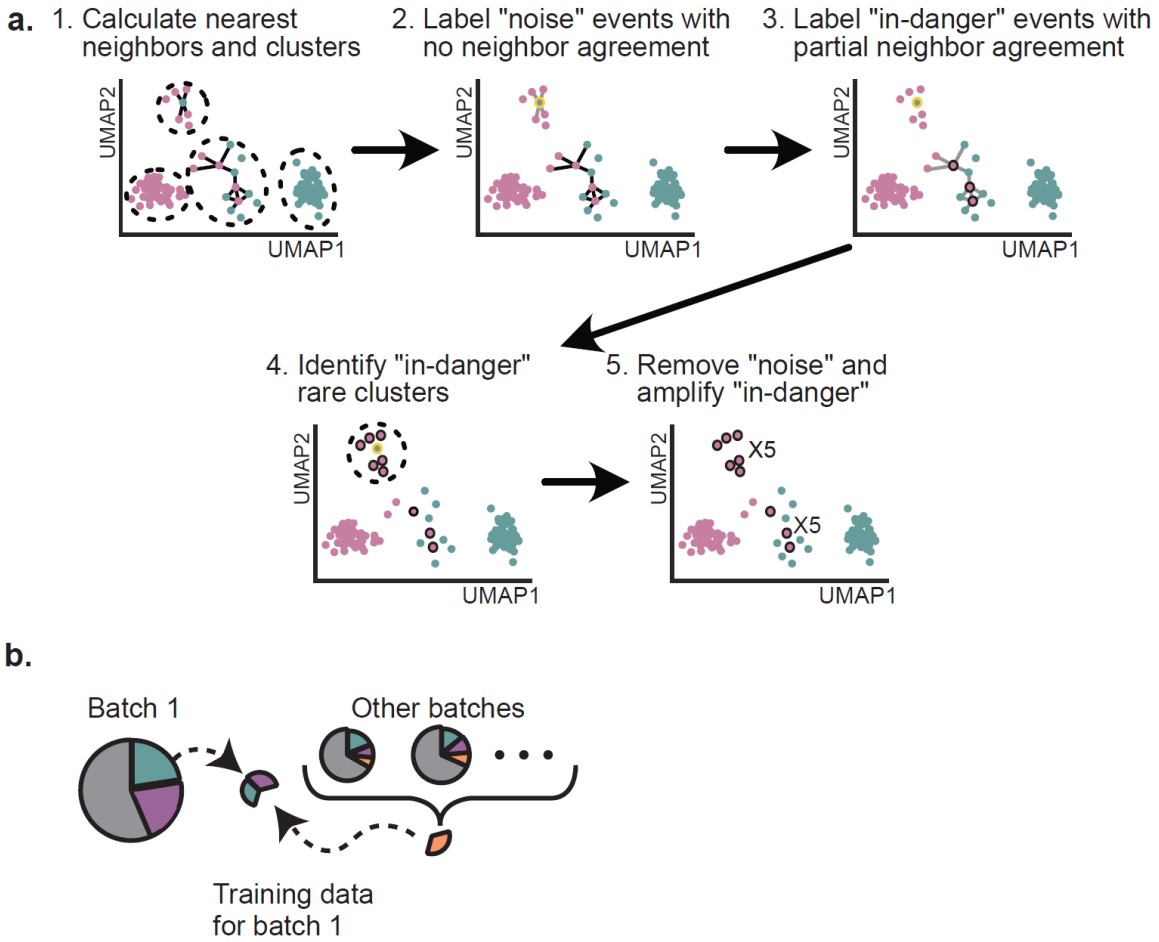
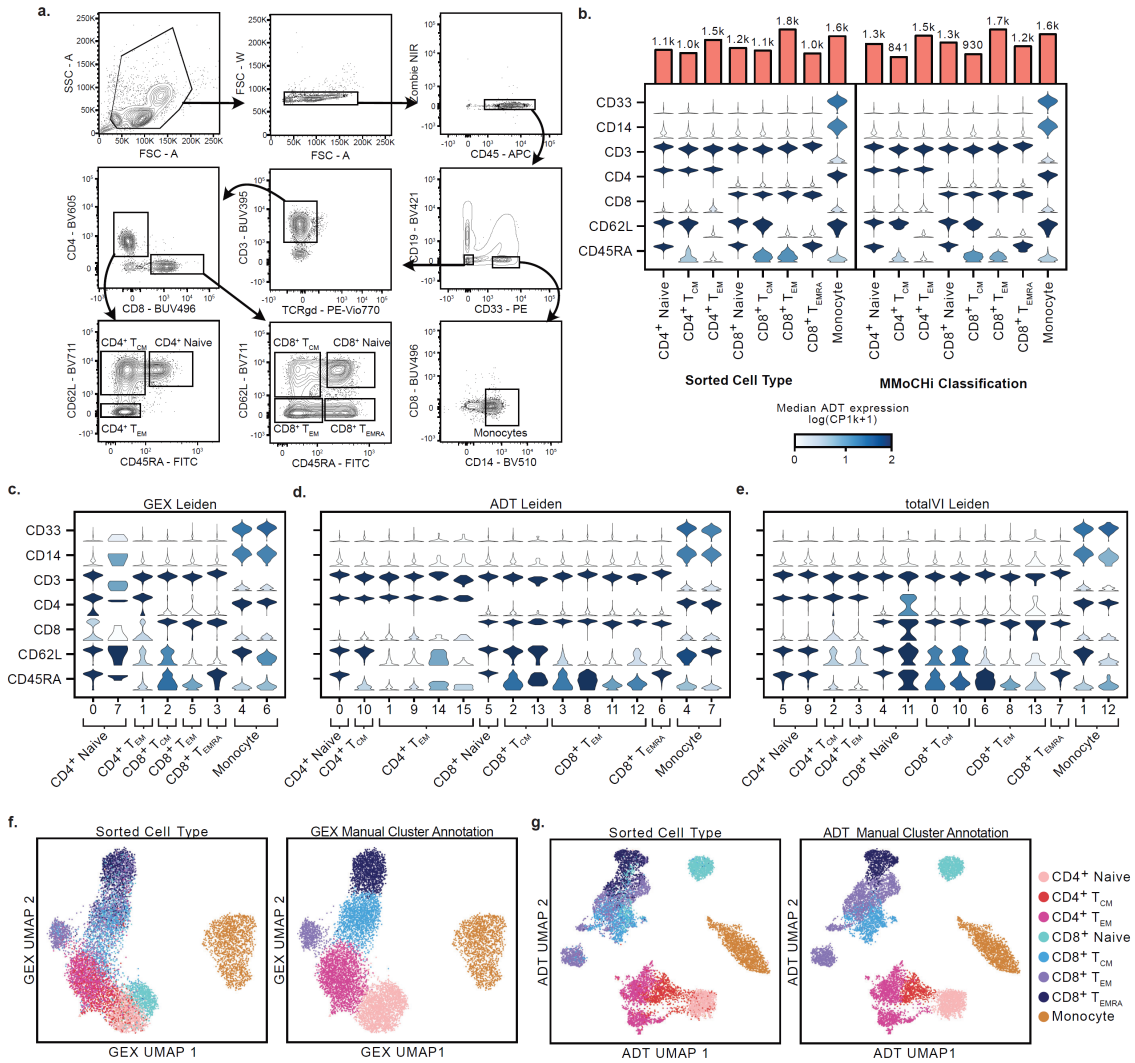


