nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-019-0535-3

Supervised classification enables rapid annotation of cell atlases

Hannah A. Pliner[1], Jay Shendure[1,2,3,4*] and Cole Trapnell[1,2,4*]

Single-cell molecular profiling technologies are gaining rapid traction, but the manual process by which resulting cell types are typically annotated is labor intensive and rate-limiting. We describe Garnett, a tool for rapidly annotating cell types in single-cell transcriptional profiling and single-cell chromatin accessibility datasets, based on an interpretable, hierarchical markup language of cell type-specific genes. Garnett successfully classifies cell types in tissue and whole organism datasets, as well as across species.

Single-cell transcriptional profiling (scRNA-seq) is a powerful means of cataloging the myriad cell types present in multicellular organisms[1]. The computational steps of constructing a cell atlas typically include unsupervised clustering of cells based on their gene expression profiles, followed by annotation of known cell types among the resulting clusters[2,3]. For the latter task, there are at least four key challenges. First, cell type annotation is labor intensive, requiring extensive literature review of cluster-specific genes[4]. Second, any revision to the analysis (for example, additional data, parameter adjustment) necessitates manual reevaluation of all previous annotations. Third, annotations are not easily transferred between datasets generated by independent groups on related tissues, resulting in wasteful repetition of effort. Finally, annotation labels are typically ad hoc; although ontologies of cell types exist[5,6], we lack tools for systematically applying these ontologies to scRNA-seq data. Collectively, these challenges hinder progress toward a consensus framework for cell types and the features that define them.

Toward addressing these challenges, we devised Garnett (https://cole-trapnell-lab.github.io/garnett) (Fig. 1a). Garnett consists of four components. First, Garnett defines a markup language for specifying cell types using the genes that they specifically express. The markup language is hierarchical in that a cell type can have subtypes (for example, $CD4^+$ and $CD8^+$ are subsets of T cells). Second, Garnett includes a parser that processes the markup file together with a single-cell dataset, identifying representative cells bearing markers that unambiguously identify them as one of the cell types defined in the file. Third, Garnett trains a classifier that recognizes additional cells as belonging to each cell type based on their similarity to representative cells, similar to an approach that our groups recently developed for annotating a single-cell mouse atlas of chromatin accessibility[7]. Garnett does not require that cells be organized into clusters, but it can optionally extend classifications to additional cells using either its own internal clustering routines or those of other tools. Finally, Garnett provides a method for applying a classifier trained on one dataset to rapidly annotate additional datasets.

We tested Garnett on a benchmark scRNA-seq dataset comprising 94,571 immunophenotyped peripheral blood mononuclear cells (PBMCs), generated with the 10X Chromium platform[8]. Garnett requires ≥1 marker gene for each cell type. To classify the PBMCs, we populated a marker file including each of the expected cell types using literature-based markers. As a supervised method, Garnett's accuracy will be dependent on these markers, so we devised a measure of each marker's usefulness for the purposes of Garnett classification (see Methods). We used this quality metric to exclude poorly scoring markers (ambiguity >0.5) before proceeding with classification (Supplementary Fig. 1a).

Garnett assigned 71% (3% incorrect, 26% unclassified) of cells to the correct type (cluster-agnostic type), with 34% of T cells also receiving a correct subtype classification (41% not subclassified, 23% unclassified, 2% incorrect) (Fig. 1b,c). Cells remaining unlabeled were comparably distributed among immunophenotypes. Moreover, by expanding cell type assignments to nearby cells using Louvain clustering[9] (cluster-extended type), correct assignments increased to 94% (2% incorrect, 4% unclassified), with 91% of T cells also receiving a correct subtype classification (8% not subclassified, <1% unclassified, <1% incorrect).

We next evaluated Garnett's ability to classify data not seen during training by analyzing PBMCs that were profiled to a higher molecular depth with a different library preparation method and a different Chromium system (v.2) (Supplementary Fig. 1b). Because these cells were unsorted, we manually assigned cell types to clusters based on classic markers (Supplementary Fig. 1c–e). Although trained on sparser molecular data from a different method and instrument, classification accuracy remained high, with 80% (3% incorrect, 17% unclassified) of cells correctly labeled with cluster-agnostic type and 95% (3% incorrect, 2% unclassified) with cluster-extended type (Supplementary Fig. 1f,g). Of note, when trained on these more deeply profiled v.2 cells, Garnett also accurately classified the more sparsely profiled v.1 cells (83% correct with cluster-agnostic type and 95% correct with cluster-extended) (Supplementary Fig. 1h,). We furthermore used the PBMC datasets to explore the limits of Garnett and found that the algorithm was robust to rare and missing cell types and low data quality (Supplementary Figs. 2 and 3 and Supplementary Note).

