Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

Diego Adhemar Jaitin *et al.*

Science **343**, 776 (2014);
DOI: 10.1126/science.1247651

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

contractility of cardiomyocytes. These pathways in cardiac muscle cells are highly conserved across all vertebrates, which explains the common, canonical crude oil toxicity syndrome observed in a diversity of fish species from habitats that range from tropical freshwater (zebrafish) to boreal marine (herring).

In conclusion, the oil-induced disruption of cardiomyocyte repolarization via K^+ channel blockade and sarcolemmal and SR Ca^{2+} cycling should call attention to a previously underappreciated risk to wildlife and humans, particularly from exposure to cardioactive PAHs that are also relatively enriched in air pollution. I_{Kr} inhibition by DWH crude oil from the MC252 well was robust, and its properties are consistent with direct channel pore block. These K^+ channel targets and their unique gating properties play a critical role in cardiac action potential repolarization and are highly conserved across the animal kingdom (*20*). These results lead us to believe that PAH cardiotoxicity was potentially a common form of injury among a broad range of species during and after the DWH oil spill. The early life stages of fish and other vertebrates may have been particularly vulnerable, given that even a transient and sublethal effect of PAHs on the embryonic heartbeat can cause permanent secondary changes in heart shape and cardiac output (*32*). Moreover, the underlying ion channel currents that drive the electrical properties of cardiomyocytes in tunas and mammals (such as heart rates), are similar (*26*, *33*). Thus, we suggest the extension of our current oil toxicity results to mammalian, cardiomyocytes may be warranted to better understand PAH threats to human health.

References and Notes

1. M. G. Carls, J. P. Meador, *Hum. Ecol. Risk Assess. Int. J.* **15**, 1084–1098 (2009).
2. C. H. Peterson *et al.*, *Science* **302**, 2082–2086 (2003).
3. C. E. Boström *et al.*, *Environ. Health Perspect.* **110** (suppl. 3), 451–488 (2002).
4. J. J. Stegeman, J. J. Lech, *Environ. Health Perspect.* **90**, 101–109 (1991).
5. J. P. Incardona *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E51–E58 (2012).
6. M. G. Carls, S. D. Rice, J. E. Hose, *Environ. Toxicol. Chem.* **18**, 481 (1999).
7. R. A. Heintz, J. W. Short, S. D. Rice, *Environ. Toxicol. Chem.* **18**, 494 (1999).
8. J. P. Incardona, T. K. Collier, N. L. Scholz, *Toxicol. Appl. Pharmacol.* **196**, 191–205 (2004).
9. J. P. Incardona *et al.*, *Environ. Health Perspect.* **113**, 1755–1762 (2005).
10. J. P. Incardona *et al.*, *Environ. Sci. Technol.* **43**, 201–207 (2009).
11. J. H. Jung *et al.*, *Chemosphere* **91**, 1146–1155 (2013).
12. J. P. Incardona, T. K. Collier, N. L. Scholz, *J. Expo. Sci. Environ. Epidemiol.* **21**, 3–4 (2011).
13. J. D. Neilson, S. E. Campana, *Can. J. Fish. Aquat. Sci.* **65**, 1523–1527 (2008).
14. B. A. Block *et al.*, *Nature* **434**, 1121–1127 (2005).
15. J. R. Rooker *et al.*, *PLOS ONE* **7**, e34180 (2012).
16. Materials and methods are available as supplementary materials on *Science* Online.
17. A. R. Diercks *et al.*, *Geophys. Res. Lett.* **37**, L20602 (2010).
18. L. M. Hondeghem, L. Carlsson, G. Duker, *Circulation* **103**, 2004–2013 (2001).
19. J. M. Nerbonne, R. S. Kass, *Physiol. Rev.* **85**, 1205–1253 (2005).
20. P. Kannankeril, D. M. Roden, D. Darbar, *Pharmacol. Rev.* **62**, 760–781 (2010).
21. D. M. Bers, *Nature* **415**, 198–205 (2002).
22. M. B. Cannell, H. Cheng, W. J. Lederer, *Science* **268**, 1045–1049 (1995).
23. H. A. Shiels, E. V. Freund, A. P. Farrell, B. A. Block, *J. Exp. Biol.* **202**, 881–890 (1999).
24. H. A. Shiels, A. Di Maio, S. Thompson, B. A. Block, *Proc. Biol. Sci.* **278**, 18–27 (2011).
25. M. Vornanen, H. A. Shiels, A. P. Farrell, *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **132**, 827–846 (2002).
26. G. L. Galli, M. S. Lipnick, B. A. Block, *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **297**, R502–R509 (2009).
27. F. Brette, J. Leroy, J. Y. Le Guennec, L. Sallé, *Prog. Biophys. Mol. Biol.* **91**, 1–82 (2006).
28. F. Brette *et al.*, *Biochem. Biophys. Res. Commun.* **374**, 143–146 (2008).
29. F. Brette, L. Sallé, C. H. Orchard, *Biophys. J.* **90**, 381–389 (2006).
30. H. A. Shiels, E. White, *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **288**, R1756–R1766 (2005).
31. H. Tie *et al.*, *Br. J. Pharmacol.* **130**, 1967–1975 (2000).
32. C. E. Hicken *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7086–7090 (2011).
33. J. M. Blank *et al.*, *J. Exp. Biol.* **207**, 881–890 (2004).