Figure 6 **MMoCHi applied to other multimodal datasets.** MMoCHi applied to Ab-seq data of sorted T and NK cells (a-b), a scRNA-seq of a high-grade glioma biopsy (c-d), or spatial transcriptomics of a human lymph node using the 10x Xenium platform (e-h). a, c, e. MMoCHi hierarchies used for classification (see Supplementary Table 17 for full specification) and schematics for extraction of multimodal data for classification (see Methods). b, d, f. Row-normalized heatmaps comparing MMoCHi classification to manual annotation. Color represents proportion of cells in each MMoCHi classified population that were manually annotated as each subset. Yellow columns indicate subsets that were not detected by manual annotation. g. Images displaying DAPI staining intensity on a human lymph node, overlaid with detected cell boundaries (see Methods). Cell boundaries are colored by MMoCHi classification or manual annotation. h. Magnified sections of the images in g, as outlined by the red boxes.

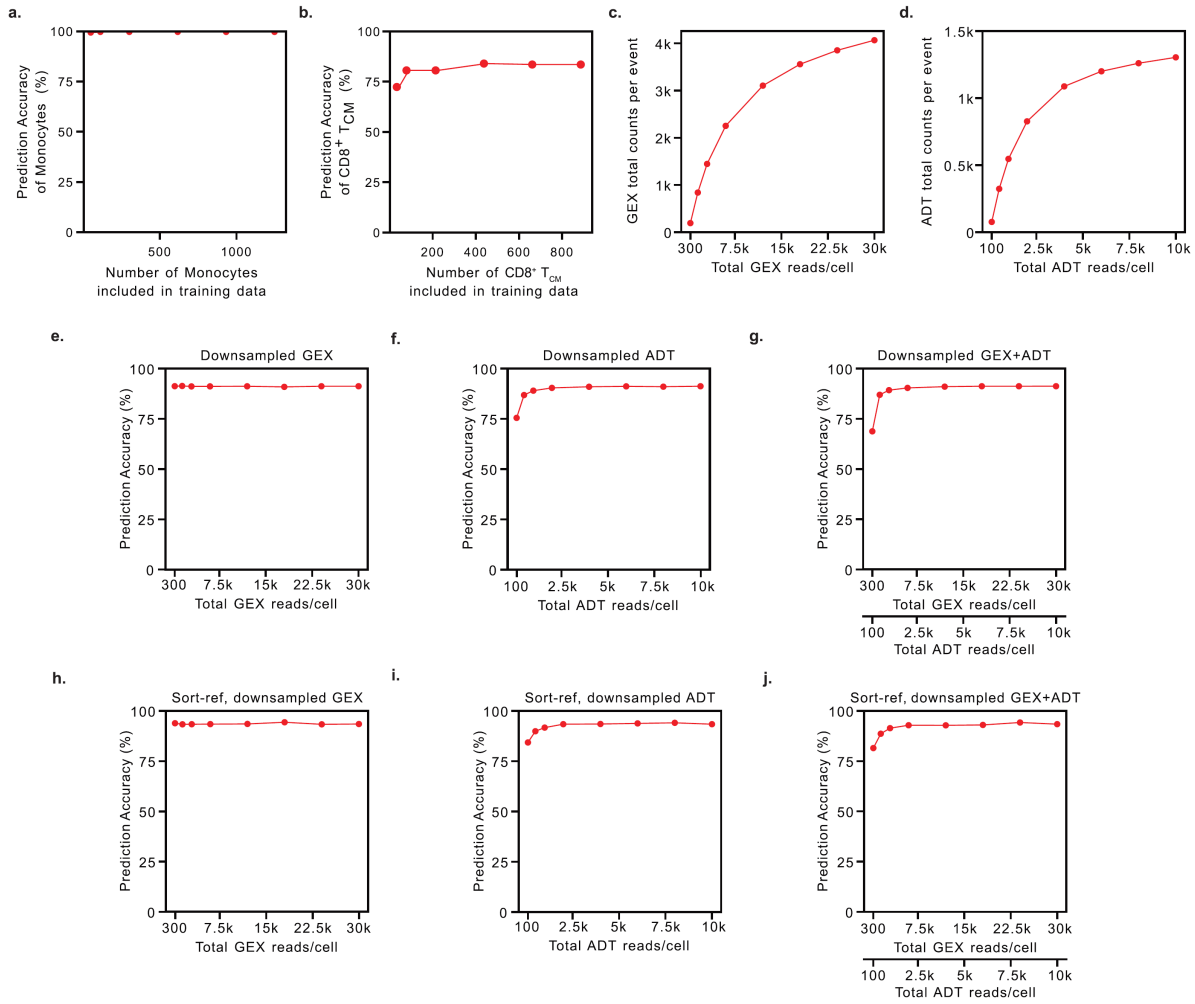
EXTENDED DATA FIGURES



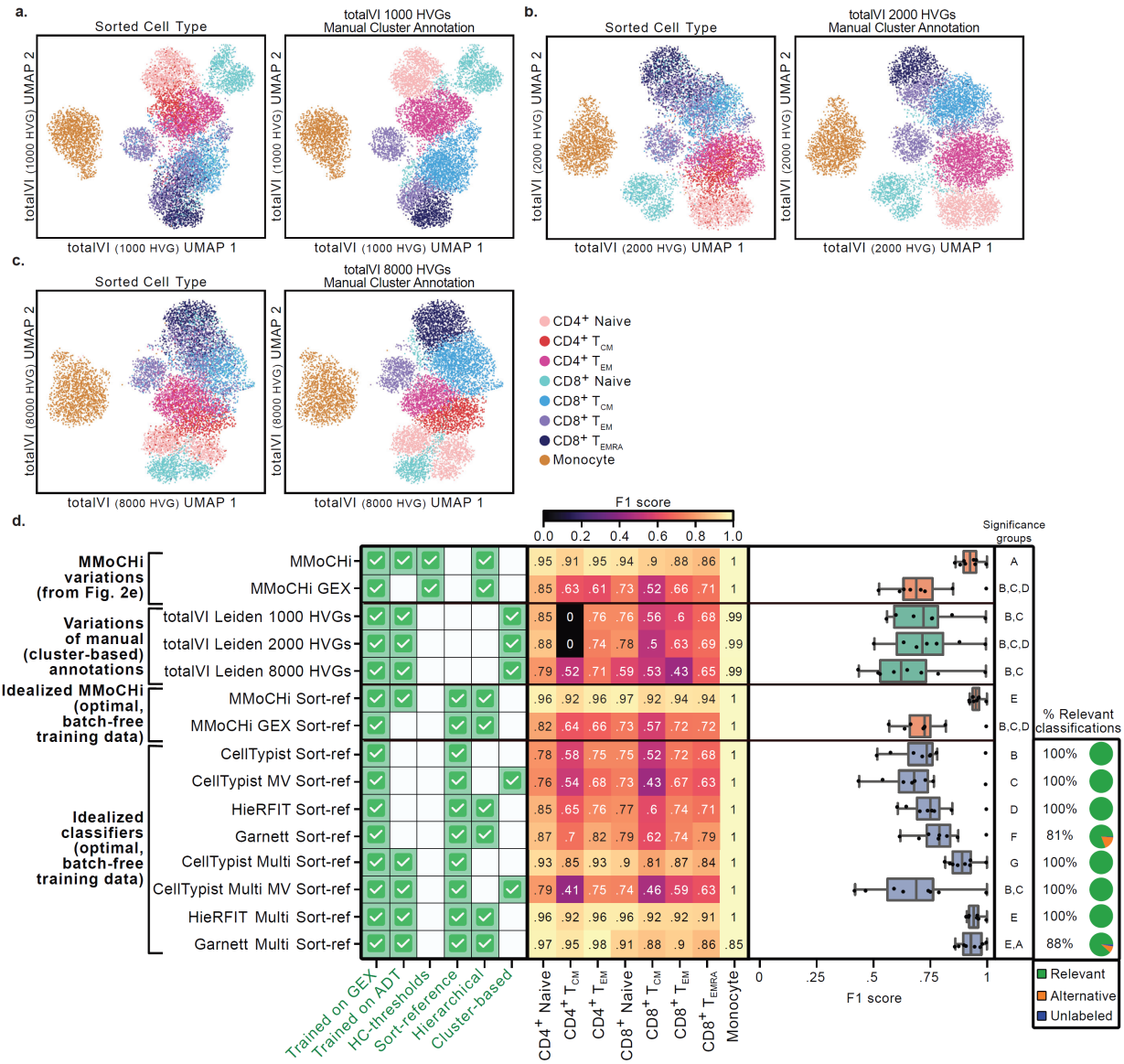
Extended Data Figure 1 **Selection and cleanup of training data.** a. Prior to training, high-confidence events were resampled to remove potentially mislabeled events ("noise") and amplify potentially challenging to classify events ("in danger"). Nearest neighbors are calculated on high-confidence events (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 events 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. Of $CD45^+$ cells 6.58% were $CD4^+$ Naive, 14.1% were $CD4^+$ T_{CM}, 5.33% were $CD4^+$ T_{EM}, 7.01% were $CD8^+$ Naive, 3.45% were $CD8^+$ T_{CM}, 5.14% were $CD8^+$ T_{EM}, 4.52% were $CD8^+$ T_{EMRA}, and 17.8% were 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.