To assess Garnett's ability to catalog cell types in complex solid tissues, we analyzed lung tissue data from two recently reported cell atlases, the Mouse Cell Atlas (MCA)[3] and Tabula Muris (TM)[2]. We defined a single hierarchy of expected lung cell types based on those studies and compiled marker genes from literature to recognize them in each dataset (marker files are Supplementary Files, consensus cell type names in Supplementary Table 1). Overall, Garnett's classifications agreed with both the MCA (58% correct, 29% unclassified with cluster-agnostic type; 65% correct, 23% unclassified with cluster-extended type; see Supplementary Fig. 4a,b) and TM (71%

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [3]Howard Hughes Medical Institute, Seattle, WA, USA. [4]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA. *e-mail: shendure@uw.edu; coletrap@uw.edu

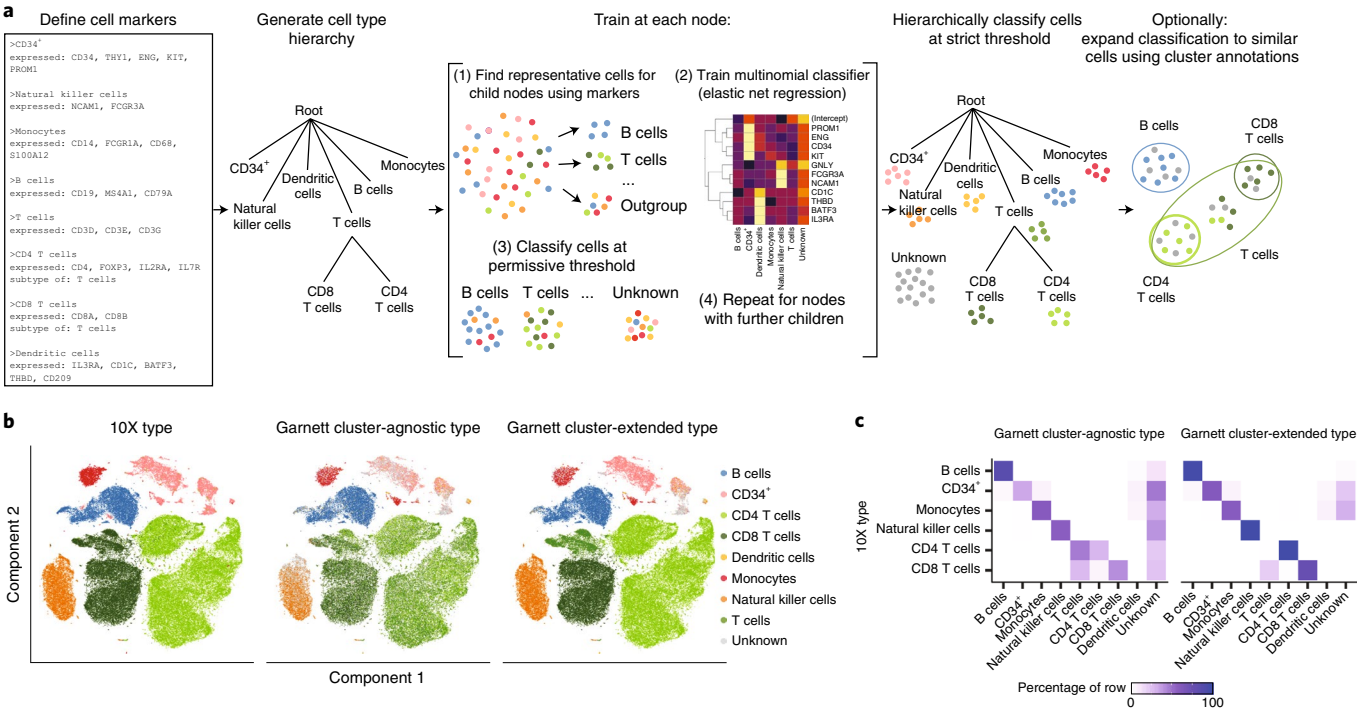

Fig. 1 | Garnett accurately classifies peripheral blood mononuclear cells. a, Overview of the Garnett algorithm (Methods). **b**, t-SNE plots of 10X Genomics' 100,000 cell PBMC dataset ($n = 94{,}571$ cells). The first panel is colored by cell type based on FACS sorting, the second panel is colored by cluster-agnostic cell type according to Garnett classification and the third panel is colored by the Garnett cluster-extended type, which labels cells based on the composition of their cluster or community. **c**, A heatmap of data in **b** comparing the labels based on FACS (rows) with the cluster-agnostic (left) and cluster-extended (right) cell type assignments by Garnett (columns). Color represents the percentage of cells of a certain FACS type labeled each type by Garnett. t-SNE, t-distributed stochastic neighbor embedding.