Acknowledgments: We thank D. Bers and K. Ginsburg of University of California, Davis; R. Kochevar of Stanford University; and R. Ricker of NOAA for constructive advice on an earlier draft of the manuscript. We are also grateful to C. Farwell, A. Norton, and E. Estess of the Monterey Bay Aquarium, along with husbandry teams for maintenance of bluefin and yellowfin tunas and handling of oil experiments within Tuna Research and Conservation Center; the Government of Mexico for permitting the collection of tunas in their waters; T. Dunn and the crew of the F/V Shogun; and H. Shiels, G. Galli, and S. Thompson for experimental help during preliminary studies of PAH compounds. This work was funded as a contributing study to the DWH–MC252 incident Natural Resource Damage Assessment. Additional funds were provided by the Monterey Bay Aquarium Foundation and Stanford University. All experiments with crude oil were approved by Stanford University Institutional Animal Care and Use Committee procedures. All data necessary to understand this manuscript are presented in the main text or supplementary materials.

Supplementary Materials

www.sciencemag.org/content/343/6172/772/suppl/DC1
Materials and Methods
Figs. S1 to S14
Table S1
References (*34*–*41*)

3 July 2013; accepted 11 December 2013
10.1126/science.1242747

Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

Diego Adhemar Jaitin,[1]* Ephraim Kenigsberg,[2,3]* Hadas Keren-Shaul,[1]* Naama Elefant,[1] Franziska Paul,[1] Irina Zaretsky,[1] Alexander Mildner,[1] Nadav Cohen,[2,3] Steffen Jung,[1] Amos Tanay,[2,3]†‡ Ido Amit[1]†‡

In multicellular organisms, biological function emerges when heterogeneous cell types form complex organs. Nevertheless, dissection of tissues into mixtures of cellular subpopulations is currently challenging. We introduce an automated massively parallel single-cell RNA sequencing (RNA-seq) approach for analyzing in vivo transcriptional states in thousands of single cells. Combined with unsupervised classification algorithms, this facilitates ab initio cell-type characterization of splenic tissues. Modeling single-cell transcriptional states in dendritic cells and additional hematopoietic cell types uncovers rich cell-type heterogeneity and gene-modules activity in steady state and after pathogen activation. Cellular diversity is thereby approached through inference of variable and dynamic pathway activity rather than a fixed preprogrammed cell-type hierarchy. These data demonstrate single-cell RNA-seq as an effective tool for comprehensive cellular decomposition of complex tissues.

