nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-00896-6

Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID

Akira Cortal[1], Loredana Martignetti[1], Emmanuelle Six[2] and Antonio Rausell[1,3] ✉

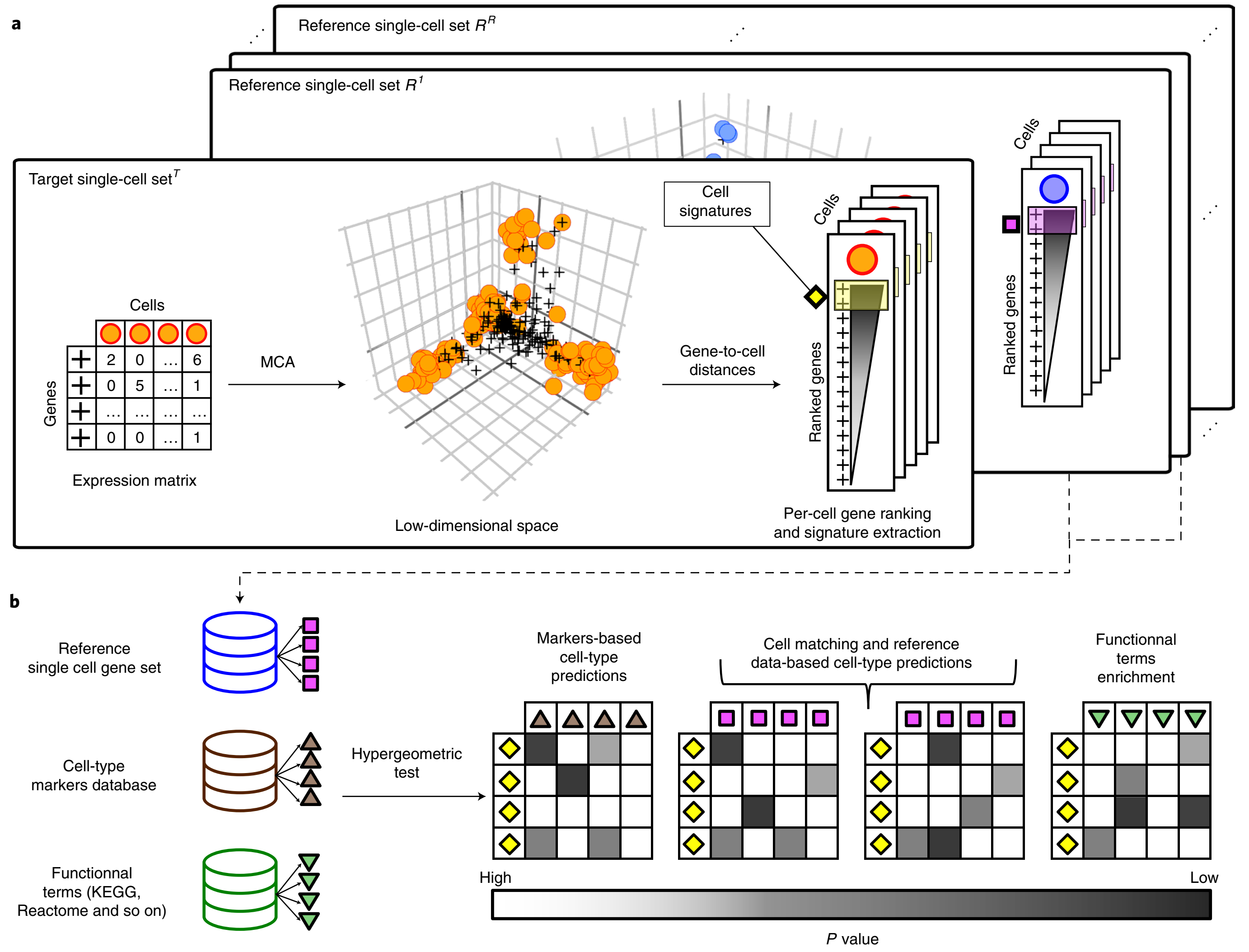
Because of the stochasticity associated with high-throughput single-cell sequencing, current methods for exploring cell-type diversity rely on clustering-based computational approaches in which heterogeneity is characterized at cell subpopulation rather than at full single-cell resolution. Here we present Cell-ID, a clustering-free multivariate statistical method for the robust extraction of per-cell gene signatures from single-cell sequencing data. We applied Cell-ID to data from multiple human and mouse samples, including blood cells, pancreatic islets and airway, intestinal and olfactory epithelium, as well as to comprehensive mouse cell atlas datasets. We demonstrate that Cell-ID signatures are reproducible across different donors, tissues of origin, species and single-cell omics technologies, and can be used for automatic cell-type annotation and cell matching across datasets. Cell-ID improves biological interpretation at individual cell level, enabling discovery of previously uncharacterized rare cell types or cell states. Cell-ID is distributed as an open-source R software package.

High-throughput single-cell technologies, such as single-cell RNA-sequencing (scRNA-seq), are currently being used for the complete cellular characterization of human tissues and cell types. The Human Cell Atlas project[1], the National Institutes of Health (NIH) Human BioMolecular Atlas Program (HuBMAP)[2] and the LifeTime[3] initiative are remarkable examples of current scientific ambitions in this direction. One of the main goals of these studies is the identification of previously unknown cell types or cell states with potential physiological roles in health and disease. However, the computational characterization of cell heterogeneity is rendered more complex by the high dimensionality and high levels of technical and biological noise associated with single-cell measurements[4]. One common strategy for enhancing the signal-to-noise ratio involves the use of a low-dimensional representation of cells from which the most salient relative differences may emerge[5]. The techniques most widely used for this purpose include principal component analysis (PCA), independent component analysis, *t*-distributed stochastic neighbor embedding (*t*-SNE) and uniform manifold approximation and projection (UMAP)[6]. Molecular characterization of the observed cell heterogeneity is then typically performed by analyzing differential gene expression between the groups of cells resulting from a clustering step[7]. However, this reliance on cell clusters results in gene signatures being assigned at cell subpopulation level rather than at full single-cell resolution. One of the main limitations of cluster-based approaches is that the gene signature analysis is bound to a certain level of resolution, at which the cell heterogeneity is partitioned into nonoverlapping classes[8]. An exhaustive exploration of transcriptional heterogeneity requires a statistically robust per-cell gene signature assessment for each cell in a dataset .

Results

The Cell-ID method. We present here Cell-ID, a multivariate approach to extracting a gene signature for each individual cell in a study (Fig. 1 and Methods). Cell-ID is based on multiple correspondence analysis (MCA), a statistical technique that provides a simultaneous representation of observations (for example, cells) and variables (for example, genes) in low-dimensional space[9–12]. Originally, MCA was applied to binary and fuzzy-coded data, and this has been so far an obstacle for its use on omics data. We describe here a linear transformation of gene expression values in a continuous scale between 0 and 1 that allows unlocking the use of this technique for quantitative single-cell data analysis, while keeping the mathematical properties of MCA. The MCA representation is a particular type of a general concept formally known as biplot[13,14] (Supplementary Note 1). Here a target matrix $M_{N,P}$ representing, for example, the transformed expression levels of N cells and P genes, is optimally approximated through singular-value decomposition as the product of two matrices, that is $M_{NP} = \Phi_{NJ} G_{JP}^{T}$, where Φ_{NJ} and G_{PJ} represent the coordinates of the N cells and the P genes, respectively, in a J dimensional space. The two sets of J dimensions on which Φ_{NJ} and G_{PJ} are expressed, are colinear; that is, they convey the same directions in a common vector space. Furthermore, in the MCA biplot a barycentric relation exists between the cell and the gene projections by which the coordinate of a gene k in a dimension j corresponds to the weighted average of the N cell coordinates in dimension j (that is, its center of mass), where cell weights are given by their relative expression conditioned on the gene k (Methods). The barycentric relation among cells and genes is a distinctive feature of MCA biplots and represents a big advantage as compared to other types of biplot such as those produced by PCA (Supplementary Note 1) as well as over alternative low-dimensional transformations providing only cell projections. Thus, in the MCA biplot, analytical distances can be calculated not only between cells and between genes, but also between each cell and each gene to estimate its association (Fig. 1a). Thus, the closer a gene g is to a cell c, the more specific to such a cell it can be considered. Gene-to-cell distances can then be ranked for each individual cell, and the top-ranked genes may be regarded as a unique gene signature representing the identity card of the cell (Fig. 1a). We show here that the unbiased per-cell gene signatures extracted by Cell-ID reproduce the gene signatures previously established for well-known cell types. Cell-ID

[1]Clinical Bioinformatics Laboratory, Université de Paris, INSERM UMR1163, Imagine Institute, Paris, France. [2]Laboratory of Human Lymphohematopoiesis, Université de Paris, INSERM UMR1163, Imagine Institute, Paris, France. [3]Molecular Genetics Service, AP-HP, Necker Hospital for Sick Children, Paris, France. ✉e-mail: antonio.rausell@institutimagine.org


a
Reference single-cell set R^R
Reference single-cell set R^1
Target single-cell set ^{T}
Cells
Genes
Expression matrix
MCA
Low-dimensional space
Cell signatures
Gene-to-cell distances
Cells
Ranked genes
Per-cell gene ranking and signature extraction
b
Reference single cell gene set
Cell-type markers database
Functionnal terms (KEGG, Reactome and so on)
Hypergeometric test
Markers-based cell-type predictions
Cell matching and reference data-based cell-type predictions
Functionnal terms enrichment
High
Low
P value

	●	●	●	●
+	2	0	…	6
+	0	5	…	1
+	…	…	…	…
+	0	0	…	1

Fig. 1 | Overview of the Cell-ID approach. a, Cell-ID performs a dimensionality reduction of the gene expression matrix through MCA, where both cells and genes are projected in a common orthogonal space. In such space, the closer a gene is to a cell, the more specific it is to it. Thus, a gene ranking is obtained for each cell in a dataset based on their distance in the MCA space. The top-ranked genes for a given cell define its gene signature, which can be regarded as a unique cell identity card. Per-cell gene signatures can be independently extracted for a collection of single-cell datasets for downstream analyses. **b**, Per-cell gene signatures from a dataset can be evaluated for their enrichment against (1) collections of preestablished cell-type markers, to perform automatic cell-type annotation, (2) per-cell gene signatures from independent single-cell datasets, allowing cell matching and label transferring and (3) gene sets representing biological functions or molecular pathways, allowing functional annotation and interpretation of cell states.

signatures are, in turn, reproducible across independent datasets, despite strong batch effects, and can be used for cell identity tracing across different donors, model organisms, tissues of-origin and single-cell omics protocols.