correct, 22% unclassified with cluster-agnostic type; 87% correct, 8% unclassified with cluster-extended type; see Supplementary Fig. 4c,d) annotations, which were derived by manual inspection of genes enriched in each cluster. Moreover, a Garnett model trained on the MCA accurately classified the TM cells and vice versa (trained on MCA: 82% correct, 5% unclassified with cluster-agnostic type; 86% correct, 2% unclassified with cluster-extended type; trained on TM: 46% correct, 30% unclassified with cluster-agnostic type; 56% correct, 21% unclassified with cluster-extended type; see Supplementary Fig. 4e–h).

We next sought to evaluate whether Garnett was similarly useful for annotating single-cell chromatin accessibility (scATAC-seq) datasets, which we have generally found to be more challenging to manually annotate than scRNA-seq datasets. We and colleagues recently used regularized, multinomial regression to classify clusters of cells based on chromatin accessibility[7]. We adapted Garnett to classify cells based on scATAC-seq-derived 'gene activity scores', a measure of open chromatin around each gene[10]. Applying it to our recent scATAC-seq atlas of the mouse[7], Garnett labeled 39% of cells concordantly with our previous assignments (cluster-extended; 22% incorrect; 39% unclassified) (Supplementary Fig. 5). A caveat is that the marker file was informed by our previous literature-based annotation of the dataset by a related method, but these analyses nonetheless illustrate the potential of Garnett to enable the rapid annotation of not only scRNA-seq but also scATAC-seq datasets.

We next applied Garnett to the task of discriminating all cell types of a whole animal, L2 stage *Caenorhabditis elegans*[11]. We defined a cell hierarchy that discriminated 29 major cell types, as well as subtypes of neuron, using the marker genes from the original study. Of cells previously assigned, Garnett labeled 87% of cells concordantly for major cell type (cluster-extended; 8% incorrect, 5% unclassified), with rectum cells being frequently mislabeled as non-seam hypodermis (Fig. 2a,b and Supplementary Figs. 6,7). Of 4,186 neurons originally assigned subtypes, 53% were subtyped correctly and a further 18% were labeled as neurons of unknown subtype (cluster-agnostic; 8% incorrect) (Fig. 2c). These analyses demonstrate Garnett can scale to classifying the cell types found in a whole animal.

To evaluate how Garnett would perform on a complex system with a deep hierarchy, we generated a four-level classifier with 144 cell definitions for mouse nervous system based on the data and taxonomy presented in ref. [12] (Supplementary Fig. 8). We found Garnett performed very well at the higher levels, but often underclassified cells at the lower, more specific levels (for example, classifying a cerebellum neuron as a neuron) (Supplementary Fig. 9a–e and Supplementary Fig. 10). The size and complexity of this hierarchy facilitated exploration of the properties of markers chosen by the elastic-net regression to discriminate among cell types. Garnett tended to select genes that were more highly expressed and more specific than other genes (Supplementary Fig. 11).

Finally, as tissue-specific gene expression patterns are largely conserved across vertebrates[13], we wondered whether Garnett models trained on mouse data could be used to classify human cell types. We applied the Garnett model trained on the MCA lung dataset to scRNA-seq data from human lung tumors[14] (Fig. 3a,b, Supplementary Fig. 12 and Supplementary Table 1). Over 92% of alveolar, B cells, T cells, epithelial (ciliated) cells, endothelial cells and fibroblasts were accurately assigned by the Garnett MCA model. Of the 9,756 cells annotated as myeloid[14], Garnett labeled 44% as monocyte/macrophage/dendritic cell and a further 16% granulocytes, leaving 34% unclassified. 22% of the dataset was labeled 'unknown', of which 55% were identified as tumor cells in