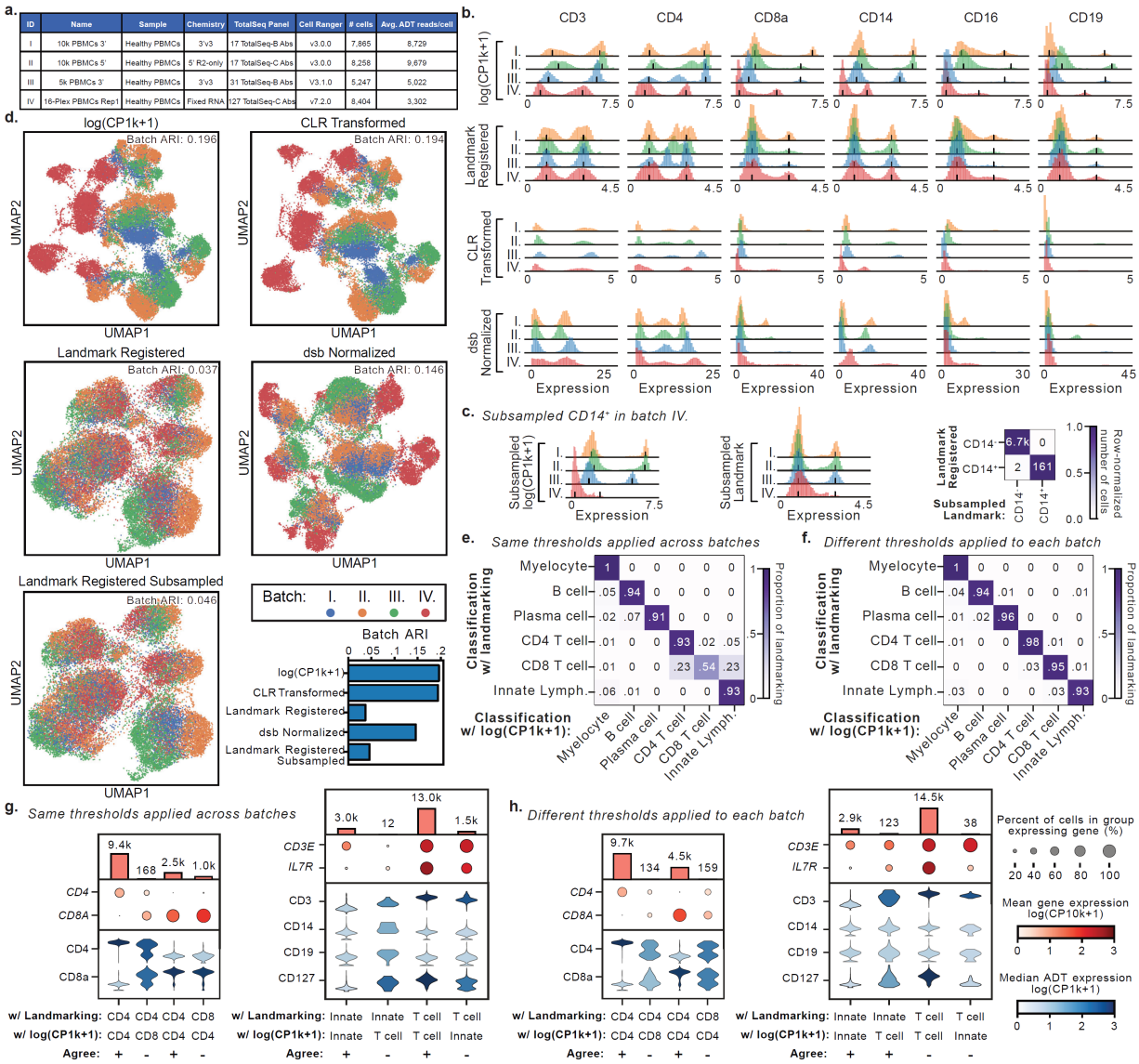


Extended Data Figure 3 **MMoCHi classification is robust to subsampled training data or downsampled reads.** a-b. Classification accuracy of a monocytes (a) or CD8$^+$ TCM following subsampling of each subset prior to training data selection (only 40-50% of which are identified by high-confidence thresholding). Sorted cell type labels (derived from hashtag-oligo; HTO) were used as "truth" and accuracy was calculated using a held-out data. c-d. Line plots depicting the effect of downsampling aligned reads of gene expression (GEX; c) or antibody derived tag (ADT) expression (d) on the average total counts in the expression matrix per event. c-h. Balanced classification accuracy of sorted T cell memory subsets and monocytes, with MMoCHi classifiers trained using high-confidence thresholding on downsampled events (e-g) or trained using the HTO-derived sort labels as reference and tested on held-out data (h-j). Classifiers were exposed to downsampled GEX and normal ADT expression (e,h), normal GEX and downsampled ADT expression (f,i), or downsampled GEX and ADT expression (g,j) for both training and prediction.



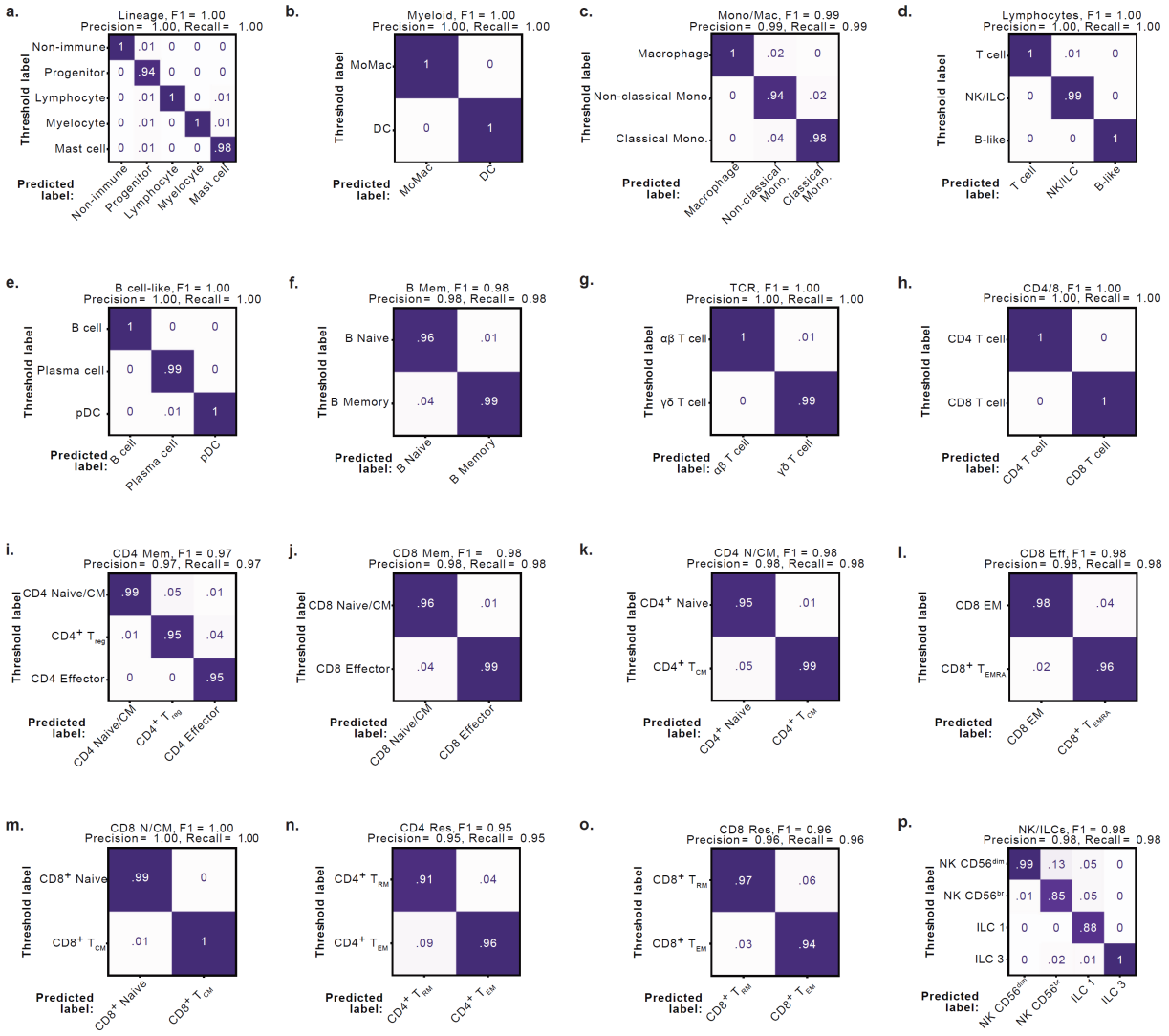
Extended Data Figure 4 **MMoCHi benchmarked against other variations and controls.** a-c. UMAPs of totalVI embeddings calculated using various numbers of most highly variable genes colored by sorted cell type (left) or manual cluster-based annotation (right). d. Performance comparison using F1 scores, 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 (right). Features of each method are labeled for context (left) using green checks: “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 HTO-derived sorted cell type labels were used to train the classifier. "Hierarchical"—whether annotation was performed on multiple levels. “Cluster-based”—whether annotations were applied to unsupervised clusters. For scRNA-seq classifiers, the percent of classifications that were relevant to this analysis, were alternative (but potentially valid) classifications or were unlabeled are shown using pie charts (far right). Statistical significance was calculated using a Friedman rank-sum test, matched by subset,

followed by multiple comparisons using paired Wilcoxon signed-rank tests and FDR correction. Significance groups for each method are shown by lettering, where methods not sharing a letter are significantly different ($p < 0.05$).