Understanding the heterogeneous and stochastic nature of multicellular tissues is currently approached through a priori defined cell types that are used to dissect cell populations along developmental and functional hierarchies (*1*–*3*). This methodology heavily relies on enumeration of cell types and their precise definition, which can be controversial (*4*–*7*) and is based in many cases on indirect association of function with cell-surface markers (*5*–*8*). Perhaps the best understood model for cellular differentiation and diversification is the hematopoietic system. The developmental tree branching from hematopoietic stem cells toward distinct immunological functions was carefully worked out through many years of study, and effective cell-surface markers are available to quantify and sort the major hematopoietic cell types. Even in this well-explored system, however, it is becoming increasingly difficult to explain modern genome-wide and in vivo data with refined cell types' hierarchy and functions that extend beyond the classical myeloid and lymphoid cell types. For example, dendritic cells (DCs) are antigen-presenting

[1]Department of Immunology, Weizmann Institute, Rehovot 76100, Israel. [2]Department of Computer Science and Applied Mathematics, Weizmann Institute, Rehovot 76100, Israel. [3]Department of Biological Regulation, Weizmann Institute, Rehovot 76100, Israel.

*These authors contributed equally to this work.
†These authors contributed equally to this work.
‡Corresponding author. E-mail: amos.tanay@weizmann.ac.il (AT); ido.amit@weizmann.ac.il (IA)

cells that were originally characterized through their morphology (*9*) but are now understood to represent a highly heterogeneous group (*10*) with multiple functions, regulatory circuits, and phenotypes (*6*, *7*, *9*). Despite considerable efforts and progress by use of the marker-based approach, much of the known functional heterogeneity within the DC group is not truly compatible with any of the DC subclassification schemes (*6*, *7*, *11*). Such lack of definitive models for cell types and states is common in many fields of biology.

An attractive alternative to marker-based cellular dissection of complex tissues is to characterize in vivo cell-type compositions through unsupervised sampling and modeling of transcriptional states in single cells. This natural approach was so far difficult to implement because of many technical limitations that are being progressively alleviated with the advent of single-cell RNA sequencing (RNA-seq) (*12*–*20*). Sampling and sequencing RNA from dozens of single cells was recently used to estimate stochastic transcriptional variation in stationary cultured cells (*14*) or during a dynamic process (*12*–*14*, *16*, *19*). An unsupervised framework for dissecting transcriptional heterogeneity within complex tissues may therefore be envisioned, provided that many thousands of cells can be assayed routinely by using single-cell RNA-seq and that data from such experiments can be normalized and modeled effectively even when cells represent highly diverse cell types and states.

We developed an automated massively parallel RNA single-cell sequencing framework