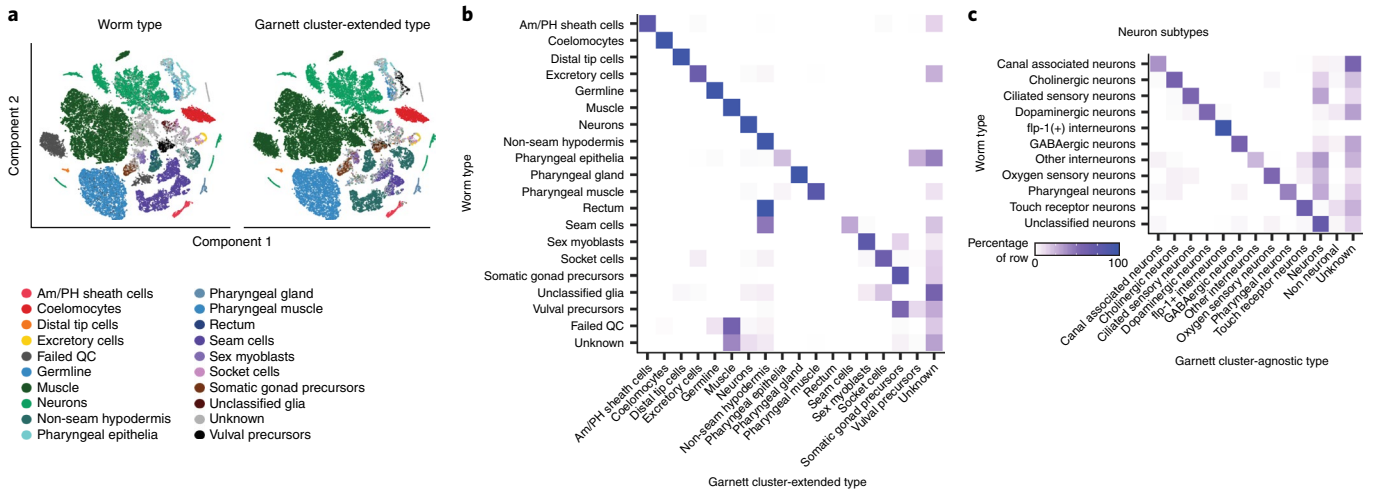

Fig. 2 | Garnett can discriminate among cell types across a whole animal, across species and between normal and pathological tissue. Garnett classification results for sci-RNA-seq data from whole *C. elegans*, published in ref. [11]. **a**, t-SNE plots of the whole worm dataset ($n = 42{,}035$ cells). Left panel is colored by published type from ref. [11], right panel colored by the major (top level) Garnett cluster-extended classification. Garnett cluster-agnostic type is available in Supplementary Fig. 7. **b**, Heatmap comparing the reported cell types versus the Garnett cluster-extended cell types. Color represents the percentage of cells of a certain reported type labeled as each type by Garnett. **c**, Heatmap comparing the reported neuron subtypes versus the Garnett cluster-agnostic neuron subtypes. Am/Ph, amphid and phasmid; QC, quality control.

the original study. As expected, given that they are not represented in the original marker file nor in the MCA lung dataset, 88% of all cells annotated as tumor cells in the original study were labeled as 'unknown' by Garnett. These analyses demonstrate that Garnett has the potential to operate across related species, and is not necessarily confounded by the presence of pathological cell states when trained on normal healthy tissue.

Cell type annotation is a critical and rate-limiting step in cell type atlas construction, as illustrated by recent studies that resorted to labor-intensive, ad hoc literature review to achieve this end[2,3,7,11,12,15]. Garnett is an algorithm and accompanying software that automates and standardizes the process of classifying cells based on marker genes. While other algorithms for automated cell type assignment have been published[3,16] we believe that Garnett's ease-of-use and lack of requirement of pre-classified training datasets will make it an asset for future cell type annotation. One existing method, scMCA, trained a model using MCA data that can be applied to newly sequenced mouse tissues. scMCA reported a slightly higher accuracy than Garnett[3], likely because of a training procedure that relies on manual annotation of cell clusters. But a key distinction is that the hierarchical marker files on which Garnett is based are interpretable to biologists and explicitly relatable to the existing literature. Furthermore, together with these markup files, Garnett classifiers trained on one dataset are easily shared and applied to new datasets, and are robust to differences in depth, methods and species.

We anticipate the potential for an 'ecosystem' of Garnett marker files and pre-trained classifiers that: (1) enable the rapid, automated, reproducible annotation of cell types in any newly generated dataset, (2) minimize redundancy of effort, by allowing for marker gene hierarchies to be easily described, compared and evaluated and (3) facilitate a systematic framework and shared language for specifying, organizing and reaching consensus on a catalog of molecularly defined cell types. To these ends, in addition to releasing the Garnett software, we have made the marker files and pre-trained classifiers described in this manuscript available at a wiki-like website that facilitates further community contributions, together with a web-based interface for applying Garnett to user datasets (https://cole-trapnell-lab.github.io/garnett).