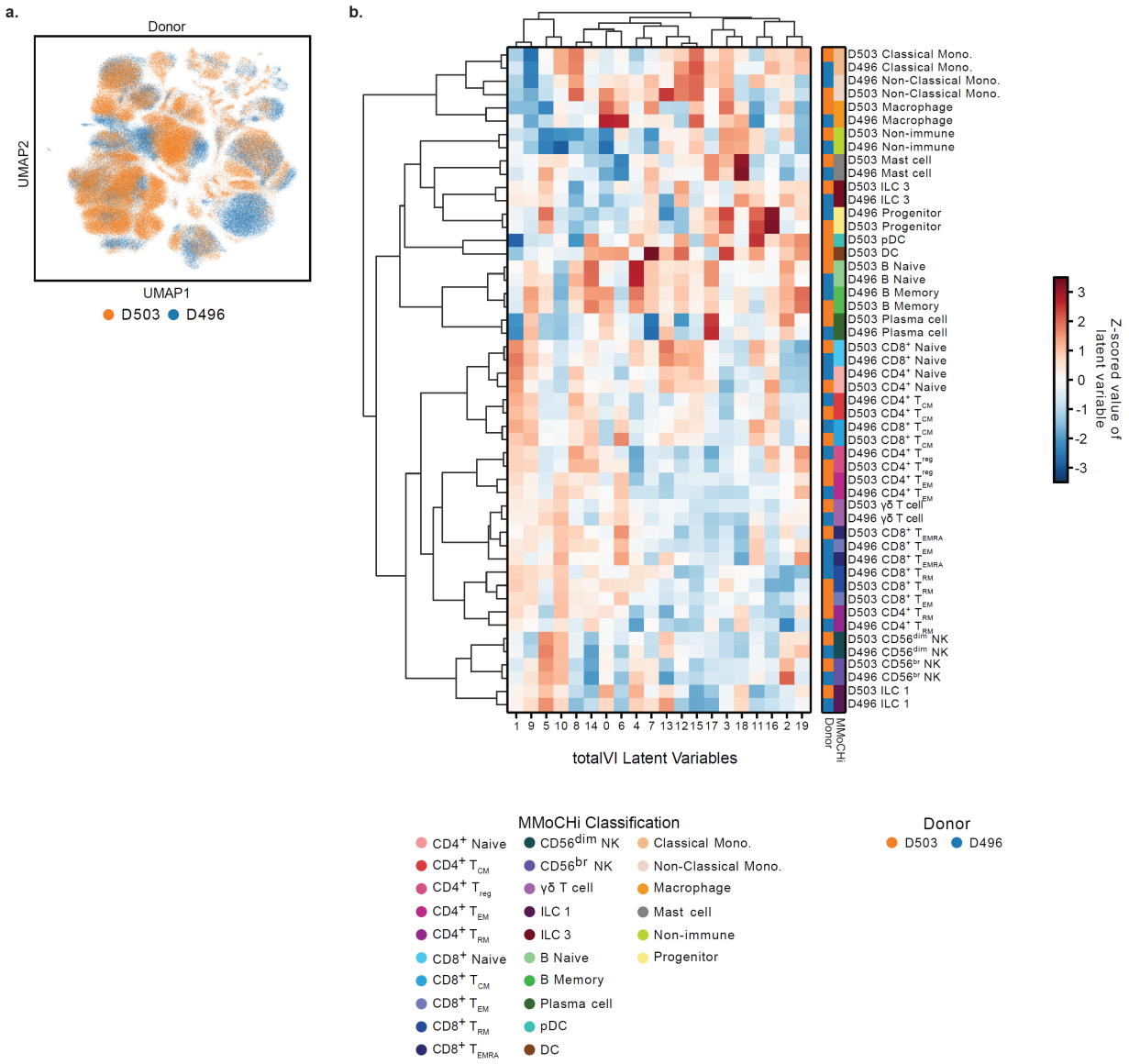


Extended Data Figure 5 **Landmark registration integrates ADT expression and allows for consistent thresholding.** a. Metadata for each dataset highlighting features that may contribute to batch effects. b. Histograms displaying normalized Antibody Derived Tag (ADT) expression distribution for selected proteins. Histograms for each batch are grouped by normalization technique. Peaks identified for landmark registration are displayed on the log(CP1k+1)-transformed expression. Aligned peaks are displayed on the landmark registered expression. c. Histograms as in (b) with the number of $CD14^+$ cells in batch IV subsampled to 5%. Heatmap displaying the number of $CD14^+$ cells by landmark-registration before and after subsampling. d. UMAPs of the expression of the 16 ADTs in all panels, colored by batch to display integration. Adjusted Rand Index (ARI) measures similarity of clustering and batch identity, where lower values indicate stronger integration. e-f. Row-normalized heatmaps displaying MMoCHi classifications (see Supplementary Table 7 for hierarchy specification) when run using landmark-registered or log(CP1k+1)-transformed ADT expression. g-h. Plots depicting expression of selected markers on populations classified by MMoCHi with or without landmark registration. Dot plots display gene expression. Dot size represents the percent of cells in the group expressing a