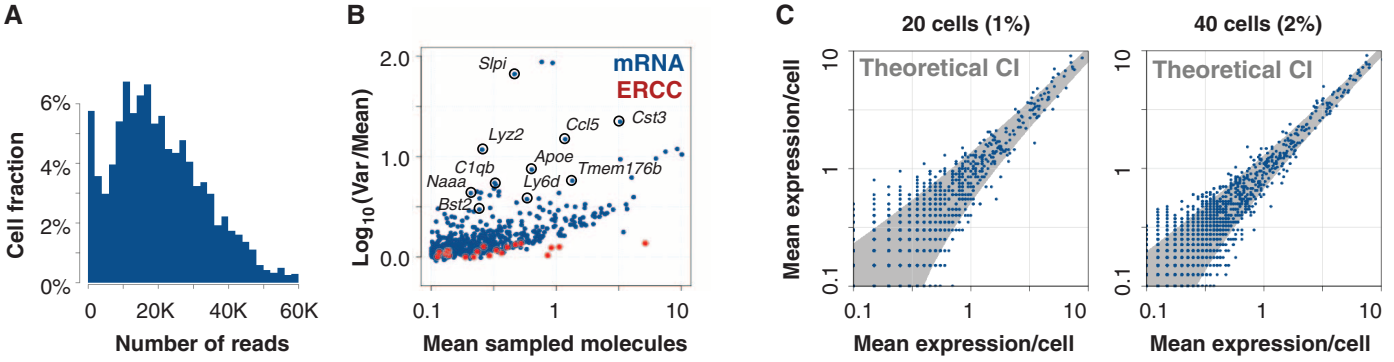
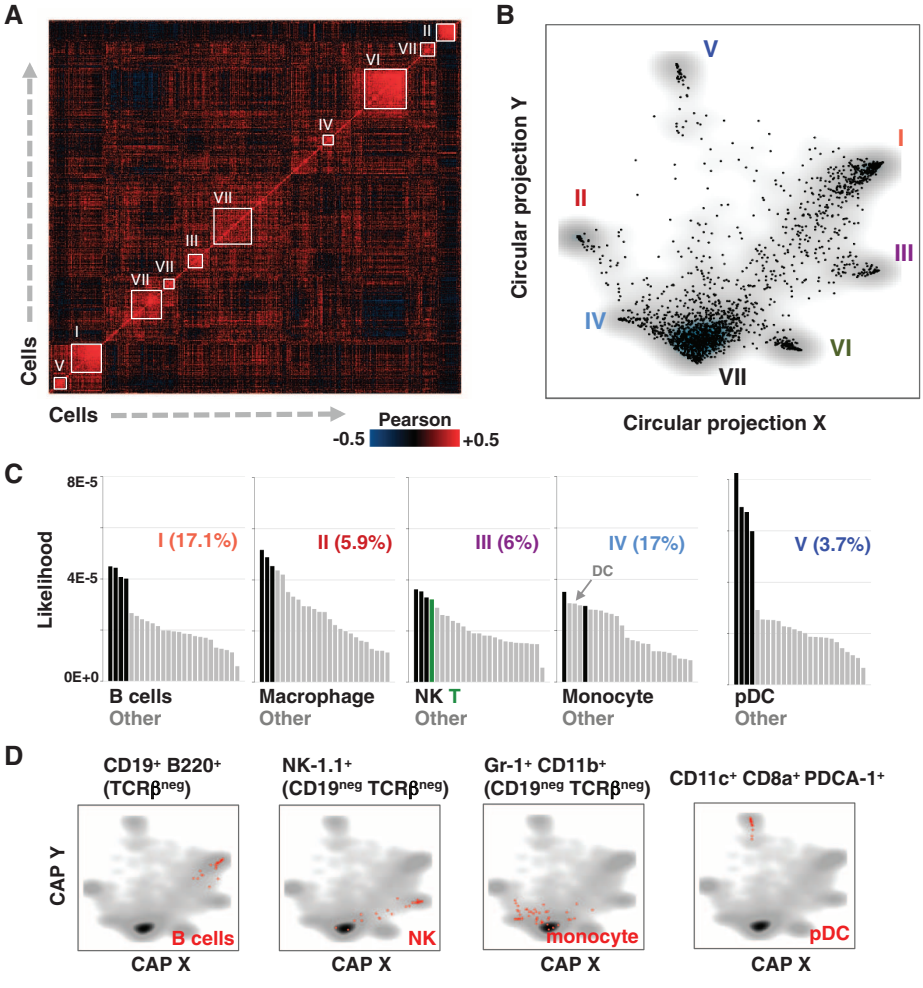


Fig. 1. Massively parallel single-cell RNA-seq. (**A**) Distribution of mapped reads per cell in a multiplexed 1536-cell experiment. (**B**) Mean and variance in mRNA (blue) and spike-in controls (red). (**C**) Mean mRNA counts in replicated pooled population of homogeneous (FACS-sorted) pDCs.

Fig. 2. Single-cell dissection of immune cell types. (**A**) Color-coded correlation matrix of single-cell mRNA profiles. Groups of strongly correlated cells that are used to initialize a probabilistic mixture model are numbered and marked with white frames. (**B**) Circular a posteriori projection (CAP) plot summarizing the predictions of the probabilistic mixture model for the CD11c^{+} cells. Each cell is projected onto the two-dimensional sphere according to the posterior probability of its association with the model's classes. The dimensions of the CAP plot should not be interpreted linearly or as principal components. (**C**) Bar plots depicting similarities of mean RNA counts in inferred types and Immgen expression profiles. The most correlated group of Immgen profiles is colored specifically as indicated for each type. (**D**) Shown are CAP plots depicting single-cell RNA-seq data sets acquired from marker-based FACS sorting for single pDCs, B cells, NK cells, and monocytes. Sorted cells are shown in red; density of the CD11c^{+} pool is shown in gray.