Consistency of MCA representation of cells and genes. We first evaluated the consistency of MCA-based low-dimensional representations of cells and genes on 100 simulated scRNA-seq datasets (Supplementary Note 2). Consistency was demonstrated at three levels. The cell representation achieved by MCA dimensionality reduction was largely equivalent to that achieved by PCA on the same dataset: Spearman's correlation coefficients, for the correlation between MCA and PCA coordinates, with median and standard deviation values across datasets of 1.00 ± 0.02, 1.00 ± 0.02, 0.99 ± 0.02, 0.99 ± 0.02 and 0.99 ± 0.02 for their first five principal axes, respectively (Supplementary Fig. 1 and Supplementary Note 2). Second, the per-cell gene rankings provided by MCA are consistent with the expression values for the 50 neighboring cells in MCA space, as reflected by their log(fold change) in expression relative to the other cells (Supplementary Fig. 2a,b). Third, MCA-based per-cell gene rankings are robust to high levels of dropout events. Thus, genes with no expression detected in a cell may nevertheless rank highly for the cell concerned if more strongly expressed in the surrounding cells than in more distant cells (Supplementary Fig. 2a,b and Supplementary Note 2). Our results highlight the advantages of a multivariate approach in which per-cell gene assessments are implicitly weighted by their cell neighborhood in low-dimensional space. Such patterns would be missed if per-cell gene rankings were naively obtained either (1) from the log(fold change) in expression in the target cell relative to background cells (Supplementary Fig. 2c,d) or (2) from highest-to-lowest expression values within a cell, with random ranks for ties, as in another published approach (AUCell[15]) (Supplementary Fig. 2e,f).

Identification of cell types using reference marker lists. We also showed that Cell-ID could extract per-cell gene signatures, recovering characteristic marker lists associated with well-known cell types[16] (Supplementary Note 3). To this end, we used two independent sets of human blood mononuclear cells for which individual cells were confidently annotated with an actual cell type through concomitant measurements of single-cell protein marker levels: (1) cord blood mononuclear cells (CBMCs) profiled with a CITE-seq protocol[17]; and (2) peripheral blood mononuclear cells profiled with a REAP-seq protocol[18]. Cell-ID per-cell gene signatures were significantly enriched in the lists of markers associated with the corresponding cell type (Fig. 2a). This enrichment made it possible to recognize cell types with high rates of precision (87 and 90%) and recall (84 and 73%) for both datasets (multinomial $P < 10^{-16}$ for all

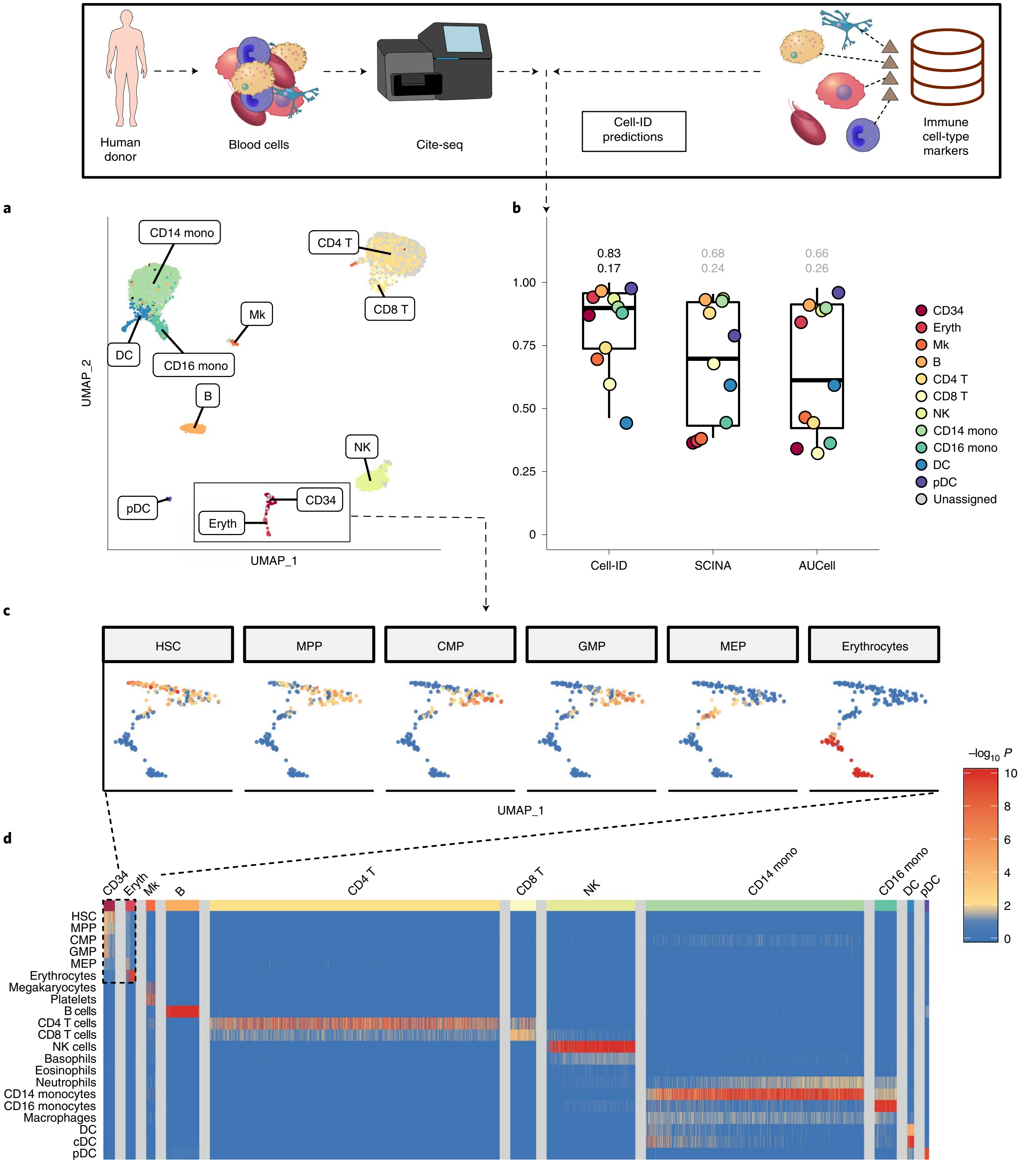


Fig. 2 | Cell-ID cell-type prediction of human CBMCs using preestablished marker lists. a, UMAP representation of $n = 8,005$ CBMCs profiled by CITE-seq[17]. Cells are colored according to Cell-ID cell-type predictions using preestablished immune cell signatures. **b**, F1 score achieved by Cell-ID, AUCell and SCINA cell-type predictions for each of the cell types reported in the original publication[17]. Boxplots summarize the F1 scores for each method and show the median (center line), interquartile range (hinges) and 1.5 times the interquartile range (whiskers). Numbers above boxplots denote the global performance (macro F1 score, upper digits) and its standard deviation (lower digits), where the maximum F1 and the minimum standard deviation values across methods are highlighted in black. **c**, Enlarged UMAP representation on Erythrocytes and CD34$^+$ cells showing that Cell-ID multi-class cell assignments capture transient cell states consistent with HSC differentiation hierarchy. Cells are color-coded according to the $-\log_{10}$ enrichment P value (after Benjamini–Hochberg correction for the number of signatures tested) obtained by Cell-ID, in tests of the association of their gene signature with the cell-type signatures associated to precursor cell types: HSC, MPP, CMP, GMP, MEP and erythrocytes. Color scale for cells extends from blue ($P = 1$) to dark red ($P = 1 \times 10^{-10}$). **d**, Heatmap representing, for each individual cell (displayed in columns), the $-\log_{10}$ transformed P value obtained by Cell-ID for each of the evaluated marker lists (displayed in rows). Heatmap color code extends from dark blue ($P = 1$) to yellow ($P = 10^{-2}$) to dark red ($P = 10^{-10}$), with $P < 10^{-10}$ fixed at this value). Nonsignificant associations ($P > 10^{-2}$ after Benjamini–Hochberg correction for the number of gene signatures tested) are shown in blue. Columns in the heatmap were grouped by the reference cell-type label. CD34, CD34$^+$ HSCs; Eryth, Erythrocytes; Mk, Megakaryocytes; B, B cells; CD4 T, CD4$^+$ T cells; CD8, CD8$^+$ T cells; CD14, CD14$^+$ monocytes; CD16 Mono, CD16$^+$ monocytes; NK, natural killer cells; DC, dendritic cells; cDC, conventional dendritic cells; pDC, plasmacytoid dendritic cells.

