nature methods

ARTICLES

https://doi.org/10.1038/s41592-019-0355-5

# Cell composition analysis of bulk genomics using single-cell data

Amit Frishberg[1,7], Naama Peshes-Yaloz[1,7], Ofir Cohn[1], Diana Rosentul[1], Yael Steuerman[1], Liran Valadarsky[2], Gal Yankovitz[1], Michal Mandelboim[3,4], Fuad A. Iraqi[5], Ido Amit[2], Lior Mayo[1,6], Eran Bacharach[1,8]* and Irit Gat-Viks[1,8]*

**Single-cell RNA sequencing (scRNA-seq) is a rich resource of cellular heterogeneity, opening new avenues in the study of complex tissues. We introduce Cell Population Mapping (CPM), a deconvolution algorithm in which reference scRNA-seq profiles are leveraged to infer the composition of cell types and states from bulk transcriptome data ('scBio' CRAN R-package). Analysis of individual variations in lungs of influenza-virus-infected mice reveals that the relationship between cell abundance and clinical symptoms is a cell-state-specific property that varies gradually along the continuum of cell-activation states. The gradual change is confirmed in subsequent experiments and is further explained by a mathematical model in which clinical outcomes relate to cell-state dynamics along the activation process. Our results demonstrate the power of CPM in reconstructing the continuous spectrum of cell states within heterogeneous tissues.**

Single-cell RNA sequencing (scRNA-seq) provides a powerful approach to understand the composition of different cell identities within a complex tissue, including discrete cell types, cell states that arise transiently during the progression of time-dependent processes, and continuous dynamic transitions within the space of possible cell states[1,2]. The frequency of cell types and cell states may vary between genetically distinct individuals, environments, chemical perturbations, and disease states. To investigate this variation at high resolution, it is possible to generate scRNA-seq profiles for each sample of interest and then use it to evaluate the frequency of the different cell types and states[3–5]. However, such studies are costly and time-consuming, and have therefore been performed only on a limited scale.

An alternative strategy is to construct a comprehensive collection of reference scRNA-seq profiles representing various cell types and cell states. Deconvolution algorithms can then use the reference profiles to computationally predict the abundance of different cell types and states within a given sample, based on only the bulk expression data from that sample[2,6–8]. This strategy should in principle avoid the scaling issues associated with multiple scRNA-seq experiments, but in practice, using a large number of reference profiles typically results in reduced prediction accuracy[9]. A standard solution is to cluster the single-cell reference profiles into a relatively small number of cell-group reference profiles[10–12]. However, although this clustering-based approach may provide a rough quantification of discrete cell types and states, the continuous cell-state space remains sparse and fragmented. Therefore, there is a substantial need for a deconvolution method that can exploit the rich spectrum of single-cell reference profiles.

Here we introduce the Cell Population Mapping (CPM) method, which provides an advantageous alternative to existing deconvolution approaches, particularly in providing a fine-resolution mapping. Similarly to approaches described in recent studies[10–12], CPM constructs its reference collection from scRNA-seq profiles derived from one or a few relevant samples, and then exploits this collection to infer cell composition within additional, bulk-profiled samples. However, instead of focusing on quantifying a few dozen discrete cell subtypes, CPM analyzes thousands of single-cell profiles scattered across the wide landscape of cell states. Using synthetic data, we demonstrate that deconvolution with CPM significantly improves the quantification of both gradual and abrupt changes in cell abundance over the continuous space of cell types and states. Furthermore, by analyzing complex changes in lung tissues, across influenza-virus-infected mice of various genetic backgrounds, we demonstrate the effectiveness of CPM in probing phenotypic diversity in large cohorts.

## Results

**Overview of CPM.** We developed CPM, a method based on computational deconvolution for identifying a cell population map from bulk gene expression data of a heterogeneous sample. In our framework, the cell population map is the abundance of cells over a cell-state space. Cell type is defined as the core characteristics of a cell, whereas a cell state can be thought of as the current phenotype of a given cell type (for example, it can refer to the proliferation, activation, or differentiation state of the cell)[1]. The cell-state space specifies each cell state as a point in a multi-dimensional space; as cells undergo changes from one state to another, they travel through the space along a trajectory between these two states[13]. Unlike existing computational methods that focus on reconstruction of the cell-state space from scRNA-seq data[1], CPM takes as its input the previously reconstructed cell-state space of certain scRNA-seq data,

[1]School of Molecular Cell Biology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel. [2]Department of Immunology, The Weizmann Institute of Science, Rehovot, Israel. [3]National Center for Influenza and Respiratory Viruses, Central Virology Laboratory, Sheba Medical Center at Tel HaShomer, Ramat-Gan, Israel. [4]Department of Epidemiology and Preventive Medicine, School of Public Health, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [5]Department of Clinical Microbiology and Immunology, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [6]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel. [7]These authors contributed equally: Amit Frishberg, Naama Peshes-Yaloz. [8]These authors jointly supervised this work: Eran Bacharach, Irit Gat-Viks. *e-mail: eranba@tauex.tau.ac.il; iritgv@post.tau.ac.il

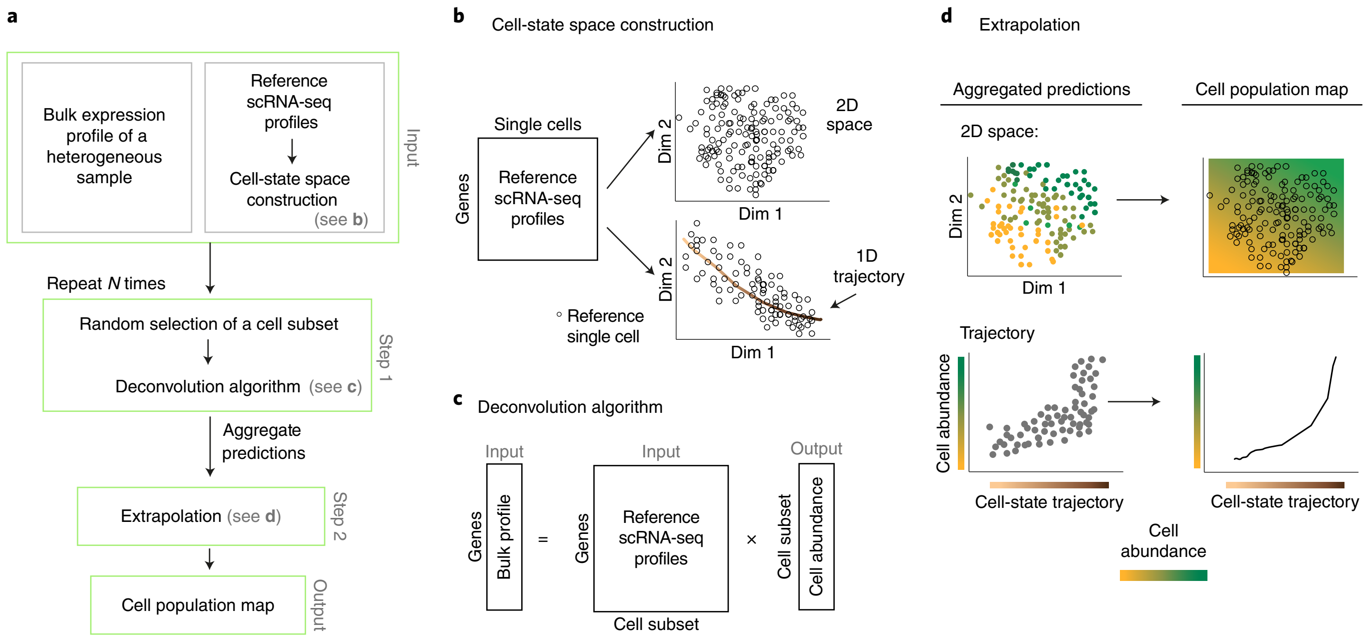

**Fig. 1 | Overview of the CPM algorithm.** **a**, A flowchart of the CPM pipeline. **b–d**, Illustration of the steps outlined in **a**; cell-state space construction (**b**), deconvolution (**c**), and extrapolation (**d**). Dim, dimension.

and then relies on this input to infer the abundance of each point in this space within a given bulk cell population.