(MARS-Seq) (figs. S1 to S6) (*21*) that is designed for in vivo sampling of thousands of cells by multiplexing RNA-seq while maintaining tight control over amplification biases and labeling errors. The method is based on fluorescence-activated cell sorting (FACS) of single cells into 384-well plates and subsequent automated processing that is done mostly on pooled and labeled material, leading to a dramatic increase in throughput and reproducibility. To explore the new technique, we sequenced RNA from more than 4000 mouse spleen single cells (table S1), focusing initially on a heterogeneous cell population enriched for expression of the CD11c surface marker. We hypothesized that this strategy for cell acquisition will sample a diverse collection of splenic cell types while focusing on the challenging DC populations (*6*, *7*).

Our methodology uses three levels of barcoding (molecular-, cellular-, and plate-level tags) to facilitate molecule counting with a high degree of multiplexing. The strategy is to characterize cell subpopulations first by classifying single cells on the basis of low-depth RNA sampling and then studying transcriptional profiles at high resolution by integrating data from dozens to hundreds of cells within each unsupervised class. As shown in Fig. 1A, multiplexing 1536 cells in one sequencing lane provided an average of 22,000 aligned reads per cell, and after extensive normalization, these can be used to unambiguously define 200 to 1500 distinct RNA molecules from each cell. Our labeling and filtering scheme ensures that spiked-in technical controls show cell-to-cell variance that is compatible with the theoretical (binomial) sampling noise, comparing favorably with previously reported techniques (Fig. 1B) (*18*). This technical stability substantially increases the information content of the sampled transcriptional states, which can be directly modeled as unbiased samples of the cells' mRNA pool. In contrast to technical spike-in controls or the bulk of detected genes, we observed high cellular variance for a large number of genes, many of which are well known cell type–specific markers, suggesting that this attests for the high degree of heterogeneity within the splenic cell population (Fig. 1B) and promoting the idea of classifying cells into subpopulations on the basis of covariation of such heterogeneous markers.