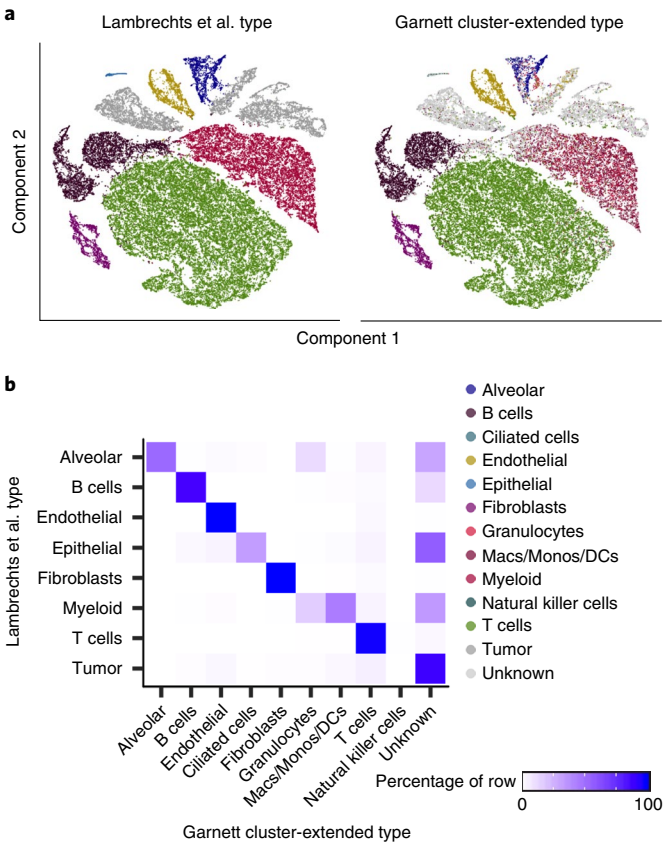

Fig. 3 | Garnett accurately classifies across species and distinguishes normal and pathological tissue. a, Garnett cluster-extended results for human lung tumors from ref. [14] classified based on a Garnett classifier trained on lung cells from the MCA. t-SNE plots of the human lung tumor dataset ($n = 52{,}698$ cells). First panel is colored by published type from ref. [14], second panel is colored by the Garnett cluster-extended classification. **b**, Heatmap comparing the reported cell types versus the Garnett cluster-extended cell types from **a**. Color represents the percentage of cells of a certain reported type labeled as each type by Garnett. DCs, dendritic cells.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0535-3.

Received: 1 February 2019; Accepted: 12 July 2019;
Published online: 09 September 2019

References

1. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. *Nat. Protoc.* **13**, 599–604 (2018).
2. Tabula Muris Consortium *Nature* **562**, 367–372 (2018).
3. Han, X. et al. *Cell* **173**, 1307 (2018).
4. Zhang, X. et al. *Nucleic Acids Res.* **47**, D721–D728 (2019).
5. Diehl, A. D. et al. *J. Biomed. Semant.* **7**, 44 (2016).
6. Bard, J., Rhee, S. Y. & Ashburner, M. *Genome Biol.* **6**, R21 (2005).
7. Cusanovich, D. A. et al. *Cell* **174**, 1309–1324 (2018).
8. Zheng, G. X. Y. et al. *Nat. Commun.* **8**, 14049 (2017).
9. Levine, J. H. et al. *Cell* **162**, 184–197 (2015).
10. Pliner, H. A. et al. *Mol. Cell* **71**, 858–871 (2018).
11. Cao, J. et al. *Science* **357**, 661–667 (2017).
12. Zeisel, A. et al. *Cell* **174**, 999–1014.e22 (2018).
13. Merkin, J., Russell, C., Chen, P. & Burge, C. B. *Science* **338**, 1593–1599 (2012).
14. Lambrechts, D. et al. *Nat. Med.* **24**, 1277–1289 (2018).
15. Rosenberg, A. B. et al. *Science* **360**, 176–182 (2018).
16. Alavi, A., Ruffalo, M., Parvangada, A., Huang, Z. & Bar-Joseph, Z. *Nat. Commun.* **9**, 4768 (2018).

Acknowledgements

We gratefully acknowledge S. Tapscott, W. Noble and D. Witten as well as members of the Shendure and Trapnell laboratories, particularly A. Hill, for their advice. Z. Pliner named the software. This work was supported by the following funding: NIH grant nos. U54DK107979 and U54HL145611 to J.S. and C.T.; NIH grant nos. DP2HD088158, RC2DK114777 and R01HL118342 to C.T.; NIH grant nos. DP1HG007811 and R01HG006283 to J.S. and the Paul G. Allen Frontiers Group to J.S. and C.T. J.S. is an Investigator of the Howard Hughes Medical Institute. C.T. is partly supported by an Alfred P. Sloan Foundation Research Fellowship. H.A.P. was supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-1256082.