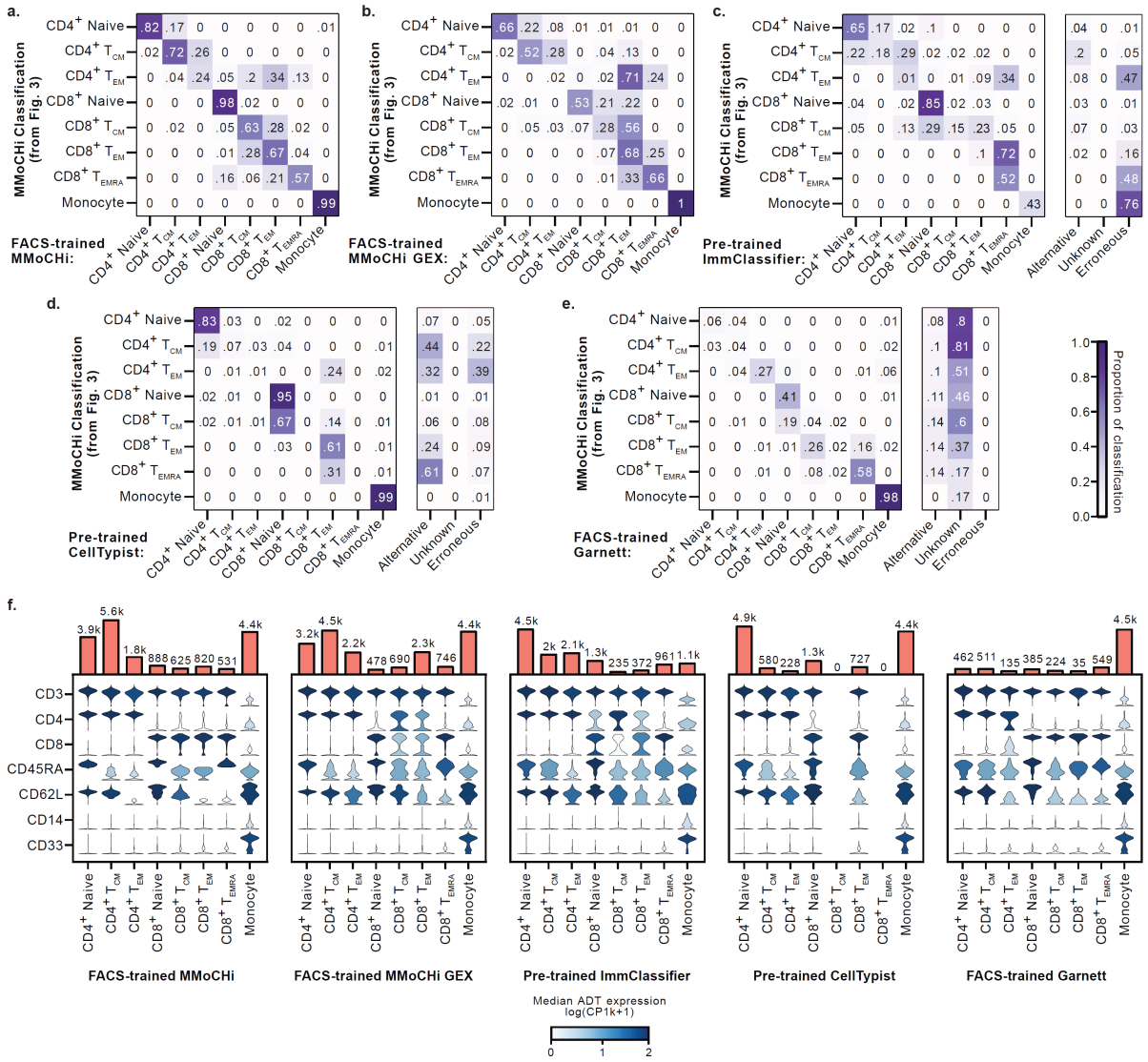
gene, and dots are colored by the mean log-normalized GEX counts per ten thousand. Violin plots display the distribution of antibody derived tag (ADT) expression for each population. Violins are colored by the median log-normalized ADT counts per thousand. The number of cells in each group are displayed above the dot plots. Populations are denoted by a "+" where annotations agree, and a "-" where they disagree. MMoCHi was run twice, with either the same high-confidence thresholds defined across all batches (e,g), or separately for each batch (f,h). CLR: Centered Log-Ratio; dsb: denoised and scaled by background.



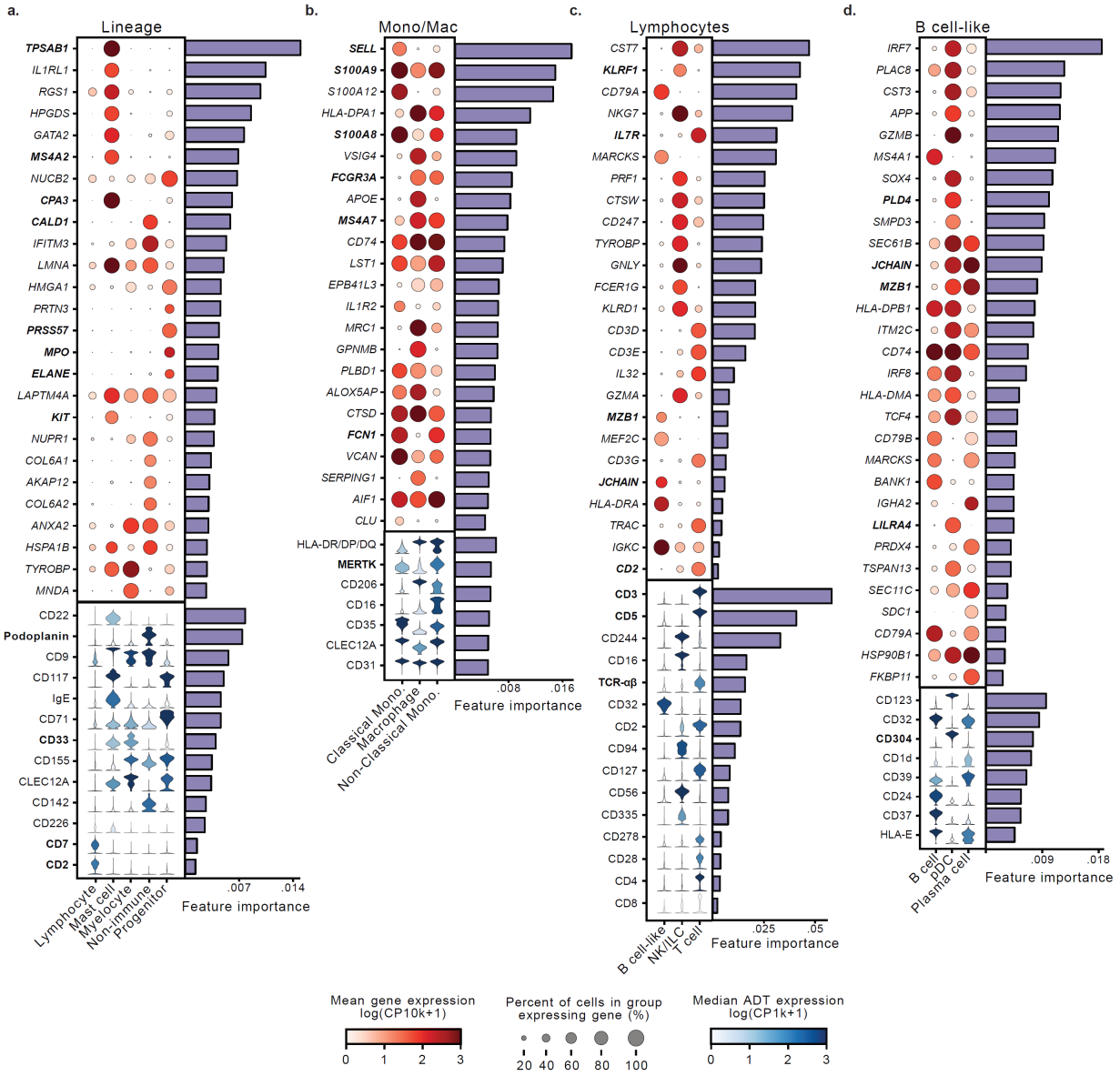
Extended Data Figure 6 **Random forests perform well at each level of MMoCHi hierarchy.** a-p. Column-normalized confusion matrices depicting the fit of each random forest classifier in the organ-donor hierarchy (Fig. 3c; Supplementary Table 8). MMoCHi classifications are compared with internally held-out high-confidence threshold labels at each node. 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 7 **MMoCHi classifies cell types across donor.** a. UMAP of donor-integrated totalVI latent space, colored by donor. b. Z-scored expression of totalVI latent variables averaged across each classified cell type in each donor (filtered by donor-subset pairs with at least 100 events). Dendrogram clustering of cell type-donors (left) and factors (top) calculated using Pearson correlation on the totalVI latent space. Rows are labeled with color by donor (left) and MMoCHi classification (right). 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.