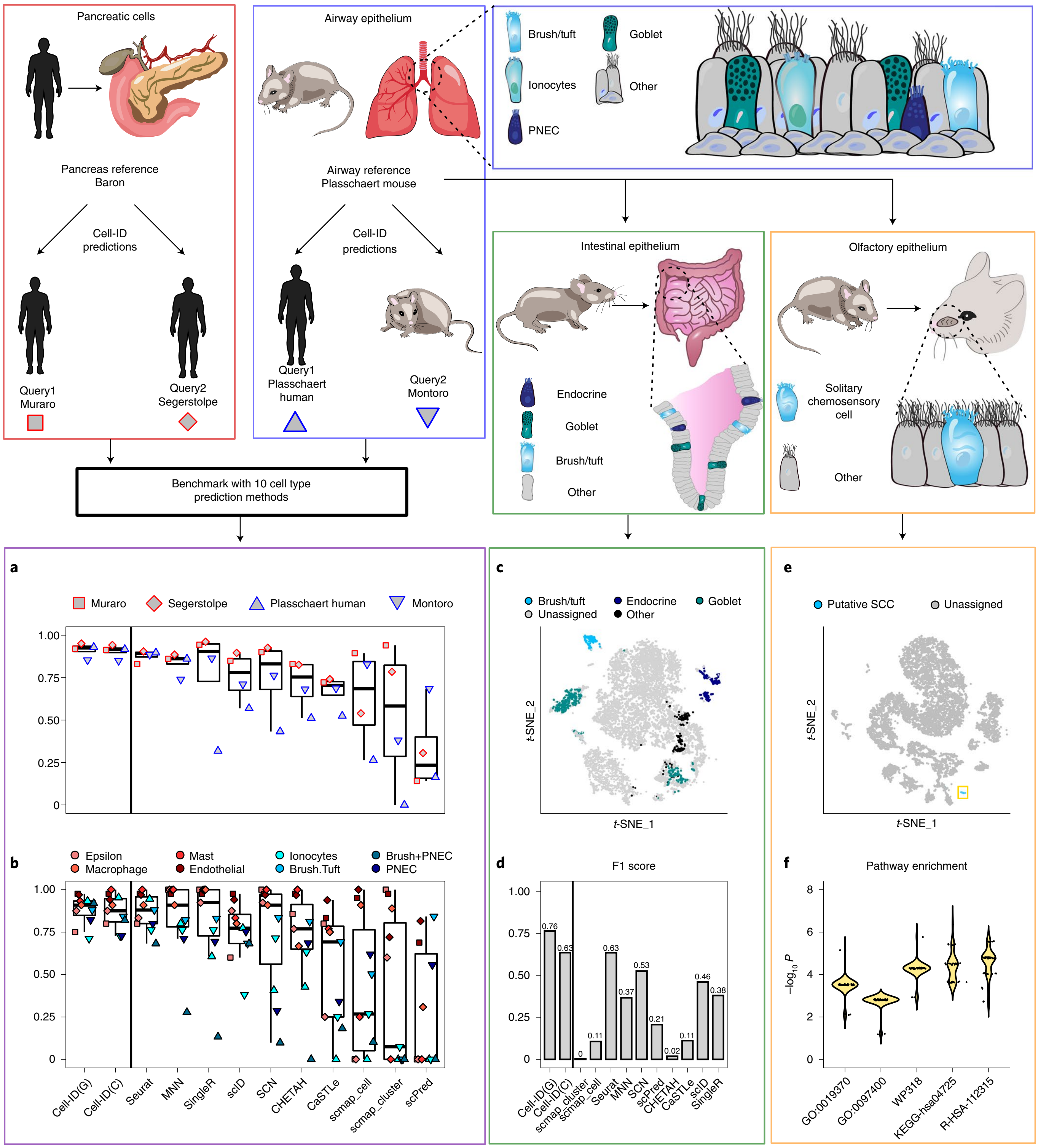
figures), outperforming reference methods for cell-type classification on the basis of marker lists (AUCell[15] and SCINA[19], Fig. 2b and Supplementary Figs. 3 and 4). In more challenging scenarios, Cell-ID was capable of nondisjoint multi-class cell-type assignments capturing smooth transitions between hematopoietic differentiation states from the most immature hematopoietic stem cells (HSC) to downstream myeloid (common myeloid progenitors (CMP)/granulocyte myeloid progenitors (GMP)) and megakaryocyte and erythrocyte progenitors (MEP) (Fig. 2c,d). Moreover, Cell-ID was consistently able to identify singleton cells, that is, rare cell types represented by only one cell within a dataset (Supplementary Note 3). The capacity of Cell-ID to recover well-established cell types at single-cell resolution supports its use for automated cell-type annotation, even for extremely rare cells, without the need for clustering.

Cell matching across datasets from the same tissue. We benchmarked the capacity of Cell-ID to match cells of analogous cell types across independent scRNA-seq datasets from the same tissue of origin, within and across species (Supplementary Note 4). Cell-ID matching across datasets is performed by a per-cell assessment in the query dataset evaluating the replication of gene signatures extracted from the reference dataset. Gene signatures from the reference dataset can be automatically derived either from individual cells (Cell-ID(c)), or from previously established groups of cells (Cell-ID(g), Methods). We thus analyzed independent human pancreatic islets[20–22] and human and mouse airway epithelium datasets[23,24] corresponding to multiple donors and diverse sequencing technologies (Fig. 3 and Supplementary Note 4). Cell-ID(c) and Cell-ID(g) consistently yielded high precision (>94 and >92%), recall (>77 and >75%) and F1 values (>88 and >87%, respectively) across all evaluated reference-to-query cell-type assignments (multinomial $P < 2.2 \times 10^{-16}$, Fig. 3a and Supplementary Fig. 5). The overall performance of Cell-ID was at least as good as that of reference methods for cell matching and label transfer[25–33] (Fig. 3a and Supplementary Fig. 6a,b), and salient scores were obtained for cell types observed at low frequencies (<2%): epsilon cells, tissue-resident macrophages, mast cells and endothelial cells from pancreatic islet samples[21,22] and pulmonary neuroendocrine cells (PNECs), brush cells and ionocytes in mouse and human airway epithelium datasets[23,24] (Fig. 3b and Supplementary Fig. 6c,d). We further probed the capacity of this tool at individual cell resolution, by extracting Cell-ID signatures from 13 Schwann cells described in a dataset of 8,629 human pancreatic cells[20]. With these signatures, we were able to recognize four and two prototypic Schwann cells that had previously gone unnoticed within two independent sets of 2,126 and 2,261 human pancreatic cells, respectively, obtained from different donors and with different sequencing protocols[21,22] (Supplementary Fig. 7).

Cell matching across samples from different tissues. We then assessed the ability of Cell-ID to recognize rare cells from the same cell type across independent scRNA-seq datasets from different tissues of origin and, thus, within diverse cell composition contexts (Supplementary Note 5). Thus, based on the unbiased gene signatures obtained from airway epithelium cells[23], Cell-ID was able to identify brush/tuft cells, endocrine cells and goblet cells in the intestinal epithelium[34] with high precision (90%), recall (73%)

Fig. 3 | Performance of Cell-ID cell matching across scRNA-seq datasets from the same or different tissue of origin, within and across species. a,**b**, F1 score (y axis) achieved by Cell-ID(g), Cell-ID(c) and ten reference methods (x axis). **a**, Results are represented for each label transferring evaluated (top left panels), corresponding to cell matching across datasets from pancreatic islets (red squares and diamonds, $n = 8,659$, 2,126 and 2,970 cells for Baron, Muraro and Segerstolpe, respectively), and across datasets from airway epithelium (blue triangles, $n = 7,662$, 7,586 and 2,970 cells for Plasschaert mouse, Montoro and Plasschaert human, respectively). Boxplots summarize method's macro F1 scores, and show the median (center line), interquartile range (hinges) and 1.5 times the interquartile range (whiskers). **b**, Results are represented for the rare cell types reported in pancreatic islets (red squares and diamonds) and airway epithelium datasets (blue triangles). A cell-type label gathering brush and PNEC was used for consistency with Plasschaert. **c**, *t*-SNE representation of 7216 cells from mouse small intestinal epithelium[34]. Cells are colored according to Cell-ID(g) cell-type predictions, using as a reference mouse airway epithelial gene signatures extracted from Plasschaert. Cells with significant enrichments in airway signatures without an analogous cell type in intestine are represented in black. Cells with no significant enrichments are displayed in gray. **d**, F1 score (y axis) achieved by Cell-ID(g), Cell-ID(c) and reference methods (x axis), for the label transferring depicted in **c**. **e**, UMAP representation of 9,126 mouse olfactory epithelium cells from Wu et al.[35]. Cells are colored according to Cell-ID(g) predictions using as a reference brush/tuft gene signatures extracted from (1) mouse airway epithelium and (2) mouse small intestinal epithelium. The 37 cells significantly enriched with airway brush/tuft gene signatures are highlighted in blue and were interpreted as putative SCCs. Identical results were obtained when using intestinal brush/tuft signatures. **f**, Violin plots of the distribution of $-\log_{10}$ enrichment P values (y axis) across the 37 cells identified in **e** are represented for five significant functional terms (x axis). GO-0019370, Gene Ontology (GO) term 'leukotriene biosynthetic process'; GO-0097400, 'interleukin-17-mediated signaling pathway'; WP318, WikiPathway 'Eicosanoid Synthesis'; KEGG-hsa04725, KEGG term 'Cholinergic synapse'; R-HSA-112315, Reactome term 'Transmission across Chemical Synapses'.