Author contributions

C.T. and J.S. conceived the project. H.A.P. wrote Garnett and led the data analysis. H.A.P., C.T. and J.S. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0535-3.

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

Correspondence and requests for materials should be addressed to J.S. or C.T.

Peer review information: Nicole Rusk was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

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

Methods

Garnett. Garnett is designed to simplify, standardize and automate the classification of cells by type and subtype. To train a new model with Garnett, the user must specify a cell hierarchy of cell types and subtypes, which may be organized into a tree of arbitrary depth; there is no limit to the number of cell types allowed in the hierarchy. For each cell type and subtype, the user must specify at least one marker gene that is taken as positive evidence that the cell is of that type. Garnett includes a simple language for specifying these marker genes, to make the software more accessible to users unfamiliar with statistical regression. Negative marker genes, that is, taken as evidence against a cell being of a given type, can also be specified. In addition, Garnett includes tools for selecting and checking the quality of markers. Garnett uses the marker information provided to select cells that are then used to train an elastic-net regression-based classifier[17], similar to the approach taken in ref. [7]. After a classifier is trained, it can be applied to other single-cell datasets run on the same or different platforms. Algorithmic details are provided below.

Constructing marker files. Garnett uses a marker file to allow users to specify cell type definitions. These definitions are then used to choose representative cells from each cell type to use when training the classifier. Full details describing the syntax of the marker file are provided as part of the software package. Briefly, the marker file consists of a series of cell type entries, beginning with a cell type name, followed by lists of expressed markers and metadata. In addition, cell types can be specified to be a subtype of another defined type; that is, hierarchical definitions. Marker files also have the capability to hold literature references for the chosen marker genes that are then included as metadata in the classifier.

Because only markers that are expressed specifically in a given cell type are useful for Garnett classification, we also provide functions for assessing the value of each of the provided marker genes. These functions estimate the number of cells that a given marker nominates for their cell type, the number of cells that become 'ambiguously' nominated to multiple cell types in a given level of the hierarchy when the marker is included, and an overall marker score G, defined as:

$$G = \frac{1}{(a+p)} \times \frac{b}{n}$$

where a is the fraction of cells nominated by the given marker that are made ambiguous by that marker, p is a small pseudocount, b is the number of cells nominated by the marker and n is the total number of cells nominated for that cell type. In addition to estimating these values, Garnett will plot a diagnostic chart to aid the user in choosing markers (for example, Supplementary Fig. 1a).

Training the classifier. Garnett's first step in training a cell type classifier is to choose representative cells to train on. Let M be an m by n matrix of input gene expression data. First, $M_{i,j}$ is normalized by size factor (the geometric mean of the total unique molecular indexes expressed for each cell j) to adjust for read depth, resulting in a normalized m by n matrix N. In addition, the gene IDs of the expression data are converted to Ensembl IDs using correspondence tables from a Bioconductor AnnotationDbi-class[18] package. Next, the input marker file is parsed and the gene IDs are also converted to Ensembl IDs as above. Finally, a tree representation of the marker file is constructed, with any designated subtypes placed as children of the parent cell type in the tree. In addition to the tree, a dataset-wide size factor is generated and saved to the tree to allow normalization to new datasets for later classification (see Classifying cells).

For each parent node in the tree, the following steps are taken: first, cells are scored as 'expressed' or 'not expressed' for each of the provided markers and an aggregate marker score is derived for each cell type for each cell (details on scoring below). Next, any metadata or hard expression cutoffs are applied to exclude a subset of cells from consideration. Last, outgroup samples are chosen (see below). After choosing the training sample, the classifier is trained (see below), and a preliminary classification is made to further train downstream nodes.

Aggregated marker scores. We devised an aggregated marker scoring system to address two challenges of single-cell RNA-seq data for the purposes of identifying representative cell types based on markers. The first challenge when choosing cells is that of differing levels of expression of different markers. If a lowly expressed but specific marker is found in a cell profile, this is better evidence of cell type than a highly expressed and less specific marker. To address this, we use the term frequency-inverse document frequency[19] (TF-IDF) transformation when generating aggregate marker scores. The TF-IDF transformed matrix is defined by,

$$T_{i,j} = \frac{N_{i,j}}{\sum_{i=1}^{m} N_{i,j}} \times \log\left(1 + \frac{n}{\sum_{j=1}^{n} N_{i,j}}\right)$$

where $N_{i,j}$ is the m by n normalized gene expression matrix defined above.