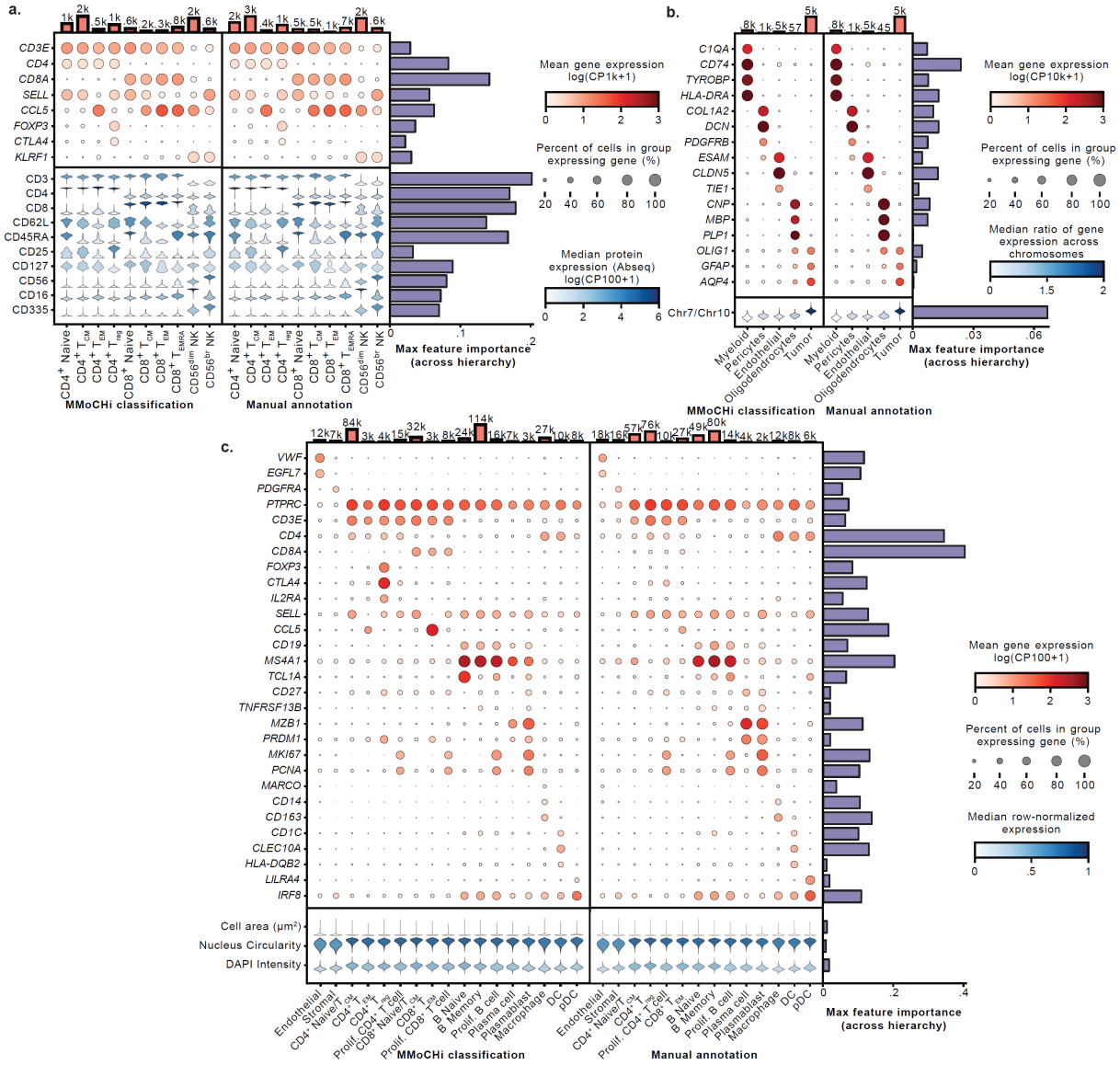


Extended Data Figure 8 **Pre-trained MMoCHi classifiers effectively annotate new data.** a–e. Row-normalized heatmaps displaying agreement between pre-trained classifiers and the MMoCHi classification obtained in Fig. 3, subset to only T cell memory subsets and monocytes from blood. Classifiers applied include: MMoCHi trained on sorted blood (FACS-trained; a), MMoCHi trained on sorted blood only using gene expression (FACS-trained GEX; b), ImmClassifier using the provided pre-trained model (c), Celltypist using a provided pre-trained model (d), and Garnett trained on sorted blood (e). Alternative (but potentially valid), Unknown (unlabeled), and Erroneous (cell types not present in sampled dataset) are indicated separately. f. Plots depicting expression of selected cell type markers, on cells grouped by their classification. 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 violin plots. T_{CM}, central memory T cell; T_{EM}, effector memory T cell; T_{EMRA}, terminally differentiated effector memory T cell.