and F1 scores (78%), outperforming reference methods for cell matching (multinomial $P < 2.2 \times 10^{-16}$ for all figures, Fig. 3c,d and Supplementary Fig. 8). From a discovery perspective, we used Cell-ID to perform cell-type scanning of two independent olfactory epithelium datasets[35,36] against brush/tuft signatures from the airway and the intestinal epithelium, which enabled us to identify putative rare solitary chemosensory cells (SCCs), a type of chemosensory cells closely related to brush/tuft cells, that had remained unclassified in the original publications (Fig. 3e,f and Supplementary Fig. 9). Thus, a total of 37 (<0.5%) and 5 (<0.6%) olfactory epithelium cells were found to display significant enrichment in airway and intestinal brush/tuft signatures (Benjamini–Hochberg corrected $P < 1.0 \times 10^{-21}$), with a median of 29% ± 0.5 and 23% ± 0.4 of genes, respectively, in common. These cells displayed high levels of expression for the characteristic SCC marker genes *Il25* and *Gnat3*, and their Cell-ID gene signatures were significantly enriched in cysteinyl leukotriene biosynthesis genes, as reported by Ualiyeva et al.[37] for SCCs (Supplementary Table 1). Our findings confirm, at single-cell



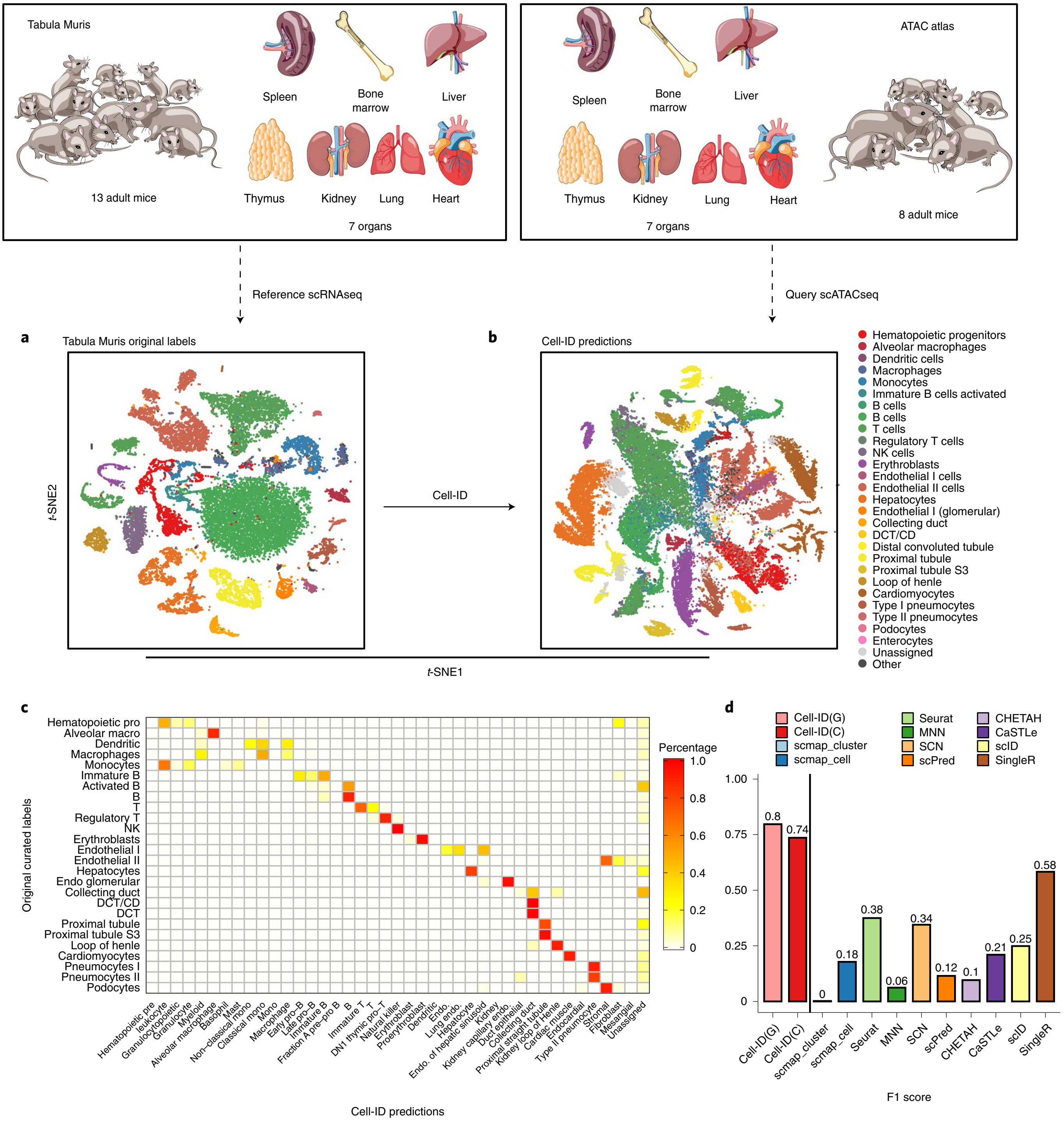


Fig. 4 | Performance of Cell-ID cell-to-cell matching across independent datasets from different single-cell omics technologies: scRNA-seq and scATAC-seq. a, *t*-SNE representation of 25,332 cells from seven organs profiled with scRNA-seq 10X genomics from the Tabula Muris mouse cell atlas. Cells are colored according to the manually annotated cell types provided by the atlas, regrouped by the categories represented in the legend in **b** and described in Supplementary Table 2. **b**, *t*-SNE representation of 50,284 cells from seven organs profiled with scATAC-seq from the Mouse ATAC atlas. Cells are colored according to the Cell-ID(g) predictions using the group gene signatures extracted from **a**, as represented in the legend. **c**, Heatmap representing the confusion matrix between the manually curated cell-type annotations from the Mouse ATAC atlas (displayed per rows) and the Cell-ID(g) cell-type predictions (displayed per columns) using the gene signatures extracted from the manually annotated cell types provided by the Tabula Muris mouse cell atlas. The color code in the heatmap represents the ratio *r* of the cell types displayed per rows that are allocated in the cell types represented per columns, ranging from white ($r=0$) to red ($r=1$). **d**, Global performance measured through the macro F1 score (*y* axis) achieved by Cell-ID(g), Cell-ID(c) and ten alternative state-of-the-art methods (*x* axis) on the label transferring from the scRNA-seq to the scATAC-seq mouse cell atlas. DCT/CD, Distal convoluted tubule/collecting duct; endo, endothelial; mono, monocytes.

resolution, the recently reported transcriptional and functional similarity between rare olfactory SCCs and brush/tuft cells from the airways and intestinal epithelia[37,38] (Supplementary Note 5).

Cell matching across datasets from different technologies. Finally, we assessed the reproducibility of Cell-ID gene signatures across datasets profiled with different single-cell omics technologies: scRNA-seq from the Tabula Muris mouse cell atlas[39] and single-cell ATAC-seq from the Mouse ATAC Atlas[40] (Supplementary Note 6). We benchmarked large-scale cell-type label transfer between the two expert-annotated atlases, collectively including 50 cell types from the eight tissues common to both: heart, kidney, liver, lung, bone marrow, spleen, thymus and large intestine (Fig. 4a,b and Supplementary Fig. 10). Both Cell-ID(c) and Cell-ID(g) matched cell types across scRNA-seq and sci-ATAC-seq datasets with high F1 scores and, together with SingleR[32], outperformed all the other reference methods evaluated (Fig. 4c,d and Supplementary Figs. 11 and 12). The capacity of Cell-ID to extract gene signatures in an automated manner that are robustly replicated across different single-cell omics technologies and cell heterogeneity backgrounds (Supplementary Notes 6 and 7), together with its computational scalability (Supplementary Note 8) pave the way for the systematic multi-omics scanning of rare cell types across tissues and whole organisms.

Discussion

Throughout this study, we found that Cell-ID was able to extract unbiased per-cell gene signatures from scRNA-seq experiments that could then be used as unique cell identity cards without the need for previous knowledge. Cell-ID signatures were consistently reproducible across diverse benchmarks collectively involving 14 independent single-cell datasets, 13 organs/tissues, more than 50 cell types, more than 200,000 cells, two model organisms, six sequencing protocols and two single-cell omics technologies. Such signatures made it possible to recognize cell identities across datasets from the same or different tissues of origin and model organisms, while overcoming batch effects arising from the use of different donors and sequencing technologies. In particular, the automatic cell-type annotations and matching across sets provided by Cell-ID were fully transparent with respect to the genes driving the hits. Such transparency improves biological interpretation at the individual cell level (Supplementary Note 9), making it possible to discover bona fide rare cells, as illustrated by the identification of Schwann cells and SCCs previously overlooked in published datasets (Supplementary Notes 4 and 5). This contrasts strongly with the capabilities of the other methods currently available, which are based on assessments of similarity over the entire transcriptome[25], cell embedding[26,33] or machine-learning approaches[27,28,30,31], in which the contributions of individual genes are difficult to interpret.