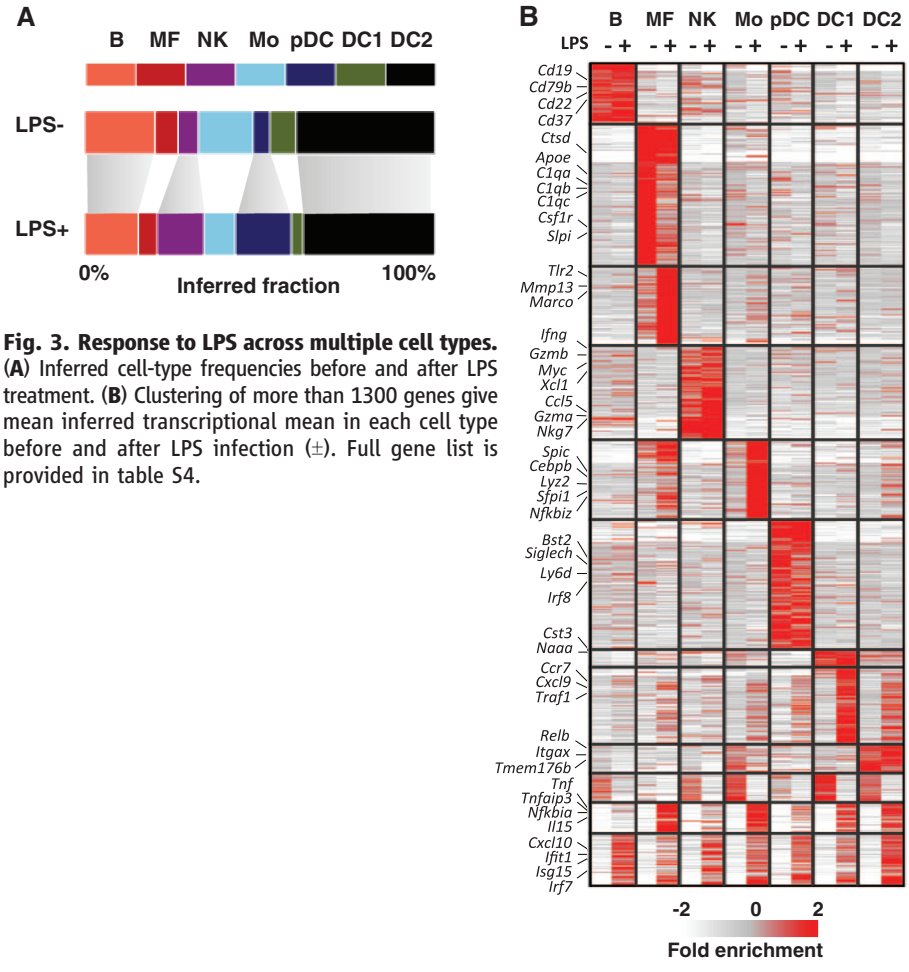


Fig. 3. Response to LPS across multiple cell types. (**A**) Inferred cell-type frequencies before and after LPS treatment. (**B**) Clustering of more than 1300 genes give mean inferred transcriptional mean in each cell type before and after LPS infection (±). Full gene list is provided in table S4.

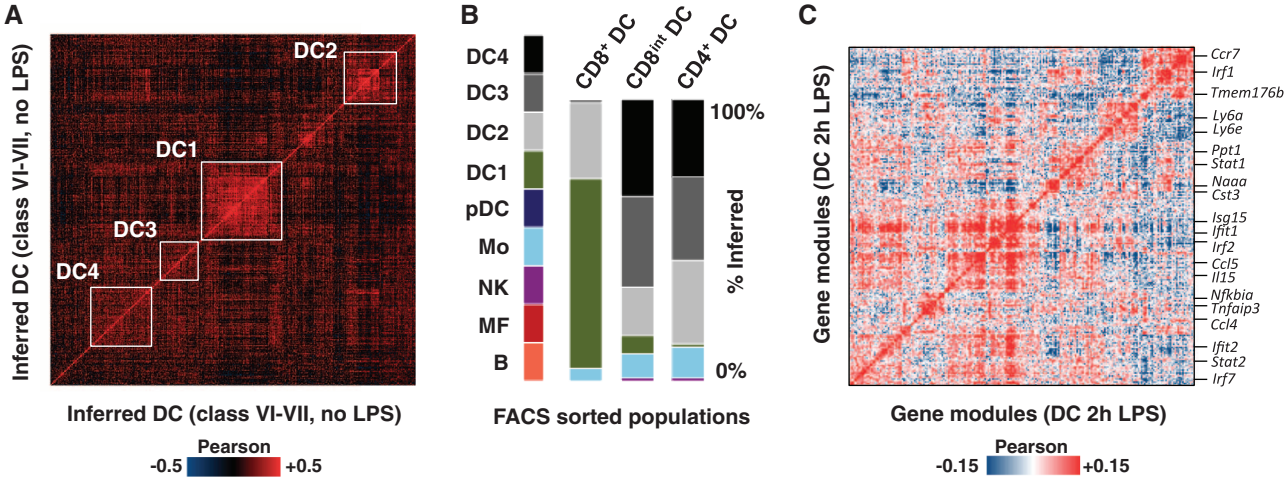


Fig. 4. Gene modules and the distribution and redistribution of DC states. (**A**) Single-cell correlation matrix for cells classified as DCs, showing detected subclasses using white frames. (**B**) Type/class distributions of single-cell RNA-seq data from three different FACS-sorted DC (CD11c enriched) populations: CD8a^{+} CD86^{+}; CD8a intermediate (int) CD86^{-}; and CD8a^{-} CD4^{+} ESAM^{+} (fig. S13A). (**C**) Gene correlation matrix is depicting potential LPS-dependent interactions between 225 genes. Key genes are indicated, with the complete list available in fig. S15.

To test how sensitive our strategy can be for characterizing the transcriptional state of subpopulations in the sample, we estimated coverage and mean mRNA molecule count reproducibility for groups of 10 to 40 single-cell profiles, representing 0.6 to 2.4% of the cells on one sequencing lane. Analysis of single-cell mRNA profiles from FACS-sorted plasmacytoid DCs (pDCs) (Fig. 1C and fig. S6) confirmed that pooling of homogeneous cell populations provides rich and highly reproducible transcriptional profiles. For a subpopulation at a frequency of 2.5%, the assay report on 1255 genes with a standard deviation of less than 35% of the mean, and on 324 genes with a standard deviation of 20% of the mean. Together, the availability of high-variance marker genes and the dynamic range provided by pooled single-cell transcriptional profiles enable unsupervised dissection and characterization of heterogeneous cell populations, opening the way for ab initio cell-type decomposition of splenic populations at a high level of detail.