The second challenge we addressed in our aggregate marker score calculation was that highly expressed genes have been known to leak into the transcriptional profiles of other cells. For example, in samples including hepatocytes, albumin transcripts are often found in low copy numbers in non-hepatocyte profiles. To address this, we assign a cutoff above which a gene is considered expressed in that cell. To determine this cutoff we use a heuristic measure defined as

$$C_i = 0.25 \times q_i$$

where C_i is the gene cutoff for gene i and q_i is the 95th percentile of T for gene i. Any gene i in cell j with a value $T_{i,j}$ below C_i is set to 0 for the purposes of generating aggregated marker scores.

After these transformations, the aggregated marker score is defined by a simple sum of the genes defined as markers in the cell marker file,

$$S_{c,j} = \sum_{k \in G_c} T_{k,j}$$

where $S_{c,j}$ is the aggregated score for cell type c and cell j and G_c is the list of marker genes for cell type c. Cells in the 75th percentile and above for aggregated marker score S in only one cell type are chosen as good representatives. Any metadata specifications (for example, the requirement that a cell type have come from a particular tissue) provided in the marker file are then used to exclude cells and generate a final training dataset.

Choosing outgroup cells. When choosing outgroup samples for training, we wanted to make sure that the outgroup set is not dominated by the most abundant cell type. So we cluster a random subset of potential outgroup cells and choose equal numbers of random cells from each cluster to make up the outgroup. Specifically, we first calculate the first 50 principal components using principal components analysis as implemented by the irlba[20] R package. Next, we calculate jaccard coefficients on a k nearest-neighbors (kNN) graph generated using $k = 20$. Last, we generate clusters using Louvain community detection on the resulting cell–cell map of jaccard coefficients. A random set of cells from each resulting community is then combined to create the outgroup.

Training with GLMnet. The classifier is trained on the normalized expression matrix N for cells chosen as representatives, and for all genes expressed in greater than 5% of cells in at least one training set and not expressed in the 90th percentile of TF-IDF transformed expression in all cell types. This last filter prevents ubiquitously expressed genes from being chosen as features. The classifier is trained using genes as features and cells as observations with a grouped multinomial elastic-net regularized ($\alpha = 0.3$)[17] generalized linear model using the package GLMnet[21] in R. Observations are weighted by the geometric mean of the counts in each of the training groups. The GLMnet regularization parameter λ is chosen using three-fold cross validation. Genes provided in the marker files are required to be included in the model and are not regularized.

Classifying cells. Because we wished to be able to use pre-trained classifiers to classify cells across datasets and platforms, we include a dataset size factor D for the training data with the classifier object. D is the geometric mean of the total read counts per cell divided by the median number of genes expressed above zero per cell. Formally, D is defined by

$$a = \sum_{i=1}^{m} M_{i,j}$$

$$D = \exp\left[\frac{1}{j}\sum_{k=1}^{j} \ln a_k\right] \times \frac{1}{\text{median}\{g\}}$$

where g is the number of genes expressed above zero per cell. When applying an existing classifier to a new dataset, we can then transform the new expression data, an m' by n' matrix M', to the scale of the training data using D

$$f_j = \frac{\sum_{i=1}^{m'} M'_{i,j}}{D \times \text{median}\{g'\}}$$

$$N' = \frac{M'}{f_j}$$

where g' is the number of genes expressed above zero per cell in the new data.

After normalization, gene IDs for the new dataset are also converted to Ensembl IDs. At each internal node in the classifier, the multinomial model for that node is applied to the data, the output probabilities of each class are normalized by dividing by the maximum probability for each cell, and the ratio of the top scoring cell type to the second-best scoring cell type is calculated. If this odds ratio is greater than the user-specified rank probability ratio (in this paper and by default, we use 1.5), the top type is assigned, otherwise the cell type is set to 'Unknown'. Optionally, Garnett will add a second set of classifications that classify an entire cluster of cells if greater than 90% of assigned cells within a cluster are the same type and greater than 5% of all cells in the cluster are classified (not 'Unknown') and greater than five cells in the cluster are classified. Cluster labels can be

provided by the user or generated by Garnett using Louvain community detection in the top 50 principal components of the expression matrix.

10X PBMCs. The 10X PBMC datasets from both v.1 and v.2 chemistry were downloaded from the 10X Genomics website. The v.1 cells are a combination of each of the pure cell type populations isolated by 10X Genomics using fluorescence-activated cell sorting (FACS) sorting ($CD14^+$ Monocytes, $CD19^+$ B cells, $CD34^+$ cells, $CD4^+$ Helper T cells, $CD4^+/CD25^+$ Regulatory T cells, $CD4^+/CD45RA^+/CD25^-$ Naive T cells, $CD4^+/CD45RO^+$ Memory T cells, $CD56^+$ Natural killer cells, $CD8^+$ Cytotoxic T cells and $CD8^+/CD45RA^+$ Naive cytotoxic T cells) preprocessed using CellRanger v.1.1.0 and published in ref. [8]. The v.2 cells are the v.2 chemistry distributed demonstration dataset labeled '8k PBMCs from a healthy donor', preprocessed using CellRanger v.2.1.0. Markers for PBMCs were those often cited in the literature. Using Garnett's marker scoring system, we excluded the markers with high ambiguity (>0.5). The final PBMC marker file used is available as Supplementary Dataset 1. Garnett classification for v.1 and v.2 was run using default parameter values defined in the preceding sections. For testing the limits of Garnett (Supplementary Fig. 2), we used a one-level marker file with no T cell subtypes defined. To downsample the reads in training and classification datasets (Supplementary Fig. 2d,e), we used the downsampleMatrix function from the DropletUtils[22,23] R package, which uses sampling without replacement per cell so that the total reads in that cell is reduced by the specified proportion. For the T cell ablation experiment (Supplementary Fig. 2g,h) we removed the T cells from the matrix and also set all values of the T cell marker genes to zero.

TM and MCA lung analysis. The TM FACS dataset from ref. [2] was downloaded from their figshare website. The MCA dataset from ref. [3] was downloaded from their figshare website. For the purposes of this analysis, only data derived from lung tissue from both datasets were used. To facilitate comparisons between each of the lung datasets used, a set of consensus cell type names were used as described in Table 1. The marker file used is available as Supplementary Dataset 2. Garnett classification was run using default parameter values for both datasets.

sci-ATAC-seq analysis. The sci-ATAC-seq data was downloaded from the website associated with ref. [7]. The input to Garnett was the previously calculated Cicero gene activity scores presented in the original publication. The final marker file used is available as Supplementary Dataset 3. Garnett classification was run using default parameter values.

Worm analysis. The worm data was downloaded from the website associated with ref. [11]. Markers were those used by the original publication to identify cell types. Using Garnett's marker scoring system, we excluded the markers with high ambiguity (Supplementary Fig. 6). The final marker file used is available as Supplementary Dataset 4. Garnett classification was run using default parameter values.

Human lung tumor analysis. The human lung tumor data was downloaded from the ArrayExpress database entry associated with ref. [14]. Because expression data were log-transformed, we exponentiated the expression data before classification. To allow for cross-species classification, we first converted the human expression data to mouse gene labels by creating a correspondence table using the biomaRt hsapiens_gene_ensembl and mmusculus_gene_ensembl databases. Only unique rows (one-to-one correspondences) were used. Ultimately 15,336 of the original 22,180 human genes could be converted to mouse labels including 89% of the genes in the MCA classifier with non-zero coefficients. The final marker file used is available as Supplementary Dataset 5. Garnett classification was run using default parameter values.

Mouse nervous system analysis. The mouse nervous system data and potential marker genes were downloaded from the website associated with ref. [12]. The final marker file used is available as Supplementary Dataset 6.

Software availability. Garnett is an R package available through github.

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

Data availability

No new data was generated for this study. All data used in this study is publicly available.

References

17. Zou, H. & Hastie, T. *J. R. Stat. Soc. Ser. B.* **67**, 301–320 (2005).
18. Carlson, M., Falcon, S., Pages, H. & Li, N. *AnnotationDbi: Annotation Database Interface*. R package v.1.44.0 (2018).
19. Jones, K. S. *J. Doc.* **28**, 11–21 (1972).
20. Baglama, J., Reichel, L. & Lewis, B. W. *irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices.* R package v.2.3.3 (2017).
21. Friedman, J., Hastie, T. & Tibshirani, R. *J. Stat. Software* **33**, 1–22 (2010).
22. Lun, A. et al. *Genome Biol.* **20**, 63 (2019).
23. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. *Nat. Commun.* **9**, 2667 (2018).

nature research

Corresponding author(s): Cole Trapnell
Jay Shendure

Last updated by author(s): Feb 4, 2019

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 Authors & Referees 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 new data was collected for this study, all data used is publicly available.
Data analysis	The Garnett software is available on GitHub at https://github.com/cole-trapnell-lab/garnett.

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/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

No new data was generated for this study. All data used in this study is publicly available.

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 new data was generated as part of this study.
Data exclusions	No data was excluded.
Replication	Classifiers were tested against new datasets, and the results are shown in the manuscript.
Randomization	There were no study groups, randomization was therefore not relevant.
Blinding	There were no study groups, blinding was therefore not relevant.

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
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

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