Formally, CPM relies on two input types (Fig. 1a): first, a bulk expression profile of the heterogeneous cell population, and second, scRNA-seq profiles of individual single cells derived from one or a few representative samples ('reference data'). We assume that the cell-state space of the reference cells is given as input and that the particular position of each reference single cell within this space is known. The cell-state space is typically obtained by dimension reduction (such as t-SNE[14]) that captures the essence of gene-regulation variation among the reference single cells (exemplified in Fig. 1b, top). It is also possible to use a well-defined trajectory within this space as an alternative 'one-dimensional (1D) space'; such a trajectory can explicitly describe the progression of cells through a biological process (Fig. 1b bottom).

CPM consists of two steps (Fig. 1a and Methods). The first step is application of a deconvolution approach (here, the support vector regression (SVR) approach) that combines the bulk profile of a complex tissue with a collection of reference scRNA-seq profiles to infer the composition of cells within the complex tissue input. The output of this step is the abundance of cells at the sub-region of each reference single cell. Such prediction poses two substantial challenges: accuracy in deconvolving a very large number of reference profiles (typically, the analysis may involve thousands of single cells[15]), and a potential bias owing to the non-uniform distribution of reference cells over the cell-state space. To address these challenges, CPM applies deconvolution using a relatively small subset of reference profiles, which are obtained by unbiased random sampling to ensure that every region in the cell space has an equal chance of being sampled (Fig. 1c). The sampling and deconvolution procedures are repeated $N$ times (here, $N = 1{,}500$), and the results are aggregated and averaged into a single inferred abundance for each reference single cell. The second step is to extrapolate, from the inferred cell abundance in each particular reference coordinate, the cell abundance in any other cell-state coordinate (Fig. 1d). In this extrapolation it is assumed that the shape of the cell distribution over the cell-state space is continuous and smooth. We refer to this smoothed continuous output as the 'cell population map'. Notably, CPM may use input bulk data either as a relative profile (that is, response between two samples) or as an absolute (single-sample) profile, thereby predicting relative or absolute cell-state abundance, respectively.

**Performance analysis.** We used a simulation framework to measure the ability of CPM to predict the cell population map at fine resolution. The simulation was based on a collection of 1,860 reference scRNA-seq profiles[16] taken from murine lungs during influenza virus infection and encompassing nine major immune and non-immune cell types: fibroblasts, epithelial cells, blood and lymphatic endothelial cells, T cells, B cells, natural killer (NK) cells, granulocytes, and cells of the mononuclear phagocyte system (MPS), encompassing monocytes (MO), macrophages (MΦs), and dendritic cells. In this reference dataset, one well-characterized trajectory is the gradual transition of cell states (within each cell type) from resting (naive)-like cells into active cells that respond to the influenza virus infection[16]. We created synthetic bulk profiles of a complex tissue by mixing these single-cell profiles according to predetermined biologically relevant functions over the cell-state space, introducing noise in the expression of genes and in the coordinates of single cells (denoted 'expression noise' and 'cell space noise', respectively). The quality of this strategy is demonstrated in Supplementary Fig. 1a and further discussed in Supplementary Note 1. CPM was compared to three alternative deconvolution methods—DCQ[17], Cibersort[18] and standard SVR[19]—whose reference collection was the 'averaged' profiles of single-cell groups (the larger the number of such single-cell groups, the higher the 'granularity' analyzed by the alternative methods; Methods). We evaluated the 'accuracy' of predictions by comparing the ground-truth cell abundance to the predicted abundance of each reference single cell.

To analyze performance, we focused on three fundamental types of simulations: the 'cell-type simulation', in which cell abundance varies from one cell type to another, but within each cell type, the abundance is uniformly distributed over the cell-state space; the 'cell-subtype simulation', consisting of a modified abundance of a subpopulation of cell states within selected cell types; and the 'gradual-change simulation', representing continuous alterations of cell abundance along the trajectory of cell-activation states (within