We have implemented a probabilistic strategy for unsupervised classification of cells into "idealized types." Hierarchical clustering (Fig. 2A) defined seeds of highly correlated cells, leading to the initialization of a probabilistic mixture model and classification of single cells into types or families of homogeneous states. Visualization of the multiclass data by using a new circular a posteriori projection technique (Fig. 2B) represented the splenic cell population as a combination of several molecular behaviors, five of which (classes I to V) being distinctively separated from a group of more loosely defined classes (classes VI and VII). The frequencies of classes I to V range between 3.7 and 17%, allowing in all cases to infer rich transcriptional states through in silico pooling of single-cell mRNA profiles within each class. Analysis of gene enrichment (table S2 and figs. S7 and S8) and comparison of these profiles with existing transcriptional profiles of classical hematopoietic populations (www.immgen.org), unambiguously linked classes I to V to B cells, natural killer (NK) cells, macrophages, monocytes, and pDCs (Fig. 2C). The remaining classes were all linked to DCs. FACS analysis using classical surface markers confirmed our in silico estimations of the frequency of B cells and pDCs within the CD11c-enriched splenic cell population (fig. S9). Further analysis and additional single-cell quantitative polymerase chain reaction experiments confirmed that "marker" genes are robustly enriched in their relevant subpopulations (figs. S10 and S11). Using classical marker-based sorting, we further validated our approach with additional single-cell RNA-seq data from FACS-sorted B cells, NK cells, pDCs, and monocytes. Projection of the new data onto the model we generated from the splenic population showed remarkable compatibility between the traditional marker-based cell-type definition and the marker-free single-cell RNA-seq technique (Fig. 2D). Analysis of splenic cell populations therefore showcased single-cell RNA-seq as a direct and unsupervised way for identifying and characterizing subpopulations within heterogeneous tissues.

We profiled additional 1536 single cells from spleens that were exposed to lipopolysaccharide (LPS) for 2 hours (*22*), aiming to test how an immediate response to an infection-mimicking stimulus can be deciphered across the heterogeneous splenic cell population. We found that the LPS-treated cells are broadly classified into similar cell types to those observed in untreated cells, with some changes in the relative representation of different types (Fig. 3A). Using the non-LPS mixture model, we classified the non-LPS and LPS-exposed cells into classes and inferred a rich transcriptional profile within each class before and after treatment. Clustering 1575 variable genes identified groups of cell type–specific response genes (such as *Tnf* and *Marco* in macrophages and *Xcl1* and *Gzmb* in NK cells), and a large group of type I interferon response genes (*Irf7*, *Stat2*, *Ifit1*, *Cxcl10*, and hundreds more) activated pervasively in all or almost all cell types (Fig. 3B, fig. S12, and tables S3 and S4).

With thousands of samples readily available, single-cell RNA-seq is poised to go beyond the classical cell-types hierarchies that are outlined by current marker-based approaches, examining complex relations between cell subpopulations or continuous spectra of types. Analysis of 1031 single cells that were associated with DC-related classes (VI and VII) in our unsupervised $CD11c^+$ model (Fig. 4A) indicated that although 15% of these cells (class DC1) are strongly linked together, the remaining bulk of DCs could not be organized along a clear clustering hierarchy (*11*). Nevertheless, we found strong support for substantial internal organization within the remaining DC population (DC2 to DC4) (table S5), including a group of cells coexpressing *Relb*, *Nfkbia*, and additional associated genes (DC2) (fig. S13). More generally, we have identified several gene modules that represent combinatorial pathway activity within the DC bulk (fig. S14), indicating that despite the lack of a clear hierarchy, the DC cell population is governed by a high degree of transcriptional organization. Additional single-cell sequencing of $CD8^+$ $CD86^+$, $CD8^{int}$ $CD86^-$, and $CD4^+$ FACS-sorted populations (Fig. 4B) showed that this organization can be approached to a limited extent with existing marker-based classification. Remarkably, exposure to LPS reorganizes the DC population substantially, with a large number of gene modules being activated in a highly heterogeneous fashion (Fig. 4C and fig. S15). According to our analysis, certain specific $CD4^+$ DC subpopulations are activating the Irf4, tumor necrosis factor, and transforming growth factor–β pathways (fig. S16 and table S6), whereas other pathways (such as Irf7) are activated pervasively (table S5). This combinatorial activity of pathways within the LPS-exposed DC pool is not represented in preexisting DC subtypes according to our data. Committed and developmentally stable myeloid and lymphoid cell types maintain their identity during immediate response to infection while responding through generic and cell type–specific pathways. These pathways create substantial cell-to-cell variance and define new cell subpopulations within each of these cell types (fig. S17), forming diversity that may have functional implications. Observation of transcriptional subpopulations, however, does not necessarily imply the existence of further committed and preprogrammed cell subtype hierarchy.