Cell-ID is computationally efficient, can be scaled up for use with large datasets and allows many-to-many dataset comparisons without the need for data integration steps (Supplementary Note 8). It can, thus, be used for the systematic screening of each individual cell in newly sequenced datasets against (1) reference databases of marker lists associated with well-established cell types (for example, PanglaoDB[41], CellMarkers[42]), (2) reference single-cell atlas databases with manually curated cell-type annotations (Tabula Muris[39], mouse ATAC atlas[40], human cell atlas[1]) and (3) molecular signature collections, functional ontologies and pathway databases (for example, MSigDB[43], Gene Ontology[44], Reactome[45], KEGG[46], Wikipathways[47]). The automatic single-cell annotations provided by Cell-ID will alleviate the need for the often-tedious visual inspection of prototypic marker levels based on expert knowledge. Yet, we stress that Cell-ID gene signatures are relative to the cell heterogeneity background from which they were assessed (Supplementary Note 7). Thus, for the purpose of cell-to-cell matching and label transferring across independent sets, comparable cell heterogeneity backgrounds are preferable (for example, tissue-level samples).

The MCA approach implemented in Cell-ID could be in principle generalized to single-cell readouts other than transcription levels, for example surface protein levels, chromatin accessibility or DNA methylation, for the assessment of per-cell molecular signatures. However, in the scenario where the signatures at different omics levels of a given cell type or cell state are not redundant among them, Cell-ID cell matching across independent datasets profiled with different modalities may be compromised. Thus, when different single-cell modalities capture diverse aspects of the cell heterogeneity within a sample, the simultaneous profiling of the different omics components in the same cells would be particularly relevant[48]. Here integrative multimodal approaches may uncover cell type and/or cell state heterogeneity with enhanced resolution[49,50]. In this setting, the Cell-ID approach could still be applied on each modality to extract and interpret complementary molecular signatures at individual cell level, while taking advantage of cell-type labeling or unsupervised clustering resulting from an integrated multimodal approach. From a discovery perspective, the identification of individual cells presenting distinctive and reproducible molecular signatures consistent with the phenotypes studied would constitute a first step toward the in-depth experimental characterization of putative new human cell types or cell states in health and disease. Cell-ID is distributed as an open-source R software package including detailed tutorials: https://github.com/RausellLab/CelliD. A development version of Cell-ID software is also available in Bioconductor (devel branch 3.13: https://bioconductor.org/packages/CelliD). Scripts to reproduce all the analyses and figures presented in this paper are provided at https://github.com/RausellLab/CellIDPaperScript.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-00896-6.

Received: 5 August 2020; Accepted: 15 March 2021;
Published online: 29 April 2021

References

1. Teichmann, S. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
2. National Institutes of Health. The Human BioMolecular Atlas Program: HuBMAP *NIH Common Fund Program* https://commonfund.nih.gov/HuBMAP (2021).
3. The LifeTime Initiative *LifeTime FET Flagship* https://lifetime-fetflagship.eu/ (2021).
4. Lähnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
5. Sun, S., Zhu, J., Ma, Y. & Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol.* **20**, 269 (2019).
6. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnol.* **37**, 38–44 (2019).
7. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.* **7**, 1141 (2018).
8. 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).
9. Greenacre, M. J. *Theory and Applications of Correspondence Analysis* (Academic Press, 1984).
10. Greenacre, M. & Blasius, J. (eds). *Multiple Correspondence Analysis and Related Methods* (Chapman & Hall/CRC, 2006).
11. Aşan, Z. & Greenacre, M. Biplots of fuzzy coded data. *Fuzzy Set. Syst.* **183**, 57–71 (2011).
12. Rausell, A., Juan, D., Pazos, F. & Valencia, A. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl Acad. Sci. USA* **107**, 1995–2000 (2010).

13. Gabriel, K. R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58**, 453–467 (1971).
14. Greenacre, M. *Biplots in Practice* Ch. 8, 79–88 (Foundation BBVA, Rubes Editorial, 2010).
15. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
16. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
17. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
18. Peterson, V. M. et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).
19. Zhang et al. SCINA: semi-supervised analysis of single cells in silico. *Genes* **10**, 531–531 (2019).
20. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems* **3**, 346–360 (2016).
21. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
22. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Systems* **3**, 385–394.e3 (2016).
23. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
24. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
25. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–359 (2018).
26. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
27. De Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.* **47**, e95 (2019).
28. Lieberman, Y., Rokach, L. & Shay, T. CaSTLe–classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS ONE* **13**, e0205499–e0205499 (2018).
29. Boufea, K., Seth, S. & Batada, N. N. scID uses discriminant analysis to identify transcriptionally equivalent cell types across single-cell RNA-seq data with batch effect. *iScience* **23**, 100914 (2020).
30. Tan, Y. & Cahan, P. SingleCellNet: a computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Systems* **9**, 207–213.e2 (2019).
31. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. ScPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.* **20**, 264–264 (2019).
32. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
33. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
34. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
35. Wu, Y. et al. A population of navigator neurons is essential for olfactory map formation during the critical period article a population of navigator neurons is essential for olfactory map formation during the critical period. *Neuron* **100**, 1066–1082.e6 (2018).
36. Fletcher, R. B. et al. Deconstructing olfactory stem cell trajectories at single-cell resolution. *Cell Stem Cell* **20**, 817–830.e8 (2017).
37. Ualiyeva, S. et al. Airway brush cells generate cysteinyl leukotrienes through the ATP sensor P2Y2. *Science Immunol.* **5**, eaax7224–eaax7224 (2020).
38. Bankova, L. G. et al. The cysteinyl leukotriene 3 receptor regulates expansion of IL-25–producing airway brush cells leading to type 2 inflammation. *Science Immunol.* **3**, eaat9453 (2018).
39. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
40. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e18 (2018).
41. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, baz046 (2019).
42. Zhang, X. et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47**, D721–D728 (2019).
43. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cells* **1**, 417–425 (2015).
44. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
45. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
46. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, 457–462 (2015).
47. Slenter, D. N. et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46**, D661–D667 (2018).
48. Efremova, M. & Teichmann, S. A. Computational methods for single-cell omics across modalities. *Nat. Methods* **17**, 14–17 (2020).
49. Hao, Y. et al. Integrated analysis of multimodal single-cell data. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.12.335331 (2020).
50. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).

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

Methods

Preprocessing and normalization of single-cell datasets. All scRNA-seq datasets analyzed in the study took as input the raw count gene expression matrices provided by the original sources. Library size normalization was carried out by rescaling counts to a common library size of 10,000. The log transformation was performed after adding a pseudo-count of 1. All analyses throughout the paper were restricted to a background set of 19,308 and 21,914 protein-coding genes from human and mouse, respectively, obtained from BioMart Ensembl release 100, v.April 2020 (GrCH38.p13 for human and GRCm38.p6 for mouse[51,52]). Genes expressing at least one count in less than five cells were removed. No filtering of cells was done unless the original sources provided ‘doublet’ or contamination annotations, which were in such cases filtered out. In the case of sci-ATAC gene activity score matrices, the same preprocessing as for scRNA-seq data was applied.

Overview of the Cell-ID approach. The main steps of the Cell-ID workflow are schematized in Fig. 1. First, the library size normalized and log-transformed count matrix is transformed into a fuzzy-coded indicator matrix where expression values are represented in a continuous scale between 0 and 1. Second, Cell-ID performs a dimensionality reduction of the indicator matrix using MCA where both cells and genes are represented into the same vector space[10]. Third, per-cell gene rankings are calculated from the gene-to-cell distances in MCA space, where the top closest genes to a cell will define its gene signature. If a grouping of cells is provided, per-group gene rankings may be obtained in an analogous way by using the geometric centroid in MCA space of the cells belonging to a given group. The enrichment of per-cell and/or per-group gene signatures is then evaluated through hypergeometric tests against (1) reference marker gene lists, and/or (2) per-cell and/or per-group gene signatures extracted through Cell-ID from reference single-cell datasets. Per-cell and per-group gene signatures represent thus identity cards allowing automatic cell type and functional annotation, as well as cell matching across datasets. Each of these steps is described in detail in the following sections.