Extended Data Figure 9 **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) nodes in the organ donor hierarchy (Fig. 3c). The top important features associated with each subset ($\log_2$(fold change) > 1.5) are displayed to include representation of features specific to each subset. All features displayed are within the top 1% of important features. Features used for high-confidence thresholding are highlighted in bold. 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. Mac, Macrophage; Mono, monocyte; NK, natural killer cell; ILC, innate lymphoid cell; pDC, plasmacytoid dendritic cell.



Extended Data Figure 10 **Marker gene and protein expression on manually annotated and MMoCHi classified populations.** a-c. Expression of select cell type markers on cells grouped by either their MMoCHi classification (left) or their manual cluster annotation (right) for the Ab-seq (a), Glioma (b), and Xenium (c) datasets. 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. Violin plots display the distribution of protein expression (a), the ratio of gene expression across chromosomes (b), or row-normalized physical attributes of cells (c), for each MMoCHi-classified or manually-annotated population. Violins are colored by the median value. T_{CM}, central memory T cell; T_{reg}, regulatory T cell; T_{EM}, effector memory T cell; T_{EMRA}, terminally differentiated effector memory T cell, NK, natural killer cell; DC, dendritic cell; pDC, plasmacytoid dendritic cell; Prolif., Proliferating.

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. Shahi, P., Kim, S. C., Haliburton, J. R., Gartner, Z. J. & Abate, A. R. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci Rep* **7**, 44447 (2017).
8. Cheng, C., Chen, W., Jin, H. & Chen, X. A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell–Cell Communication. *Cells* **12**, (2023).
9. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24**, 550–572 (2023).
10. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
11. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).
12. 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).
13. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
14. Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* **18**, 272–282 (2021).
15. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* **7**, 1141 (2020).
16. 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).
17. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
18. 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).
19. Kaymaz, Y. *et al.* HieRFIT: a hierarchical cell type classification tool for projections from complex single-cell atlas datasets. *Bioinformatics* **37**, 4431–4436 (2021).
20. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* **16**, 983–986 (2019).
21. 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).
22. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 31 (2020).
23. Osumi-Sutherland, D. *et al.* Cell type ontologies of the Human Cell Atlas. *Nat Cell Biol* **23**, 1129–1135 (2021).

24. 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).
25. Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* **16**, 1007–1015 (2019).
26. Fang, P. *et al.* Immune cell subset differentiation and tissue inflammation. *J Hematol Oncol* **11**, 97 (2018).
27. Gattinoni, L. *et al.* A human memory T cell subset with stem cell–like properties. *Nat Med* **17**, 1290–1297 (2011).
28. 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).
29. Thome, J. J. C. *et al.* Spatial Map of Human T Cell Compartmentalization and Maintenance over Decades of Life. *Cell* **159**, 814–828 (2014).
30. 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).
31. 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).
32. 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).
33. 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).
34. Hahne, F. *et al.* Per-channel basis normalization methods for flow cytometry data. *Cytometry A* **77**, 121–131 (2010).
35. 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).
36. Staats, J., Divekar, A., McCoy, J. P. & Maecker, H. T. Guidelines for Gating Flow Cytometry Data for Immunological Assays. in *Immunophenotyping: Methods and Protocols* (ed. McCoy, J., J. Philip) 81–104 (Springer, New York, NY, 2019). doi:10.1007/978-1-4939-9650-6_5.
37. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053–1058 (2018).
38. Mulè, M. P., Martins, A. J. & Tsang, J. S. Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nat Commun* **13**, 2099 (2022).
39. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
40. Chen, X., Wang, M. & Zhang, H. The use of classification trees for bioinformatics. *WIREs Data Mining and Knowledge Discovery* **1**, 55–63 (2011).
41. 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).
42. Evren, E. *et al.* Distinct developmental pathways from blood monocytes generate human lung macrophage diversity. *Immunity* **54**, 259-275.e7 (2021).
43. Karnell, J. L. *et al.* CD19 and CD32b Differentially Regulate Human B Cell Responsiveness. *J Immunol* **192**, 1480–1490 (2014).
44. 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).
45. Mason, D. Y. *et al.* CD79a: A Novel Marker for B-Cell Neoplasms in Routinely Processed Tissue Samples. *Blood* **86**, 1453–1459 (1995).
46. Xu, C. *et al.* MARCKS regulates tonic and chronic active B cell receptor signaling. *Leukemia* **33**, 710–729 (2019).

47. 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).
48. 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).
49. Trzupek, D. *et al.* Single-cell multi-omics analysis reveals IFN-driven alterations in T lymphocytes and natural killer cells in systemic lupus erythematosus. *Wellcome Open Res* **6**, 149 (2022).
50. Levitin, H. M. *et al.* De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Molecular Systems Biology* **15**, e8557 (2019).
51. Tosolini, M. *et al.* Assessment of tumor-infiltrating TCRVγ9Vδ2 γδ lymphocyte abundance by deconvolution of human cancers microarrays. *Oncoimmunology* **6**, e1284723 (2017).
52. 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).
53. 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).
54. Mimitou, E. P. *et al.* Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat Methods* **16**, 409–412 (2019).
55. Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* **10**, e63632 (2021).
56. 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).
57. 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).
58. Reimegård, J. *et al.* A combined approach for single-cell mRNA and intracellular protein expression analysis. *Commun Biol* **4**, 1–11 (2021).
59. Takei, Y. *et al.* Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590**, 344–350 (2021).
60. 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).
61. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).
62. 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, Bonn, Germany, 2005). doi:10.1145/1102351.1102430.
63. 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).
64. 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).
65. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).
66. Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* **39**, 813–818 (2021).
67. Melsted, P., Ntranos, V. & Pachter, L. The barcode, UMI, set format and BUStools. *Bioinformatics* **35**, 4472–4473 (2019).
68. Yuan, J. & Sims, P. A. An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. *Sci Rep* **6**, 33883 (2016).
69. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet* **23**, 169–181 (2022).

70. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
71. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
72. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nat Med* **27**, 904–916 (2021).
73. Bredikhin, D., Kats, I. & Stegle, O. MUON: multimodal omics analysis framework. *Genome Biology* **23**, 42 (2022).
74. 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).
75. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).