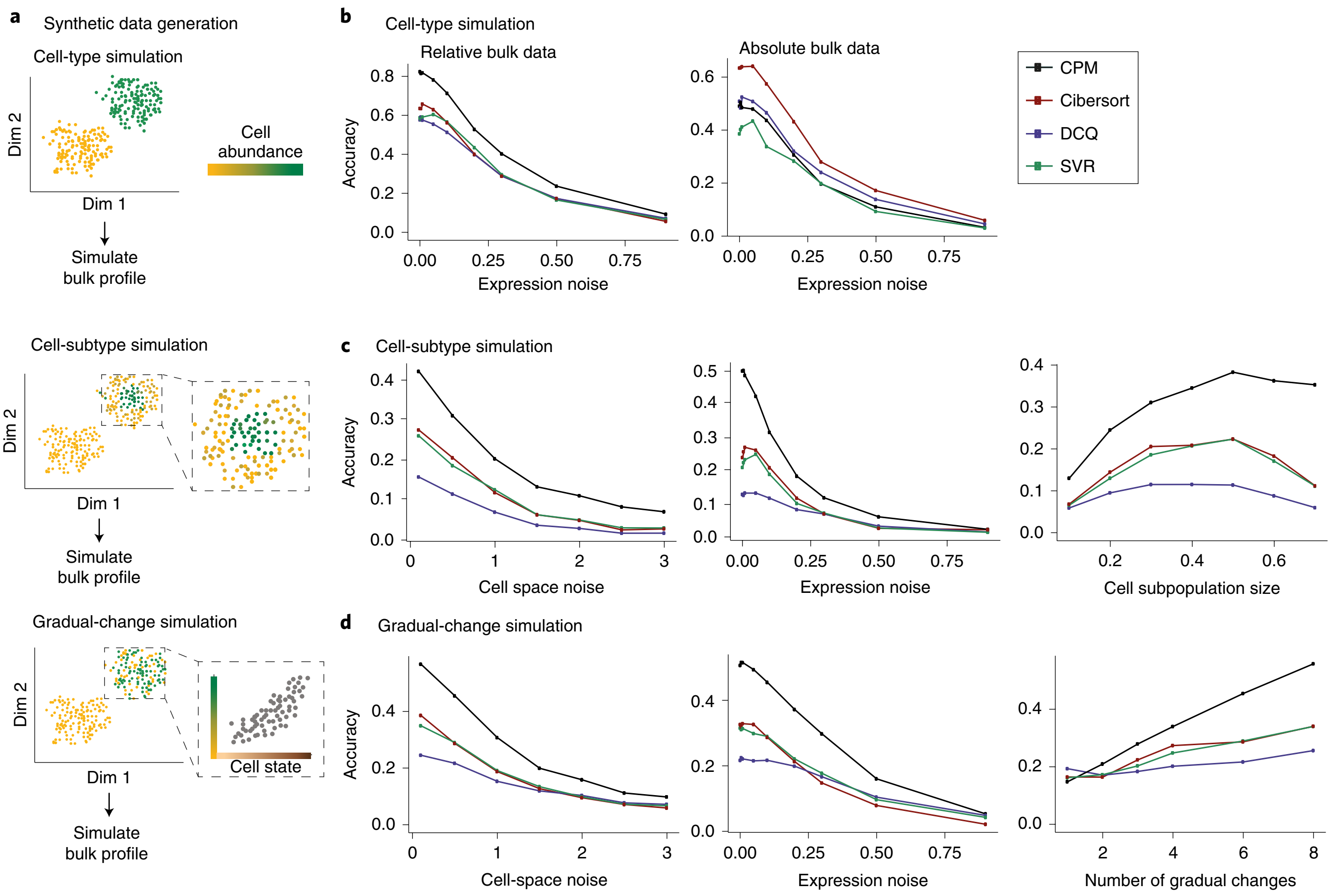

**Fig. 2 | Performance assessed via synthetic data. a**, We generated synthetic data by introducing changes in the percentage of discrete cell types ('cell-type simulation', top); changes in the percentage of cell subtypes, within cell types ('cell-subtype simulation', middle); or changes in the percentage along a trajectory of cell states ('gradual-change simulation', bottom). Illustration and abbreviation are as in Fig. 1. **b–d**, Accuracy of inferring cell abundance for the three simulation types; cell-type simulation (**b**), cell-subtype simulation (**c**), and the gradual-change simulation (**d**). Accuracy is defined as the Pearson correlation coefficient between predicted and true cell abundance and is shown across varying data parameters, such as the level of inaccuracy ("noise") that was synthetically generated (*x* axis) for alternative deconvolution methods. Results are shown for bulk relative profiles (**b** (left), **c**, and **d**) or absolute profiles (**b**, right). The alternative methods were applied with a reference dataset that was generated using granularity of four cell groups.

selected cell types) (Fig. 2a). Overall, the cell-type simulation is focused on inter-cell-type variation, whereas the cell-subtype and gradual-change simulations are focused on intra-cell-type variation, which arises from differences among cell states within the same cell type.

Consistent with previous observations, changes in discrete cell types were accurately modeled by the alternative deconvolution methods (Fig. 2b and Supplementary Fig. 1b). However, in the case of intra-cell-type changes in the composition of cell states (the 'cell-subtype' and 'gradual-change' simulations), CPM showed consistent improvement in prediction accuracy compared to existing deconvolution methods (relative bulk data, Fig. 2c,d and Supplementary Fig. 2a,b; absolute bulk data, Supplementary Fig. 3a,b) within a reasonable running time (Supplementary Fig. 2c,d). Unsurprisingly, CPM was able to capture the continuous nature of the input tissue, unlike alternative deconvolution methods, which could provide only a discrete approximation with lower accuracy (Supplementary Figs. 2e and 3c). Furthermore, CPM outperformed the existing methods in its ability to handle a high cell-state complexity and in its 'scalability' to a large number of reference profiles (Supplementary Fig. 2f,g and detailed in Supplementary Note 1). Quantitatively similar results were also observed for varying parameter settings (Supplementary Fig. 4), using different cell-state space solutions (Supplementary Fig. 5a), and for regions of different local density within the cell-state space (Supplementary Fig. 5b,c). Of note, CPM may lose power with lower sequencing depth (Supplementary Fig. 6 and discussed in Supplementary Note 1).

**Cell-state-specific relationships between infection symptoms and cell abundance.** We applied CPM to investigate in vivo influenza virus infection across the Collaborative Cross (CC) recombinant inbred strains[20], a panel of mouse lines designed to mimic the phenotypic and genotypic diversity seen in human populations. To this end, we generated bulk transcriptional expression profiles derived from lung tissues of 38 infected and 34 phosphate-buffered saline (PBS)-treated control mice (typically one or two individuals of each CC strain; Supplementary Table 1 and Methods). We transformed absolute bulk profiles into relative bulk profiles using a common control profile as the normalizer, and then applied CPM to each of these bulk profiles using the aforementioned single-cell measurements of the same experimental setting (lung tissues from influenza-virus-infected mice[16]) as the reference data. As the cell-state space, CPM used the continuous sequence of cell-activation states that were previously defined for each of the nine immune and non-immune cell types in this reference dataset[16]. Altogether, CPM calculated a relative cell population map consisting of relative cell abundance in each cell state for each individual mouse. We found that CPM predictions varied considerably between individuals (see

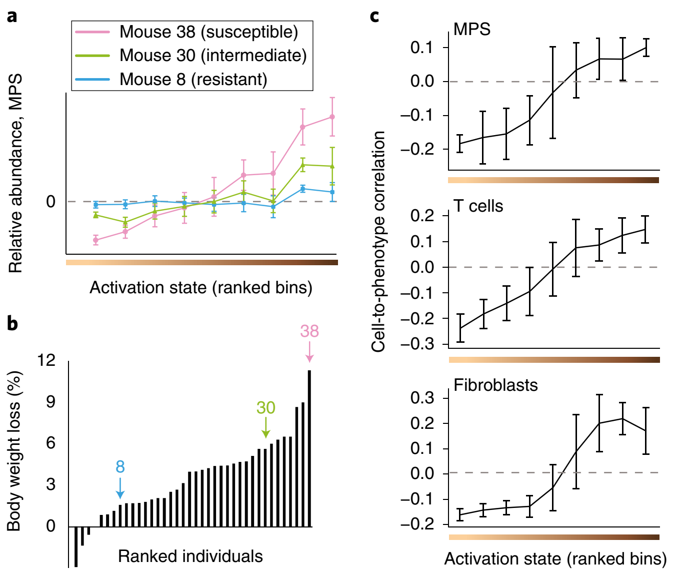

**Fig. 3 | Cellular heterogeneity during in vivo influenza virus infection, reconstructed by CPM. a**, CPM-inferred relative MPS abundance values averaged over cells from each activation-state bin (bins were ranked with increasing activation states from left to right), for three representative infected individuals. $n = 103$ cells; error bars, s.d. **b**, Percentage body weight loss of 38 infected individuals, ranked by disease severity. Marked individuals are the three shown in **a**. **c**, Cell-to-phenotype Pearson correlation coefficients across the 38 infected CC mice, averaged by activation state bins, presented for MPS cells (top, 103 cells), T cells (middle, 378 cells) and fibroblasts (bottom, 375 cells). Error bars, s.d.

examples in Fig. 3a) and that this variation was robust across the *N* deconvolution repeats (Supplementary Fig. 7a).

Although the inferred cell population maps demonstrated substantial variation, the extent to which these cell-state changes related to the clinical outcome of disease remained unclear. To elucidate this point, we monitored one of the main clinical symptoms of murine infection, namely, body weight loss (measured at 2 d post-infection (p.i.); Fig. 3b), and calculated the correlation (across individuals) between this outcome and the inferred relative abundance of each reference cell (denoted the 'cell-to-phenotype correlation'). By splitting the reference cells into consecutive activation-state intervals (within each cell type), we could assess the variation in cell-to-phenotype correlations over the activation trajectory (illustrated in Supplementary Fig. 7b). Intriguingly, cell-to-phenotype correlations across infected mice clearly showed a gradual increase over the trajectory of cell-activation states, from negative correlations at the lower (naive-like) range to positive correlations at the upper (activated) range (mainly in T cells, MPS, and fibroblasts; Fig. 3c). In fact, no particular threshold could be found that split the activation-state trajectory into two discrete groups in which cell-to-phenotype correlations did not gradually change. Similar conclusions about the gradual change in cell-to-phenotype correlations were obtained using a second public dataset of influenza virus infection[21] and using additional computational analyses (Supplementary Note 1 and Supplementary Fig. 7c–h). As expected, the use of unrelated (uninfected) reference datasets and alternative deconvolution methods did not yield the same conclusions (Supplementary Note 1 and Supplementary Fig. 7i,j). Taken together, these findings highlight the advantage of a CPM model that is based on a continuous space of cell states, and further emphasize the importance of using reference and bulk data derived from a similar experimental setting.

**Experimental validation of predictions.** To test for the presence of a gradual change in cell-to-phenotype relationships as predicted by CPM, we performed flow cytometry analyses of lung cells from influenza-infected CC mice at 2 d p.i. We focused on MO/MΦ cell types, which constitute a major fraction of the total MPS population. To determine the activation states of MO/MΦs, we used flow cytometry with two established cell-activation markers, CD64 and Ly6C[22]. The use of these markers enabled us to quantify the distribution of cells over a trajectory of activation states ranging from non-inflammatory ($\text{CD64}^{\text{low}}\text{Ly6C}^{\text{low}}$) to inflammatory ($\text{CD64}^{\text{high}}\text{Ly6C}^{\text{high}}$) MO/MΦs. As expected, the fraction of inflammatory MO/MΦs was higher in infected mice than in PBS-treated controls (Fig. 4a,b and Supplementary Table 1). Encouraged by this observation, we used the flow cytometry measurements to calculate the cell-to-phenotype correlation, that is, the correlation between the clinical readout (weight loss at 2 d p.i.) and the MO/MΦ cell fractions enumerated by flow cytometry across infected individuals. The correlation analysis yielded several lines of evidence that validated the reconstruction by CPM: inflammatory MO/MΦs showed a positive cell-to-phenotype correlation ($r^2 = 0.67$, Fig. 4c left); non-inflammatory MO/MΦs had a negative cell-to-phenotype correlation ($r^2 = -0.54$; Fig. 4c, right); cell-to-phenotype correlations increased with each of the two separate activation markers (Fig. 4d); and flow cytometry analysis confirmed a gradual increase in correlation values over the CD64-cell-state continuum for both $\text{Ly6C}^{\text{high}}$ and $\text{Ly6C}^{\text{low}}$ cell states (Fig. 4e). The lack of cell-to-phenotype correlation obtained when we used the total MO/MΦs count ($r^2 = 0.1$; Supplementary Fig. 8a) further validated the contention that cell-to-phenotype relationships depend on particular cell-activation states, thus accentuating the importance of fine-resolution deconvolution mapping. The observed association between activated inflammatory MO/MΦs and severe physiological responses (Fig. 4c, left) has been reported previously[23–25], whereas the opposite trend of naive MO/MΦs (Fig. 4c, right) and the continuous transition between negative and positive correlations over the activation process (Fig. 4e) have not been described previously.

**Inferring dynamics with a Markov model.** Given that the CPM-reconstructed map yielded accurate predictions for MPS cells, we next investigated the temporal dynamics over the activation trajectory for these cells. Our results showed that the association between cell abundance and weight loss varied in a gradual manner along the MPS-activation process (Figs. 3c and 4e), but that the total MPS counts did not correlate with body weight loss (Supplementary Fig. 8a). A parsimonious explanation for this observation is that the phenotypic diversity is associated with inter-individual variation in temporal dynamics along the activation process; for example, inter-individual variation in onset times, or in cell-state progression rates. Like scRNA-seq data[26], the CPM-reconstructed data provided valuable information that allowed such temporal dynamics to be computationally reconstructed. For instance, we focused here on cell-state progression, and because its underlying mechanism is a stochastic process, we assumed a Markov process of naive-to-activation transitions between consecutive cell states (Supplementary Fig. 8b and Methods). We used this model to predict the probability of transition ('transition rate') between sequential states in each individual mouse (see examples in Supplementary Fig. 8c). With the assumption that the activation-onset time in all individuals is the same, comparison of the inferred transition rates might reveal complex transition-rate-to-phenotype relationships (Supplementary Fig. 8d). By calculating transition rates based on CPM predictions, we found that weight loss was indeed positively correlated with transition rates over a wide range of the MPS activation axis (at its early and intermediate parts; Supplementary Fig. 8e). These CPM-predicted transition rates closely matched the rates calculated from flow cytometry measurements (Supplementary Fig. 8e,f). Overall, this theoretical approach suggests a mechanistic model of in vivo

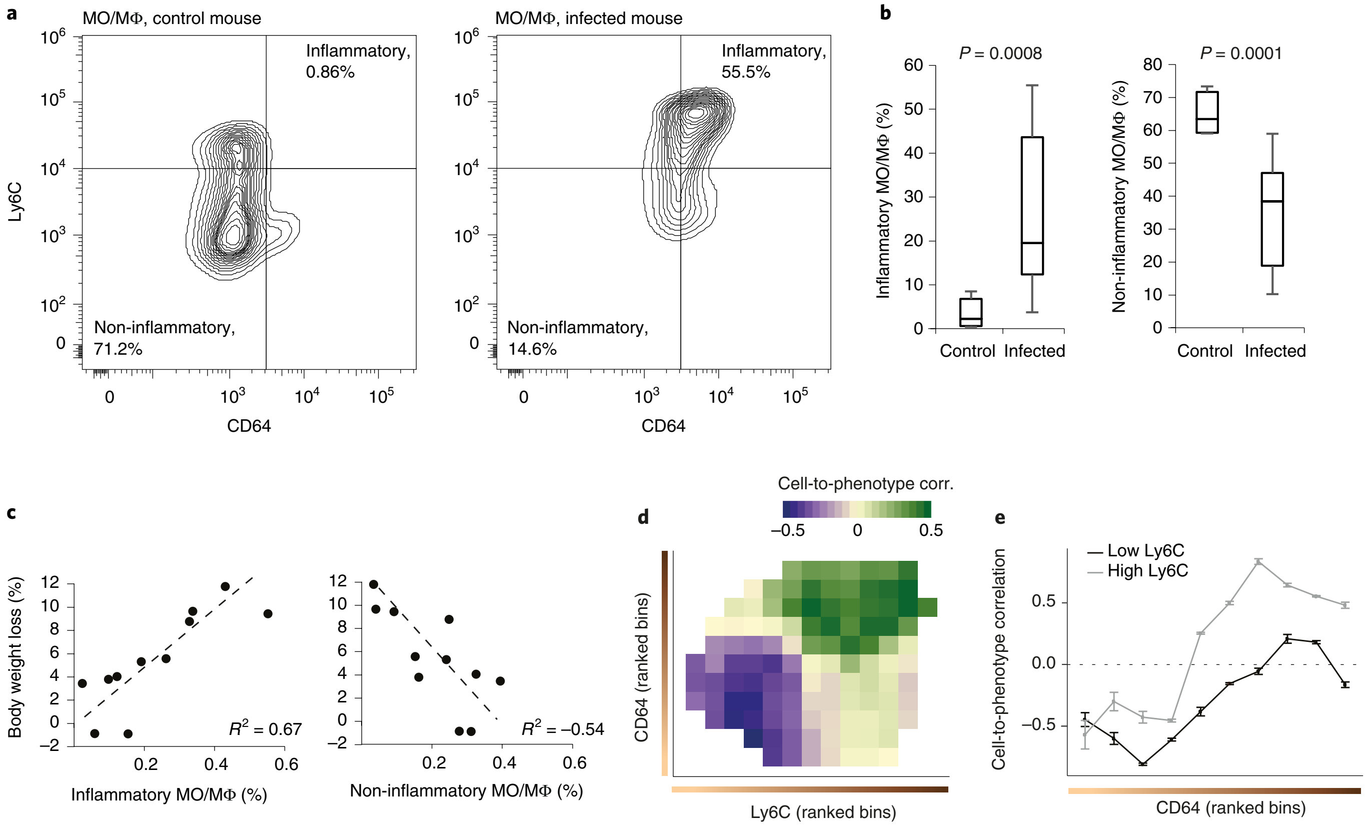

**Fig. 4 | Confirmation of gradual changes in relationships of cells to physiology over a trajectory of cell-activation states.** Flow cytometry analyses of lung-derived MO/MΦs, stained for CD64 and Ly6C activation markers. Cell percentages were calculated relative to the entire population of lung MO/MΦs. **a**, Representative analyses of control and infected animals (CC line 5001 A). Statistics for the remaining individuals are shown in **b**. **b**, Box plots showing the percentages of inflammatory MO/MΦs (CD64$^+$Ly6C$^+$; left) and non-inflammatory MO/MΦs (CD64$^-$Ly6C$^-$; right) in 11 infected and 5 control CC animals. Boxes represent the 25th, 50th, and 75th percentiles; whiskers show maxima and minima. One-sided *t*-test. **c**, Percentage body weight loss of infected individual mice as a function of their percentage inflammatory (left) and non-inflammatory (right) MO/MΦs. **d**, Cell-to-phenotype correlation coefficients (calculated across the infected mice), binned and ranked according to the levels of the activation markers (CD64, Ly6C) and color-coded in each bin. **e**, Cross-sections of the 2D map in **d**. Data are mean ± s.d. across *n* = 100 bootstrapped samples.

influenza-outcome diversity and demonstrates a general strategy for uncovering inter-individual variation in temporal dynamics.

## Discussion

CPM enables the reconstruction of cellular heterogeneity at fine resolution in many bulk-profiled samples, relying on reference single-cell data from only one or a few representative samples (Fig. 1). We attribute the success of CPM to two main advantages. First, the resolution of cell states enables the identification of alterations that are otherwise impossible to detect. Second, CPM's continuous cell-state mapping is well suited for biological scenarios in which the underlying mechanisms drive continuous transitions between cell types or states (for example, a differentiation/developmental continuum, or temporal dynamics during fast response to environmental stimuli). Indeed, we have demonstrated how the association between disease severity and cell-state dynamics provides new insights into the progression of infection.

We believe that the basic framework of CPM will spur the development of additional methods and will prove useful for additional analyses. First, CPM infers the quantities of cell states, a prerequisite for additional studies such as calculations of cell-state-specific expression within complex tissues[27] and reconstruction of temporal dynamics of cell-state progression. Second, an exciting extension of CPM would be the development of a 'functioning tissue' model through integration of cell states together with their spatial organization. To address this, CPM would require significant extension by incorporating the 3D organization as another factor of the cellular state. This is a difficult task because a given spatial location may include different cell types and states. Third, although CPM outperforms existing methods, we expect that additional extensions would enable improved performance (for example, using multiple splicing isoforms rather than the canonical isoform of each gene). Finally, CPM provides a framework to study cellular heterogeneity within the massive body of existing bulk genomics data such as in TCGA[28] and GTeX[29]. As extensive single-cell catalogues (for example, the Human Cell Atlas[30]) are currently in construction, it may soon become possible to analyze cellular heterogeneity in bulk expression data without requiring expertise in single-cell technologies (discussed in Supplementary Note 1).

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0355-5.

Received: 10 June 2018; Accepted: 12 February 2019;
Published online: 18 March 2019

## References

1. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).

2. Chen, X., Teichmann, S. A. & Meyer, K. B. From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Annual Review of Biomedical Data Science* **1**, 29–51 (2018).
3. Krieg, C. et al. High-dimensional single-cell analysis predicts response to anti-PD-1immunotherapy. *Nat. Med.* **24**, 144–153 (2018).
4. Shalek, A. K. & Benson, M. Single-cell analyses to tailor treatments. *Sci. Transl. Med.* **9**, eaan4730 (2017).
5. Kim, K.-T. et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome. Biol.* **17**, 80 (2016).
6. Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.* **25**, 571–578 (2013).
7. Baron, M. et al. A single-cell transcriptomic map of the human and mouse Pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360 (2016).
8. Frishberg, A., Brodt, A., Steuerman, Y. & Gat-Viks, I. ImmQuant: a user-friendly tool for inferring immune cell-type composition from gene-expression data. *Bioinformatics* **32**, 3842–3843 (2016).
9. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969–1979 (2018).
10. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck. *Cancer Cell* **171**, 1611–1624 (2017).
11. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
12. Schelker, M. et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* **8**, 2032 (2017).
13. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
14. Rostom, R., Svensson, V., Teichmann, S. A. & Kar, G. Computational approaches for interpreting scRNA-seq data. *FEBS Lett.* **591**, 2213–2225 (2017).
15. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
16. Steuerman, Y. et al. Dissection of influenza infection in vivo by single-cell RNA sequencing. *Cell Syst.* **6**, 679–691.e4 (2018).
17. Altboum, Z. et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.* **10**, 720 (2014).
18. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
19. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.* **9**, 1871–1874 (2008).
20. Welsh, C. E. et al. Status and access to the collaborative cross population. *Mamm. Genome* **23**, 706–712 (2012).
21. Bottomly, D. et al. Expression quantitative trait loci for extreme host response to influenza a in pre-collaborative cross mice. *G3 (Bethesda)* **2**, 213–221 (2012).
22. Yu, Y.-R. A. et al. A protocol for the comprehensive flow cytometric analysis of immune cells in normal and inflamed murine non-lymphoid tissues. *PLoS ONE* **11**, e0150606 (2016).
23. Ferris, M. T. et al. Modeling host genetic regulation of influenza pathogenesis in the collaborative cross. *PLoS Pathog.* **9**, e1003196 (2013).
24. Dengler, L. et al. Cellular changes in blood indicate severe respiratory disease during influenza infections in mice. *PLoS ONE* **9**, e103149 (2014).
25. Coates, B. M. et al. Inflammatory monocytes drive influenza a virus-mediated lung injury in juvenile mice. *J. Immunol.* **200**, 2391–2404 (2018).
26. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
27. Shen-Orr, S. S. et al. Cell type–specific gene expression differences in complex tissues. *Nat. Methods* **7**, 287–289 (2010).
28. Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
29. eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat. Genet.* **49**, 1664–1670 (2017).
30. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).

## Acknowledgements

This work was supported by the European Research Council (637885) (to A.F., N.P.-Y., O.C., D.R., Y.S., and I.G.-V.), and partially supported by the Israeli Centers of Research Excellence (I-CORE) Center No 41/11 (to A.F. and Y.S.), the Broad-Israel Science Foundation (ISF) 1168/14 (to D.R. and N.P.-Y.), ISF 1824/13 (to E.B.), partial fellowships from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University (to A.F. and Y.S.), and a Shulamit Aloni Scholarship (to Y.S.). Research in the Gat-Viks laboratory was supported by ISF 288/16. I.G.-V. is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and an Alon Fellow. We thank O. Danziger and H.J. Abu-Toamih Atamni (Tel Aviv University) for help with mouse work, and S. Smith for scientific editing.

## Author contributions

A.F., N.P.-Y., E.B. and I.G.-V. conceived and designed the study. N.P.-Y., O.C., D.R., L.V., F.I., M.M., L.M., I.A., and E.B. developed experimental protocols and conducted the experiments. A.F. developed computational methods and performed bioinformatics analysis. Y.S. performed bioinformatic analyses. A.F., N.P.-Y., E.B., and I.G.-V. wrote the manuscript with input from all other authors.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41592-019-0355-5.

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

**Correspondence and requests for materials** should be addressed to E.B. or I.G.

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

## Methods

**The CPM algorithm.** CPM takes as input both bulk transcriptome and reference data. The input bulk transcriptome is represented in a column vector of expression values across all genes. Bulk expression values can be either the measured expression values in a single heterogeneous tissue or the relative values between two experiments, such as diseased versus healthy heterogeneous tissues. The input reference data consist of scRNA-seq profiles, represented in a matrix of the format $R_{ij}$, where the *ij*th entry is the expression level of gene *j* in single cell *i*. A row vector $R_i$ in this matrix is an RNA-seq signature of a certain reference single cell *i* (referred to as a 'reference profile'). We further assume that the main split of cells into broad cell-type categories is known, and that low-quality cells were already removed[16]. In addition, we assume that each cell type is associated with a pre-determined neighborhood structure, denoted the 'cell-state space' structure. The cell-state space is represented as a set of coordinates, where the *i*th entry represents the position of reference single cell *i* within the cell-state space. The cell space should be constructed in a pre-processing step through various single-cell analysis techniques. For instance, it is possible to exploit standard dimension reduction methods that provide cell positions in a low-dimensionality space (such as t-SNE or using several principal components[14]). Alternatively, it is possible to determine the position of cells along a certain trajectory of cell states and use this trajectory as a one-dimensional cell-state space[14].

CPM starts with a preprocessing step in which genes carrying many dropout events (here, fraction of zero expression values across single cells >90%) are filtered. Both the reference profiles and the input bulk profile are then standardized. Next, the algorithm proceeds in two steps: first, the composition of reference cells within the complex tissue is inferred, and then this information is used to predict cell abundance over the entire continuous cell-state space.

*Step 1 (deconvolution).* To infer the abundance level of the reference cells in a bulk expression profile, we solve the following linear regression: $U = \sum_i R_i \cdot \beta_i$, where $R_i$ is the expression vector of all genes in reference single cell *i*, *U* is the vector of expression levels of all genes in the complex tissue, and $\beta_i$ indicates the unknown abundance of single cell *i* in the complex tissue. As in Cibersort[18], we achieve robustness by solving this regression using linear SVR (the 'LiblineaR' v2.10-8 R package[19]) to prevent biases due to outliers. To further improve performance in the presence of a large number of reference profiles, we use a consensus approach in which the aforementioned SVR inference is repeated *N* times for *N* different subsets of the reference profiles (denoted 'reference subsets'), each reference subset consisting of *Ns* profiles. Predicted abundance values of the *N* runs are then averaged for each individual reference cell.

We apply several improvements that make step 1 more robust and accurate. First, we perform an unbiased random selection of the reference subset (without replacements) so that the selected subset is distributed uniformly over the cell-state space. To address this, each reference cell is sampled by random sampling of a 'pivot point' within the cell-state space using inverse transform sampling and then choosing an arbitrary reference cell in the proximity of this pivot point. A grid was added to the cell-state space so that all reference cells that fall in a certain entry of this grid are defined as the proximity group for a pivot that falls in this entry. The number of grid entries, calculated on each cell type separately, is the number of reference single cells divided by the cell neighborhood size. *N* was set to a value at which each reference cell would be selected to an average depth of at least *Nr* repeats. In particular, given that cells in high-density grid entries are less likely to be selected, we calculate *N* based on the highest density entry by requiring that each cell in this entry would be selected to an average depth of *Nr* repeats.

The second improvement is that each SVR run is applied to a set of genes that is tailored for a specific reference subset. The basic idea is to select a gene set that offers the best ability to distinguish between the reference profiles. Similarly to Cibersort[18], for each gene we compare its expression in different scRNA-seq profiles using one-way ANOVA (the *Nd* nearest neighbors of each cell are used to calculate the within-group variance component); each gene is then associated with the cell profile in which it attains the highest average expression, and the *Ng* top ANOVA-score genes associated with each cell profile are selected. In particular, *Ng* is defined as the number that minimizes the 'condition number' that is calculated with the R 'kappa' function.

*Step 2 (extrapolation).* To infer the abundance of a given candidate cell state, CPM averages the predicted abundances of its *Nd* nearest-neighbor reference cells. This leads to a smoothed cell abundance over the entire cell-state space. We refer to this solution as the 'cell population map'.

Overall, the methodology relies on three parameters: the number of deconvolution repeats (determined by *Nr*), the reference subset size (*Ns*) and the cell neighborhood size (*Nd*). Here we used $Nr = 5$, $Ns = 50$, and $Nd = 10$ as our default setting. The contribution of CPM is further discussed in Supplementary Note 1.

**The reference single-cell data.** The reference data are a collection of 1,860 single cells that were collected from the lungs of a C57BL/6 mouse 2 d after infection with $4.8 \times 10^3$ plaque-forming units (pfu) (in 40 µl of PBS) of the PR8 influenza virus (published data[16] from GEO accession number GSE107947). As reported previously[16], this collection already excludes poor-quality cells, and the cells were already partitioned into nine cell-type groups (in total, 92 B cells, 135 blood endothelial cells, 24 epithelial cells, 291 granulocytes (GN), 345 lymphatic endothelial cells, 375 fibroblasts, 103 mononuclear phagocyte system (MPS) cells, 117 NK cells, and 378 T cells). Furthermore, the progression of cell states through a trajectory of an antiviral-activation response was previously defined for each of the nine immune and non-immune cell types[16]. We refer to this continuum as the 'trajectory' of cell-activation states. In brief, the cell-state trajectory was constructed in two steps: first, a group of 101 generic-response genes were defined (consisting of all genes that were upregulated in all nine cell types during influenza infection); next, for each single cell, its average expression level across these generic genes was used as the activation-state trajectory[16] (in Figs. 3a,c and 4d,e and Supplementary Fig. 7c–g,i,j, this trajectory was further binned into equal intervals). Unless stated otherwise, this reference single-cell collection, together with its cell type groups and cell-state trajectory, were used as input in our analyses. We note that scRNA-seq data from a replicate infected mouse[16] was used to corroborate the results (Supplementary Fig. 7f).

**Synthetic data analysis.** Synthetic bulk profiles were generated by mimicking the heterogeneity of cells within a biological complex tissue. Each synthetic bulk profile was generated as a mixture of reference scRNA-seq profiles according to pre-designed fractions of single cells. Our pre-designed fractions of cells represent prevalent realistic scenarios, including changes in the overall level of a certain subpopulation (the cell-type and cell-subtype simulations), as well as changes along cellular trajectories such as cell-state shifts (the gradual-change simulation). We generated both absolute and relative synthetic bulk profiles and in both cases tested the entire range of the noise parameter (ranging from entirely non-informative data to almost-zero noise). The 'accuracy' was calculated as the Pearson correlation between the actual and predicted fractions of cells. Technical details of synthetic data generation and the accuracy score are available in Supplementary Note 1.

For each synthetic data collection, the performance of CPM was compared to alternative state-of-the-art deconvolution algorithms, including the digital cell quantifier (DCQ) algorithm[17], which builds on elastic net regression; Cibersort[18], which utilizes a non-iterative linear support vector regression; and a standard linear SVR. SVR was applied using L2-regularized L2-loss support (primal) vector regression because it provided similar accuracy compared to alternative settings but is faster than the alternatives. SVR was applied with the optimal setting of its C (the 'LiblineaR' R package[19]) and $\varepsilon = 0.001$ (all results were maintained with alternative ε values such as 0.1 and 0.001; Supplementary Fig. 2h). Using SVR and Cibersort, in the case of relative data we retained the negative coefficients, as previously suggested[17]. For the CPM algorithm, we further tested the effect of modifying the *Nr*, *Ns*, and *Nd* parameters. As the three compared deconvolution methods rely on a relatively small number of input reference profiles, the reference data were constructed by grouping of the scRNA-seq profiles. We used *K*-means clustering of the scRNA-seq data[16] and then used the averaged profile of each group as a reference profile. Supplementary Note 1 further describes alternative reference-construction methods whose accuracy levels are presented in Supplementary Fig. 4e. Each of the compared methods was analyzed using a variety of *K* (granularity) values. Finally, we compared CPM to an alternative approach in which cell composition is evaluated through enrichment of each individual reference profile (an 'enrichment scheme', as described previously[31]; detailed in Supplementary Note 1).

To verify that the mixture of single cells fully resembles real-data bulk profiles, we generated synthetic bulk expression values as a mixture of scRNA-seq of an uninfected C57BL/6J mouse (2,075 cells derived from the lung tissue, partitioned into nine cell types[16]), using quantities that were measured previously within the lungs of naive C57BL/6J mice (flow cytometry fractions from previous studies[22,32]). We further measured bulk lung profiles of a naive C57BL/6J mouse (Supplementary Table 1) and found a good match between measured and computationally synthesized bulk data (Supplementary Fig. 1a), supporting the validity of aggregating single cells into synthetic bulk profiles.

**Mice.** The present study used female mice aged 7–10 weeks from the Tel Aviv University (TAU) collection of Collaborative Cross recombinant inbred mice[20] and the C57BL/6J strain. The mice were raised at the Animal Facility at the Sackler Faculty of Medicine of TAU. All experimental mice and protocols were approved by the Institutional Animal Care and Use Committee (IACUC) of TAU (approval number 04-14-049), which adheres to Israeli guidelines and follows the US NIH animal care and use protocols. Mice were housed on hardwood chip bedding under a 12 h light/dark cycle at 21–23 °C. Mice were given tap water and standard rodent chow diet ad libitum from their weaning day until the end of the experiment.

**In vivo influenza virus infection.** Mouse-adapted PR8 strain, influenza virus A/Puerto Rico/8/34 (A/PR/8/34, H1N1), was persistently grown in hen egg amnion, and its effective titer was quantified. All mice were anesthetized with 7 mg ml$^{-1}$ ketamine and 1.4 mg ml$^{-1}$ xylazine at 0.1 ml per 10 g body weight, intraperitoneally. Animals were then infected intranasally with PR8 ($4.8 \times 10^3$ pfu in 40 µl PBS), whereas mock-treated ('control') animals received only 40 µl PBS. All mice were monitored daily for percentage body weight loss and clinical disease manifestations, and euthanized at 48 h post treatment. Of note, this experimental setting closely resembles that of the reference scRNA-seq data[16] (for example, the same gender, time after infection and virus strain, similar age and virus doses).

**RNA isolation, library construction and pre-processing.** To test CPM on complex tissues, murine lungs were collected immediately after mice were killed, sliced into small pieces, homogenized using a D1030-E BeadBlaster Microtube Homogenizer (Benchmark Scientific; 90 s, 4,000 r.p.m.) in the presence of QIAzol, and used for total RNA extraction using the miRNeasy Mini Kit (Qiagen). Library quality and concentration was measured using a TapeStation System (Agilent Technologies) and a Qubit Fluorometric Quantitation (Life Technologies), as described earlier[17]. mRNA sequencing libraries were constructed as described previously[17]. Absolute bulk profiles were generated through read alignment and transcript quantification, as described earlier[17] (detailed in Supplementary Note 1). Absolute profiles were transformed into 'relative' profiles using a common control profile as the normalizer, where the control profile was pooled from the PBS-treated mice. Relative profiles were calculated using log-transformed infected and control samples. Unless stated otherwise, results reported are the results of using relative profiles.

**Cell-to-phenotype correlations.** CPM was applied on each bulk RNA-seq lung sample by integrating the reference single cell collection (a total running time of ~16 min for deconvolution of 72 samples, using six cores of a Dell Latitude E6430 laptop, containing an Intel i7-3740QM CPU). Our analysis builds on the CPM-inferred abundance of each reference single cell in each individual mouse. For each reference cell (associated with a particular cell-activation state) we calculated the 'cell-to-phenotype correlation' as a Pearson correlation coefficient between the predicted abundance of cells at this cell state and the in vivo clinical phenotype at 2 d p.i. (illustrated in Supplementary Fig. 7b). The cell-to-phenotype correlation was calculated using two groups of mice: either the infected or the control (PBS-treated) mice. Cell-to-phenotype correlations were binned into nine equal intervals along the trajectory of cell activation states. Supplementary Note 1 describes several tests that were applied to support the inferred gradual changes over the cell-activation bins.

**Fluorescence-activated cell sorting and analysis.** To validate the performance of CPM, we sorted the population of macrophages from the lungs of various CC mice (Supplementary Table 1). To address this, the lungs were dissociated into single-cell suspensions using a Miltenyi Biotec lung dissociation kit (130-095-927) and gentleMACS dissociator (Miltenyi Biotec), according to the manufacturer's instructions. Isolated lung cells were then enriched for CD45$^+$ cells by positive selection (CD45 microbeads, Miltenyi Biotec, 130-052-301), incubated with blocking solution (5% normal mouse serum, 5% normal rat serum, and 1% anti-mouse CD16/CD32) for 30 min on ice, and stained with fluorochrome-conjugated antibody for CD11b (M1/70), CD64 (X54-5/7.1), I-A/I-E (M5/114.15.2), Ly6G (1A8), Ly6C (HK1.4), and CD45 (30-F11, Miltenyi Biotec). All antibodies were from Biolegend, unless otherwise mentioned (clone number in parentheses). Data were obtained using a SH800 flow cytometer (Sony Biotechnology) and analyzed with FlowJo v.10 software. Mononuclear phagocyte cells were gated as CD11b$^+$CD45$^+$Ly6G$^-$I-A/I-E$^+$, as described previously[22], and the expression levels of Ly6C and CD64 were analyzed.

**Inferring dynamics with a Markov model.** In this analysis, we rely on the assumption that cells along the activation trajectory are partitioned into $D$ equal intervals. The $i$th interval represents a discrete cell state $i$. In addition, we assume that the probability of transition between any two states (per unit time) is constant over time. The 'stochastic matrix' $Q_{DxD}$ encodes the probabilities of transitions $q_{ij}$ from state $i$ to state $j$ per unit time (referred to as 'transition rates'). Assuming that each cell in each state $i$ may remain in the same state or switch into state $i+1$ (but not to any other cell state), it follows that for each $i$, $q_{ii}+q_{i,i+1}=1$; and for each $j\notin\{i,i+1\}$, $q_{ij}=0$. We further define a row vector $F_{1xD}=(f_1,f_2,\ldots,f_D)$, where $f_i$ is the proportion of cells in state $i$ (denoted a 'state proportions vector'). We assume that each cell resides in exactly one of the cell states, and therefore $\sum_{i=1\ldots D}f_i=1$. The state proportions vector before infection and at any time $t$ after infection are $F^{(0)}$ and $F^{(t)}$, respectively. $Q^t$ encodes the probability of transitions after $t$ units of time, and therefore $F^{(t)}=F^{(0)}Q^t$. Using known state proportions vectors $F^{(0)}$ and $F^{(t)}$ (either CPM-inferred or FACS-measured; Supplementary Fig. 8e,f, respectively), we fit the missing transition rate parameters $\{q_{i,i+1}|i=1,\ldots,D-1\}$.