MCA of the gene expression matrix. MCA is a multivariate descriptive statistical technique conceptually equivalent to PCA for qualitative/binary data[9,53]. MCA can be applied to quantitative data through an intermediate step of a so-called fuzzy coding. Here each continuous variable p is coded through user-defined functions into a number of disjoint categories Q_p where membership x to each category q is represented in a continuous scale between 0 and 1, and $\sum_{j=1}^{Q_p} x_j = 1$. Following ref. [11], fuzzy coding of a cases-by-variables matrix of continuous data can be performed in its simplest form by doubling each variable into $Q_p = 2$ categories as follows: let $M_{N,P}$ be the gene expression matrix of N cells (that is, cases) and P genes (that is, variables), with general term m_{np} gathering the expression level of gene p in cell n. For each column vector **Mp** in M, two membership functions $x^+ = f^+(m) : \Re \rightarrow \Re : [0,1]$ and $x^- = f^-(m) : \Re \rightarrow \Re : [0,1]$ (where $f^-(m) = 1 - f^+(m)$; $\Re$ being the set of real numbers) can be defined by linearly scaling between 0 and 1 the expression values for each gene across all cells as follows:

$$x_{np}^+ = \frac{m_{np} - \min(\mathbf{M_p})}{\max(\mathbf{M_p}) - \min(\mathbf{M_p})} \ ; \ x_{np}^- = 1 - x_{np}^+$$

From such functions, a fuzzy-coded indicator matrix $X_{N,K}$ can be built, representing a total of $K = 2P$ categories:

$$X_{N,K} = \begin{bmatrix} x_{11}^+ & x_{11}^- & \dots & x_{1P}^+ & x_{1P}^- \\ x_{21}^+ & x_{21}^- & \dots & x_{2P}^+ & x_{2P}^- \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N1}^+ & x_{N1}^- & \dots & x_{NP}^+ & x_{NP}^- \end{bmatrix}$$

The grand total of $X_{N,K}$ is thus NP, since each of the N cells has P sets of fuzzy-coded-values, each adding up to 1. The MCA of a fuzzy-coded indicator matrix follows that of a regular MCA[11]. From the matrix $X_{N,K}$, a matrix of relative frequencies $F_{N,K}$ is defined as:

$$F_{N,K} = \frac{1}{NP} X$$

From the row sums and column sums of F, two diagonal matrices D_r and D_c are built, respectively, with general terms

$$d_{r_{n,n}} = \sum_{k=1}^{K} f_{nk} \ ; \ d_{c_{k,k}} = \sum_{n=1}^{N} f_{nk}.$$

Let S_{NK} be the matrix of standardized relative frequencies resulting from

$$S = D_r^{-1/2} F D_c^{-1/2}$$

The singular-value decomposition of the matrix S_{NK} leads to

$$S = U D_\alpha V^T$$

where U and V contain by columns the singular vectors of norm 1 ($U^T U = 1$; $V^T V = 1$), and D_α is a diagonal matrix of singular values α_i, which are positive and displayed in descending order: $\alpha_1 \geq \alpha_2 \geq \ldots > 0$. Alternatively, U and V can be obtained as the matrices displaying by columns the eigenvectors of norm 1 of the product SS^T and $S^T S$, respectively, with eigenvalues λ_i, where $\alpha_i = \lambda_i^{1/2}$. Thus,

$$SS^T U = U D_\lambda \ ; \ S^T S V = V D_\lambda \ ; \ U^T U = 1 \ ; \ V^T V = 1$$

Alternatively, V can be calculated from U with the transition formula:

$$V = S^T U D_\alpha^{-1} = S^T U D_\lambda^{-1/2}$$

In the previous expressions, the first vectors $\vec{u}_1$ and $\vec{v}_1$ are associated with the trivial solution $\alpha_1 = \lambda_1 = 1$, and are thus removed from the analysis at this stage. After eliminating the trivial solution, the sum of all the eigenvalues λ_i from SS' (so-called total inertia in MCA terminology) equals the chi-squared statistic χ^2 of the indicator matrix $X_{N,K}$ divided by N. Thus, the orthogonal vectoral space generated by the eigenvectors of SS' can be viewed as a decomposition of the chi-squared statistic χ^2 in its independent sources of variation, each accounting for a fraction given by $\lambda_i / \sum_{i=1}^{I} \lambda_i$. At this stage, further dimensionality reduction can be performed by retaining the first J eigenvectors as the most informative components, while disregarding the rest of dimensions from downstream analysis. Here we established $J = 50$ as the default parameter throughout all the analyses performed.

The orthogonal dimensions given by the eigenvectors U and V allow to simultaneously represent both rows (that is, cells) and columns (that is, gene categories) in S into the same orthogonal vectoral space, where coordinates are obtained as follows[14]:

$$\text{Row standard coordinates} : \Phi = D_r^{-1/2} U = D_r^{-1/2} S V D_\alpha^{-1}$$

$$\text{Column principal coordinates} : G = D_c^{-1/2} S^T U = D_c^{-1/2} V D_\alpha$$

The previous expressions correspond to so-called standard and principal coordinates for rows and columns, respectively, in MCA terminology[14].

From the previous expressions, MCA can be formulated as a particular type of biplot (Supplementary Note 1), where the product of the two matrices ΦG^T approximates the contingency ratios expressed in $D_r^{-1} F D_c^{-1}$:

$$\Phi \mathrm{G}^T = (D_r^{-1/2} U)(D_\alpha V^T D_c^{-1/2}) = D_r^{-1/2} (U D_\alpha V^T) D_c^{-1/2} = D_r^{-1/2} S D_c^{-1/2} = D_r^{-1} F D_c^{-1}$$

Gene-to-cell distances in MCA space and per-cell and per-group gene signature extraction. In the vectoral space provided by MCA, a barycentric relationship is fulfilled between the rows and the column coordinates: the general term g_{kj} of G representing the coordinate of a column k in the dimension represented by the eigenvector $\vec{u}_j$ of U, corresponds to the weighted average (that is, the center of mass) of the N row coordinates ϕ_{nj} from Φ, where weights are given by $x_{nk} / \sum_{n=1}^{N} x_{nk}$, that is the frequency conditioned by columns of the corresponding values in the fuzzy indicator matrix $X_{N,K}$:

$$g_{kj} = \frac{1}{\sum_{n=1}^{N} x_{nk}} \sum_{n=1}^{N} x_{nk} \phi_{nj}$$

Thus, in MCA space, the closer a column (that is, gene category) is to a row (that is, cell), the more specific it is to it. In addition, each set of $Q_p = 2$ categories for each gene p is centered at the origin: that is $\vec{g}_p^+ + \vec{g}_p^- = \vec{0}$. At this stage, only gene category coordinates $\vec{g}_p^+$, conveying presence of gene expression relative to the maximum per gene, are retained for downstream analysis. From the previous expressions, the Euclidean distances $d_{np}(\vec{\phi}_n, \vec{g}_p^+)$ can be computed for each cell n and each gene p in the dataset. The genes g_p constituting the signature Γ_n associated to a cell n are obtained from its top γ closest genes in MCA space:

$$\Gamma_n = \left\{ g_p \,\middle|\, \forall p : \underset{p}{\mathrm{rank}} \left(d_{np} \left(\vec{\phi}_n , \vec{g}_p^+ \right) \right) \leq \gamma \right\}$$

A default value of $\gamma = 200$ was established throughout this work and ties resolved with random ranks. More generally, the genes g_p constituting the signature Γ_Θ associated to a group of cells Θ can be obtained from the Euclidean distances $d_p(\vec{\phi}_\theta, \vec{g}_p^+)$ between each gene p and the group centroid $\vec{\phi}_\theta$, obtained from the geometric center of the $\vec{\phi}_n$ vectors associated to the cells $n \in \Theta$:

$$\Gamma_\Theta = \left\{ g_p \,\middle|\, \forall p : \underset{p}{\mathrm{rank}} \left(d_{np} \left(\vec{\phi}_\theta , \vec{g}_p^+ \right) \right) \leq \gamma \right\}$$

We note, however, that Cell-ID does not perform or relies in any clustering step whatsoever. Notwithstanding, cell grouping information may optionally be used here as input, as provided by a external reference source (for example, database or publication).

Per-cell gene signature enrichment analyses against reference gene sets. The gene signatures $\mathit{\Gamma}_n$ extracted for each cell n in a dataset can be assessed through their enrichment against reference gene sets (for example, marker gene lists) associated to well characterized cell types and/or functional terms. Cell-ID evaluates such enrichment through a hypergeometric test as follows: let P be the set of genes retained in the gene expression matrix $M_{N,P}$ previously defined, after the initial steps of cell and gene filtering described above. Let W be the set of genes within a reference gene set that are contained on P ($W \subset P$). Let w be the number of genes overlapping between the signature $\mathit{\Gamma}_n$ of size γ and the gene set W:

$$w = |\Gamma_n \cap W|$$

The observed overlap w can be modeled as a random variable X distributed hypergeometrically, with probability mass function given by:

$$\mathrm{Prob}_{nW}(X = w) = \frac{\binom{W}{w}\binom{P - W}{\gamma - w}}{\binom{P}{\gamma}}$$

Only reference gene sets of size $W \geq 10$ were considered throughout this work. When the gene signature $\mathit{\Gamma}_n$ of a cell n in a dataset D is evaluated against a collection of reference gene sets $W_1, W_2, \ldots, W_\Omega$ (for example, a repository of cell-type marker lists or a pathway database), the above hypergeometric test P values are corrected by multiple testing for the number of gene sets Ω evaluated. Thus, a cell n is considered as enriched in those gene sets for which the hypergeometric test P value is $<1 \times 10^{-2}$, after Benjamini–Hochberg multiple testing correction[54]. In addition, when a disjointed classification is required, a cell n may be assigned to the gene set W_ω with the lowest significant corrected P value. On the contrary, if no significant hits are found, a cell n will remain unassigned.