We presented this framework for broad sampling of single-cell transcriptional states from tissues and demonstrated how it can be used to dissect complex functions in a bottom-up fashion. MARS-seq can be readily applied to tissues and organs in normal and disease states to redefine their cell-type and cell-state compositions and link it to detailed genome-wide transcriptional profiling. Given the inherent stochasticity and heterogeneity of multicellular tissues, this approach can prove essential for understanding how in vivo biological function emerges from complex cell ensembles.

References and Notes

1. M. Acar, J. T. Mettetal, A. van Oudenaarden, *Nat. Genet.* **40**, 471–475 (2008).
2. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, *Science* **297**, 1183–1186 (2002).
3. R. N. Germain, *Nat. Immunol.* **13**, 902–906 (2012).
4. S. C. Bendall *et al*., *Science* **332**, 687–696 (2011).
5. B. M. Bradford, D. P. Sester, D. A. Hume, N. A. Mabbott, *Immunobiology* **216**, 1228–1237 (2011).
6. F. Geissmann, S. Gordon, D. A. Hume, A. M. Mowat, G. J. Randolph, *Nat. Rev. Immunol.* **10**, 453–460 (2010).
7. D. A. Hume, *J. Leukoc. Biol.* **89**, 525–538 (2011).
8. M. C. Nussenzweig, R. M. Steinman, M. D. Witmer, B. Gutchinov, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 161–165 (1982).
9. R. M. Steinman, Z. A. Cohn, *J. Exp. Med.* **137**, 1142–1162 (1973).
10. L. Bar-On *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14745–14750 (2010).
11. D. Hashimoto, J. Miller, M. Merad, *Immunity* **35**, 323–335 (2011).
12. T. Hashimshony, F. Wagner, N. Sher, I. Yanai, *Cell Rep.* **2**, 666–673 (2012).
13. S. Islam *et al*., *Nat. Protoc.* **7**, 813–828 (2012).
14. D. Ramsköld *et al*., *Nat. Biotechnol.* **30**, 777–782 (2012).
15. Y. Sasagawa *et al*., *Genome Biol.* **14**, R31 (2013).
16. A. K. Shalek *et al*., *Nature* **498**, 236–240 (2013).
17. F. Tang, K. Lao, M. A. Surani, *Nat. Methods* **8**, (Suppl), S6–S11 (2011).
18. A. R. Wu *et al*., *Nat. Methods* **11**, 41–46 (2014).
19. Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, *Science* **343**, 193–196 (2014).
20. S. Islam *et al*., *Nat. Methods*, published online 22 December 2013 (10.1038/nmeth.2772).
21. Materials and methods are available as supplementary materials on *Science* Online.
22. I. Amit *et al*., *Science* **326**, 257–263 (2009).

Acknowledgments: Research was supported by the European Research Council and Israel Science Foundation (1782/11, 1050/12) grants to I.A. and A.T. RNA-seq data are deposited in Gene Expression Omnibus, accession no. GSE54006.

Supplementary Materials

www.sciencemag.org/content/343/6172/776/suppl/DC1
Materials and Methods
Figs. S1 to S17
Tables S7 to S9
References (*23*–*25*)
Tables S1 to S6

25 October 2013; accepted 22 January 2014
10.1126/science.1247651