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

## Code availability

CPM is implemented in the 'scBio' CRAN R package (the CPM function), available at https://cran.r-project.org/web/packages/scBio/index.html.

## Data availability

All RNA-seq data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) database under accession numbers GSE113530 and GSE117975.

## References

31. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome. Biol.* **18**, 220 (2017).
32. Singer, B. D. et al. Flow-cytometric method for simultaneous analysis of mouse lung epithelial, endothelial, and hematopoietic lineage cells. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **310**, L796–L801 (2016).

nature research

Corresponding author(s): Irit Gat-Viks and Eran Bacharach

# 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.

## Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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

□ ☒ An indication of 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 statistics 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

□ ☒ Clearly defined error bars
*State explicitly what error bars represent (e.g. SD, SE, CI)*

*Our web collection on statistics for biologists may be useful.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | HOMMER, HISAT, FlowJo v10. |
| Data analysis | R v3.4, LiblineaR R package v2.10-8, limma R package v3.32.5, doSNOW R package v1.0.16, foreach R package v1.4.4, raster R package v2.6-7, fields R package v9.6, sp R package v1.2-6. |

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 upon request. 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

The bulk RNA-seq datasets has been deposited in the Gene Expression Omnibus (GEO) database, under GEO accession numbers GSE113530 and GSE117975.

# Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

# Life sciences study design

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

| | |
|---|---|
| Sample size | 80 female mice from 20 different lines of new developed Collaborative Cross and one C57BL/6J mouse were used for sequencing. Our analysis with 20 CC lines in Nachshon et al. 2016 (Frontiers in genetics, DOI: 10.3389/fgene.2016.00172) show that 20 lines should be sufficient to provide informative associations. We also used 13 mice from 6 CC lines for FACS validation; this number were chosen since CPM-inferred cell-to-phenotype correlations can be already observed with 13 mice. |
| Data exclusions | None |
| Replication | Not relevant, since our mice had a large variety of genetic backgrounds. However we did noticed that the gene expression profiles of mice from the same strain was usually more similar than the gene expression profiles of mice from different strains. |
| Randomization | Allocation of mice from each CC strain into the infected, PBS-control and untreated groups was random. |
| Blinding | Investigators were not blinded to group allocation. However, the main data contained a set of comprehensive gene expression profiles (with more than 20000 genes). We used the same method on each one of these full profiles and analyzed the results. We could not influence the results since initially we did not know the prediction for the samples. In the FACS analysis we mostly compared mice within each group and not between them. |

# Reporting for specific materials, systems and methods

## Materials & experimental systems

| n/a | Involved in the study | |
|---|---|---|
| ☒ | ☐ | Unique biological materials |
| ☐ | ☒ | Antibodies |
| ☒ | ☐ | Eukaryotic cell lines |
| ☒ | ☐ | Palaeontology |
| ☐ | ☒ | Animals and other organisms |
| ☒ | ☐ | Human research participants |

## Methods

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

## Antibodies

| | |
|---|---|
| Antibodies used | The following antibodies were used used for FACS:<br>AF-700-anti-mouse I-A/I-E (BLG-107621; Biolegend) 1:50; PE-anti-mouse Ly-6G (BLG-127607; Biolegend) 1:100; FITC-anti-mouse Ly-6C (BLG-128005; Biolegend) 1:100; PE/Cy7-anti-mouse/human CD11b (BLG-101215 Biolegend) 1:100; APC-anti-mouse CD64 (FcγRI) (BLG-139305; Biolegend) 1:50; VioBlueAnti-mouse CD45 (130-102-430; Miltenyi) 1:100; anti-mouse CD16/32 (14-0161 eBioscience) 1:100. |
| Validation | All the antibodies used for the FACS were similar to the antibodies used by Yu, Y.-R.A., et al. (PloS 2016). We also tested non specific binding/background by comparing each antibody to its isotype control and unstained cells. All antibodies were previously validated and shown efficiency in mice; the relevant information from the manufacturers' websites:<br><br>AF-700-anti-mouse I-A/I-E (BLG-107621; Biolegend):<br>https://www.biolegend.com/en-us/products/alexa-fluor-700-anti-mouse-i-a-i-e-antibody-3413<br><br>PE-anti-mouse Ly-6G (BLG-127607; Biolegend):<br>https://www.biolegend.com/en-us/products/pe-anti-mouse-ly-6g-antibody-4777 |

FITC-anti-mouse Ly-6C (BLG-128005; Biolegend):
https://www.biolegend.com/en-us/products/fitc-anti-mouse-ly-6c-antibody-4896

PE/Cy7-anti-mouse/human CD11b (BLG-101215 Biolegend):
https://www.biolegend.com/en-us/products/pe-cy7-anti-mouse-human-cd11b-antibody-1921

APC-anti-mouse CD64 (BLG-139305; Biolegend):
https://www.biolegend.com/en-us/products/apc-anti-mouse-cd64-fcgammari-antibody-7874

VioBlueAnti-mouse CD45 (130-102-430; Miltenyi):
https://www.miltenyibiotec.com/_Resources/Persistent/8fb245a80ee24b578ce3a5ccd6aaa4e97c130b83/DS_CD45-VioBlue%2Bmouse_30F11.pdf

anti-mouse CD16/32 (14-0161 eBioscience):
https://www.thermofisher.com/order/genome-database/generatePdf?productName=CD16/CD32&assayType=PRANT&detailed=true&productId=14-0161-81

## Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

All experiments were conduct on female mice of the new developed Collaborative Cross (CC). We used mice from 20 different CC lines (111A, 1488A, 1912A, 2126A, 21B, 2513A, 2750A, 3348A, 3912A, 4438A, 5000A, 5001A, 5003A, 5004A, 5010A, 5021A, 5022A, 5023A, 57B, 72A) and the C57BL/6J strain, all aged 7 to 10 weeks.

Wild animals

The study did not involved wild animals

Field-collected samples

The study did not involved samples collected from the field

## Flow Cytometry

### Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation

Mice were sacrificed using CO2 asphyxiation and the lung tissue was dissociated into single-cell suspensions using lung dissociation kit (130-095-927 MACS, Miltenyi Biotec). 2x10^6 cells were taken to measure the portion of CD45 positive cells in the single cell suspension .
CD45 positive cells were enriched using CD45 microbeads (130-052-301 MACS, Miltenyi Biotec) following Miltenyi protocol. Cells were stained for FACS analysis -We Used  2x10^6 cells per sample. Cells were incubated in blocking solution containing 5% normal mouse serum, 5% normal rat serum, and 1% FcBlock (eBiosciences, San Diego, CA) in PBS. Cells were then stained with antibodies on ice for 30 minute. After staining, cells were washed and fixed with 0.4% paraformaldehyde in PBS.

Instrument

SONY SH800

Software

Flowjo v10. Software

Cell population abundance

N.A.

Gating strategy

The examined cell were collected from Flu infected mice or from MOCK infected mice. Cells were gated based on their FSC/SSC (cells with FSC larger than 250K or SSC larger than 250K were expelled) and their viability (FSC-H/FSC-A). Gating boundaries were chosen based on control (unstained cells) and by comparing to cells extracted from control mice (mock infected with PBS)

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.