Per-cell gene signature enrichment analyses against per-cell and per-group gene signatures extracted from reference single-cell datasets. The gene signatures $\mathit{\Gamma}_n$ extracted for each cell n in a dataset D can be assessed through their enrichment against the gene signatures Γ'_n extracted for each cell n' in a reference dataset D', an approach called here Cell-ID(c). Analogous to the previous section, Cell-ID(c) evaluates such enrichment through a hypergeometric test as follows: Let P be the set of genes retained in the gene expression matrix $M_{N,P}$ associated to dataset D as previously defined. Let $\Gamma'_{n\prime|P}$ be the set of genes of size W' within a per-cell gene signature Γ'_n extracted for a cell n' in the dataset D', which are contained on P, that is $\Gamma'_{n\prime|P} = \Gamma'_{n\prime} \cap P$.

Let w' be the number of genes overlapping between the signature $\mathit{\Gamma}_n$ of size γ and the gene set P:

$$w' = \left|\Gamma_n \cap \Gamma'_{n'|P}\right|$$

The observed overlap w' between two per-cell gene signatures can be modeled as a random variable X distributed hypergeometrically, with probability mass function given by:

$$\mathrm{Prob}_{nW}(X = \mathrm{w}') = \frac{\binom{W'}{w'}\binom{P - W'}{\gamma - w'}}{\binom{P}{\gamma}}$$

For each cell n in a dataset D, the above hypergeometric test P values are corrected by multiple testing for the number of cells N' in the reference dataset D' against which it is evaluated. Thus, a cell n in D is considered as enriched in those signatures n' in D' for which the hypergeometric test P value is $<1 \times 10^{-2}$, after the Benjamini–Hochberg correction on the number N' of tested gene signatures. In addition, when a disjointed classification is required, a cell n may be assigned to the cell n' in D' with the lowest significant corrected P value. Best hits can be used for cell-to-cell matching and label transferring across datasets. On the contrary, if no significant hits are found, a cell n will remain unassigned.

Alternatively, if a grouping $\Theta'_1, \Theta'_2, \ldots, \Theta'_\theta$ of the N' cells in D' is provided, the gene signatures $\mathit{\Gamma}_n$ for each cell n in a dataset D can be assessed through their enrichment against the corresponding per-group gene signatures $\Gamma'_{\Theta_1}, \Gamma'_{\Theta_2}, \ldots, \Gamma'_{\Theta_\theta}$ extracted from D' as described above. We call this approach Cell-ID(g). Here a cell n in D is considered as enriched in those cell groups Θ'_ϑ from D' for which the hypergeometric test P value is $<1 \times 10^{-2}$, after Benjamini–Hochberg correction for the number of groups evaluated. In addition, when a disjointed classification is required, a cell n may be assigned to the group Θ'_ϑ in D' with the lowest significant corrected P value. Best hits can be used for cell-to-group matching and group-based label transferring across datasets. On the contrary, if no significant hits are found, a cell n will remain unassigned. Cell-ID(g) can handle both disjoint and nondisjoint cell groupings (that is, overlapping groups), as well as complete or non-complete groupings (that is, when not all cells in D' have been assigned to a group).

Interpretation of the distance of genes and cells to the origin in MCA space. In the MCA biplot, genes with rather uniform values for all cells will lie in the vicinity of the origin of coordinates $\vec{0}$, while less frequently expressed genes (that is, presenting high values for a fewer number of cells) will be farther from it, in relative terms. From a complementary point of view, in MCA space observations (that is, cells) lie at the barycenter of the variable values they present[55]. Thus, the distance of a cell to the origin reflects, on the one hand, how many less frequently expressed genes it possesses (that is, the more frequently expressed genes it presents, the closer it is to the origin) and, on the other hand, how close their values are to the mean of each variable (that is, gene). Thus, two extreme scenarios are theoretically possible by which a cell would be placed close to the origin $\vec{0}$ in all principal axes: (1) if all genes are rather uniform, which in turn would imply that all cells are indeed close to the origin, or (2) the cell presents values corresponding to the mean value for all genes, that is the observation equals the 'mean observation'. In the first scenario, distances from genes to cells would be rather uniform as well, and thus per-cell gene rankings would be useless. In the second scenario, the closest genes to the 'average cell' will indeed be those presenting uniform values across cells, and thus would rank high in their gene signature.

Simulated datasets. Simulated scRNA-seq datasets were obtained with the Splatter[56] Bioconductor package (v.1.4.1, https://bioconductor.org/packages/release/bioc/html/splatter.html). For the generation of structured datasets, each simulation was set to originate from five underlying subpopulations with relative sizes of 30, 25, 20, 15 and 10% cells, respectively. No clustering or cluster labels were used in any way for the purpose of the analysis. Splatter's logistic function, modeling the probability of a gene having zero counts, was defined by a midpoint parameter $x0 = 3$ counts, to obtain a simulated dataset with about 60–70% of the count matrix content equal to zero after default normalization and filtering, a dropout rate consistent with those observed on other datasets used in this paper. Default values were used for all the other parameters. Centered and scaled PCA was performed with the base R prcomp function.

Comparative benchmark of approaches performing marker-based cell-type annotation. In the comparative benchmark assessing the method's capacity to classify cells using preestablished gene signatures, two alternative semisupervised classifiers were used: SCINA[19] and AUCell[15] (Supplementary Table 3). Cell-ID predictions were based on a reference collection of blood cell markers from the XCell repository[16] (Supplementary Table 4). Only cell types of the hematopoietic lineage supported by more than three bulk RNA-seq samples in XCell were considered: HSCs, multipotent progenitors (MPP), B cells, CD4$^+$ T cells, CD8$^+$ T cells, natural killer cells, plasmacytoid dendritic cells, CMP, GMP, MEP, erythrocytes, megakaryocytes, platelets, basophils, eosinophils, neutrophils, CD14$^+$ monocytes, CD16$^+$ monocytes, macrophages, dendritic cells (DCs) and conventional dendritic cells (cDCs). For each cell type, the marker list used included genes replicated in at least 20% of the reported sources. Raw count matrices were used as input for AUCell and log-transformed normalized matrices were used for SCINA, following their associated vignettes. SCINA was run with default parameters except for (1) the maximum number of iterations and the convergence rate, which were increased to 20 and 0.999, respectively, to ensure a stable result, and (2) the rejection parameter, which was set to true to enable cells to be labeled as unassigned when there is a low confidence on the cell-type prediction. For AUCell, the gene set with the highest area under curve (AUC) score was used to classify cells unless the maximum AUC score was <0.1, what left a cell as 'unassigned'.

Comparative benchmark of approaches performing cell-type label transferring across datasets. In addition to Cell-ID, we evaluated ten alternative approaches for cell-type label transferring across scRNA-seq (Supplementary Table 3). All methods were run using default parameters unless otherwise stated. When default settings were not explicitly defined, setting used in the associated vignettes were followed. Methods used as input either the raw or the normalized count data (after gene filtering as described above), following each method's documentation. For those methods that stipulate it, gene expression matrices were further restricted to genes in common between the reference and the query datasets. In the case of mutual nearest neighbors (MNN)[26], we transferred labels from the reference to the query datasets between closest MNN cells, and cells were left unassigned when no MNN cells were found. We modified the default parameter of k nearest neighbor of MNN to $k = 50$ as the default $k = 20$ failed to find MNN matches for a large fraction of cells, which negatively affected the benchmark metrics evaluated. Furthermore, for the selection of hypervariable genes in MNN, we used the default Seurat function for highly variable gene detection (2,000 genes), and we took the intersection between the reference and the query highly variable genes to perform the query and reference dataset integration following the package's vignette. In the case of SCN[30], all cells that were classified as 'rand' were considered as unassigned, as well as all cells classified as 'nodes' in CHETAH[27]. For Seurat, cells were labeled as unassigned when the projection score was below 0.5.

Classification performance assessment. Cell annotation performance was assessed through three complementary metrics, that is precision, recall and F1 score, which is the harmonic mean between the precision and recall. Each metric was first calculated for each cell type in the query dataset. Second, each metric (that is, recall, precision and F1 score) was calculated for the global set as the arithmetic mean of the corresponding metric across the evaluated cell types. In such a way, an overweighed contribution of largely populated cell types is avoided, thus allowing a larger contribution of more rare cell types to the metrics evaluated. Mapping across datasets of cell-type nomenclature was performed through manual curation and is reported in Supplementary Table 5. For label transferring performance assessment between datasets that present different sets of cell types (that is, the mapping between mouse airway epithelial cells against mouse small intestinal epithelial cells in this paper) calculations are based on the cell types in common to the two sets (that is, endocrine, brush/tuft and goblet cell). Therefore, cells in the query dataset from a cell type only present in the query and that are assigned to a cell type only present in the reference dataset, are not considered in the calculations. In addition, cells left as unassigned in the query dataset were considered as a false assignment when their actual cell type was indeed represented in the reference dataset, or excluded from the calculations of performance otherwise. An additional metric was evaluated in that case, referred to here as rejection rate, defined as the percentage of unassigned cells out of the cells in the query dataset that belong to a cell type that is not in the reference dataset. For interspecies label transferring across human and mouse datasets, the initial raw matrix of the query dataset was restricted to genes with one-to-one orthologs. Ortholog relations were obtained from BioMart (release 100, v.April 2020, GrCH38.p13 for human and GRCm38.p6 for mouse[51,52]) using gene symbols.

Functional enrichment. Functional enrichment analyses were performed using six sources of functional annotations: Reactome[45], KEGG[46], WikiPathways[47], Gene Ontology biological process[44], Gene Ontology molecular function and Gene Ontology cellular component, collectively gathering a total of 8,709 terms. Gene sets associated with functional pathways and ontology terms were obtained as provided at the enrichr[57] website http://amp.pharm.mssm.edu/Enrichr/#stats (Supplementary Table 6).

Visualization. UMAP[6] or *t*-SNE[58] representations were alternatively used for visualization purposes throughout the paper. However, no biological conclusions whatsoever were drawn from visual inspection of such representations. UMAP and *t*-SNE representations were obtained with Seurat default parameters and preprocessing as described in this vignette (https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html). We note that UMAP and *t*-SNE coordinates were computed from PCA coordinates (that is, not MCA coordinates).

Computational resources. All analyses presented in the paper were run in a workstation with 64-Gb RAM memory and an AMD Ryzen 2700X processor with eight 3.6-GHz physical cores, with the exception of the scalability benchmark presented in Supplementary Note 7 where an Intel Xeon Gold 6140 with 36 2.3-GHz core processors and 640 Gb of RAM was used.

General statistical methods. In all boxplot representations, the whiskers denote data within 1.5× the interquartile range of the upper and lower quartiles. All points on boxplot were horizontally, but not vertically jittered or ordered in a visually comprehensive manner. Correlations were calculated using the cor function in the R programming language v.3.6.0.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All single-cell datasets used in this paper are publicly available (Supplementary Table 7). scRNA-seq datasets for human blood cells profiled by Cite-Seq[17] and Reap-Seq[18] were downloaded from the Gene Expression Omnibus (GEO) (accession numbers GSE100866 and GSE100501, respectively). Cell-type labels for these two datasets were obtained following the Multimodal Analysis vignette of the Seurat[33] R package (https://satijalab.org/seurat/multimodal_vignette.html). Pancreas scRNA-seq datasets from Baron[20], Muraro[22] and Segerstolpe[21], as well as their associated cell-type annotations were downloaded via the scRNA-seq[59] R package as a SingleCellExperiment format R object. Plasschaert[23] mouse and human and Montoro[24] mouse airway epithelium scRNA-seq datasets, and their annotations were downloaded from GEO (GSE102580, GSE103354). Haber[34] intestinal epithelium scRNA-seq dataset was downloaded from GEO accession code GSE92332. Olfactory epithelium scRNA-seq datasets from Fletcher[36] and Wu[35] were downloaded from GEO (GSE95601, GSE120199), and their cell-type annotations were obtained from the associated GitHub repositories: https://github.com/rufletch/p63-HBC-diff and https://www.stowers.org/research/publications/odr for Fletcher[36] and Wu[35], respectively. Tabula Muris[39] 10X and Smart-seq mouse scRNA-seq datasets were downloaded from https://tabula-muris.ds.czbiohub.org/. Gene activity score matrices from the Mouse sci-ATAC-seq atlas datasets from Cusanovich[40] were obtained from http://atlas.gs.washington.edu/mouse-atac/data/, as provided by the authors and resulting from the aggregation of information across all differentially accessible chromatin sites linked to a target gene.

Code availability

Cell-ID is implemented as an R package and is available on GitHub (https://github.com/RausellLab/CelliD) under the GPL-3 open-source license. Complete documentation is provided with step-by-step procedures for MCA dimensionality reduction, per-cell gene signature extraction, cell-type prediction, label transferring across datasets and functional enrichment analysis. A development version of Cell-ID software is also available in Bioconductor (devel branch 3.13): https://bioconductor.org/packages/CelliD. In addition, R scripts to reproduce all figures in the paper are available on a dedicated GitHub repository (https://github.com/RausellLab/CellIDPaperScript).

References

51. Zerbino, D. R. et al. Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
52. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
53. Lebart, L, Morineau, A & Warwick, K. M. *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices* (John Wiley & Sons, 1984).
54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B. (Methodological)* **57**, 289–300 (1995).
55. Pagès, J. *Multiple Factor Analysis by Example Using R* (CRC Press, 2014).
56. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174–174 (2017).
57. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **14**, 128 (2013).
58. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).
59. Risso, D. & Cole, M. scRNAseq: Collection of public single-cell RNA-Seq datasets. R package v.2.4.0 http://bioconductor.org/packages/scRNAseq/ (Bioconductor, 2020).

Acknowledgements

We thank the Laboratory of Clinical Bioinformatics and the Laboratory of Human Lymphohematopoiesis for helpful discussions and support. The Laboratory of Clinical Bioinformatics was partly supported by the French National Research Agency (ANR) 'Investissements d'Avenir' Program (grant no. ANR-10-IAHU-01). The Laboratory of Clinical Bioinformatics and the Laboratory of Human Lymphohematopoiesis were partly supported by Christian Dior Couture, Dior. We also thank G. Fuentes from The Visual Thinker LLP for the creation of the illustrations in Figs. 1–4.

Author contributions

A.C. and A.R. conceived and designed research. A.C. performed research. A.C and L.M. contributed with materials/analysis tools. A.C. and A.R. analyzed data. A.C., E.S. and A.R. interpreted results. A.C., E.S. and A.R. wrote the paper. All authors read and approved the final draft of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00896-6.

Correspondence and requests for materials should be addressed to A.R.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Dr. Antonio Rausell

Last updated by author(s): Mar 11, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No novel sequencing data was generated. All data used in the manuscript is publicly available. scRNAseq bioconductor package was used to download Baron, Muraro, and Segerstolpe pancreatic datasets. All other datasets were obtained from public repositories with accession codes reported in the manuscript.
Data analysis	Analyses were performed with R 3.6.0, using the following packages: splatter (1.10.0), scRNAseq (1.99.8), ComplexHeatmap (2.0.0), patchwork (1.0.0.9000), RColorBrewer (1.1-2), pbmcapply (1.5.0), irr (0.84.1), lpSolve (5.6.13.3), plotly (4.9.2.1), ggpubr (0.2.3), forcats (0.4.0), stringr (1.4.0), purrr (0.3.4), readr (1.3.1), tidyr (1.1.1), tibble (3.0.3), tidyverse (1.2.1), r2excel (1.0.0), xlsx (0.6.1), data.table (1.13.0), singleCellNet (0.1.0), cowplot (1.0.0), reshape2 (1.4.4), pheatmap (1.0.12), dplyr (1.0.1), foreach (1.4.7), scID (2.1), CHETAH (1.2.0), scPred (0.0.0.9000), Mlmetrics (1.1.1), kernlab (0.9-27), caret (6.0-85), lattice (0.20-38), irlba (2.3.3), Matrix (1.2-18), batchelor (1.2.0), scmap (1.8.0), SingleR (1.0.0), igraph (l.2.5), xgboost (0.90.0.2), seater (1.14.0), ggplot2 (3.3.2), magrittr (1.5), CellID (1.0.0), SingleCellExperiment (1.8.0), SummarizedExperiment (1.16.0), DelayedArray (0.12.0), BiocParallel (1.19.6), matrixStats (0.56.0), Biobase (2.46.0), GenomicRanges (1.38.0), GenomeInfoDb (1.22.0), IRanges (2.20.0), S4Vectors (0.24.0), BiocGenerics (0.32.0), Seurat (3.2.0). Cell-ID package, including detailed documentation and a tutorial vignette, is available from GitHub (https://github.com/RausellLab/CelliD)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All single-cell data sets used in this paper are publicly available (Supplementary Table 7). scRNA-seq datasets for human blood cells profiled by Cite Seq and Reap Seq were downloaded from the gene expression omnibus (GEO; accession numbers GSE100866 and GSE100501, respectively). Cell-type labels for these two datasets were obtained following the Multimodal Analysis vignette of the Seurat R package (https://satijalab.org/seurat/multimodal_vignette.html). Pancreas scRNA-seq datasets from Baron, Muraro, and Segerstolpe, as well as their associated cell-type annotations were downloaded via the scRNAseq R package as a SingleCellExperiment format R object. Plasschaert mouse and human and Montoro mouse airway epithelium scRNA-seq datasets, and their annotations were downloaded from GEO(GSE102580, GSE103354). Haber intestinal epithelium scRNA-seq dataset was downloaded from GEO accession code GSE92332. Olfactory epithelium scRNA-seq datasets from Fletcher and Wuwere downloaded from GEO (GSE95601, GSE120199), and their cell type annotations were obtained from the associated Github repositories: https://github.com/rufletch/p63-HBC-diff and https://www.stowers.org/research/publications/odr for Fletcher and Wu, respectively. Tabula Murisl0X and Smart-seq mouse scRNAseq datasets were downloaded from https://tabula-muris.ds.czbiohub.org/. Gene activity score matrices from the Mouse sci-ATAC-seq atlas datasets from Cusanovich were obtained from http://atlas.gs.washington.edu/mouse-atac/data/, as provided by the authors and resulting from the aggregation of information across all differentially accessible chromatin sites linked to a target gene. Gene sets associated to functional pathways and ontology terms were obtained as provided in enrichr website http://amp.pharm.mssm.edu/Enrichr/#stats. (Supplementary Table 6)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No experiments in study
Data exclusions	No experiments in study
Replication	No experiments in study
Randomization	No experiments in study
Blinding	No experiments in study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging