An Atlas of Cellular Archetypes

George Crowley[1], Uri Alon[2], Stephen R. Quake[1,3,4 *]

[1] Department of Bioengineering, Stanford University, Stanford, California, USA
[2] Department of Molecular Cell Biology, Faculty of Biology, Weizmann Institute of Science
[3] Department of Applied Physics, Stanford University
[4] Chan Zuckerberg Initiative
* Corresponding author. Email: steve@quake-lab.org

Abstract

Single cell RNA sequencing allows effective characterization of cell types and cell states. Cell types can be reliably identified as defined clusters via marker gene expression and differential expression analysis. However, once a cell type is identified, current analysis methods provide little information about the potential variety of functions within the cell type. Here, we apply a principled computational approach to understand the functional heterogeneity within a given cell type. It has been hypothesized that cells in competitive environments will conform to Pareto task optimality theory by organizing in low-dimensional polytopes in gene expression space, and that specialist phenotypes identified via Pareto Task Inference Analysis would be reproducible and general. We used the Tabula Sapiens atlas of single-cell RNA sequencing across cell types and tissues in the human body to test this hypothesis and discovered that most cell types are well-fit by this theory. Furthermore, we derived specialist phenotypes from this method, which were used to identify archetypes of cellular function that represent essential biological tasks. We compared the specialist phenotypes from common cell types shared across tissues and found consistent results. These phenotypes are derived from an unbiased approach and do not incorporate ideas from existing biological models or theories, and yet in many cases they recapitulate our understanding of the functions of major cell types. Taken together, these results suggest a principled approach to annotating the continuum of functions within each cell type.

Main Text

Advances in single-cell RNA sequencing have made it possible to perform unbiased, high-throughput analysis of single-cell transcriptomes from entire organisms. Recently, the Tabula Sapiens Atlas has sequenced close to 500,000 cells across 24 organs of the human body, with tissue replicates across donors. Similar projects have produced parallel databases for a variety of model organisms(*1-4*), and substantial efforts have been made to compile these datasets. In principle, these large atlases allow us to understand cellular heterogeneity across the entire human body. However, a key gap is the lack of methods to summarize and interpret the biology of these cell types, specifically their function in the body. The standard workflow of single cell transcriptomics (*5-7*)—clustering cells to identify and report cell type marker genes—lacks biological functional insight, and has been described as akin to "reading a phonebook".(*8*) More broadly, the functional and molecular ontologies currently used in cell biology are subjective, incomplete, and can falsely discretize continuous heterogeneity.(*9*) Here, we demonstrate a principled method to identify biological function in continuously varying expression data. We show that this method can be broadly applied across cell types to find biologically distinct functions, and present an atlas of human *cell archetypes*—a molecular characterization of cell functions in healthy human tissue.

Generally, cells often perform several tasks and may trade-off performance at one task for performance at another. In the context of such tradeoffs, optimal phenotypes exist on a Pareto front in performance space, and occupy a polytope in trait space. In cell biology, these tasks are not always known a priori, but even without such prior knowledge one can discover the tasks. To do so, one plots the phenotypes in trait space to identify specialist phenotypes (also called archetypes) that, due to convexity, lie at the vertices of polytopes: the ends of a line segment, or the vertices of a triangle or tetrahedron. One can then systematically determine the function of these specialist archetypes by looking at differentially expressed genes in the cells closest to the vertices. This form of analysis is called Pareto Task Inference (ParTI) analysis and was introduced by Shoval et al. in 2012 to derive functional insight from a wide variety of phenotypic traits. Here we apply the approach within individual cell types in order to understand the range of tasks that the cell types perform.

We used Tabula Sapiens to test two fundamental questions: 1) Do single-cell transcriptomes of each cell type adhere to low-dimensional polytopes?; and 2) Do the vertices of these polytopes correspond to cells specializing in specific biological functions? The Tabula Sapiens Atlas is a single-cell RNA sequencing dataset of 456,101 high-quality single cell transcriptomes, covering 58,870 genes across 177 cell types, 25 tissues, and 15 donors.(*3*) We applied quality control filters to remove outlier cells on several metrics, yielding 309,193 cells across 176 cell types, 24 tissues, and 14 donors, **Table S1** and **Figure S1**. Cell type abundance filters left 113 cell types across the same number of tissues and donors, yielding 440 distinct cell type-tissue-donor strata for ParTI analysis.(*10, 11*)

In the following analysis, we fit polytopes to the data. A polytope is defined by the location of its vertices. We interpret each vertex as representing a specialist phenotype that performs

some specialized biological task(s). In this manuscript, specialist phenotypes that are reproducible across biological replicates are considered to be functional archetypes.

Most cell types are described by polytopes.

We performed dimensional reduction and model fitting independently for each distinct cell type-tissue-donor dataset, **Figure 1A**. After removing mitochondrial genes and genes known to be affected by tissue processing, we normalized the remaining genes to 10,000 counts (so that cell size effects would not dominate our results), and used principal component analysis (PCA) of protein-coding genes to represent cells in PCA space.(*3*) Pareto optimality, coupled with physical and biochemical constraints, results in reproducible geometry in trait space, **Figure 2.**

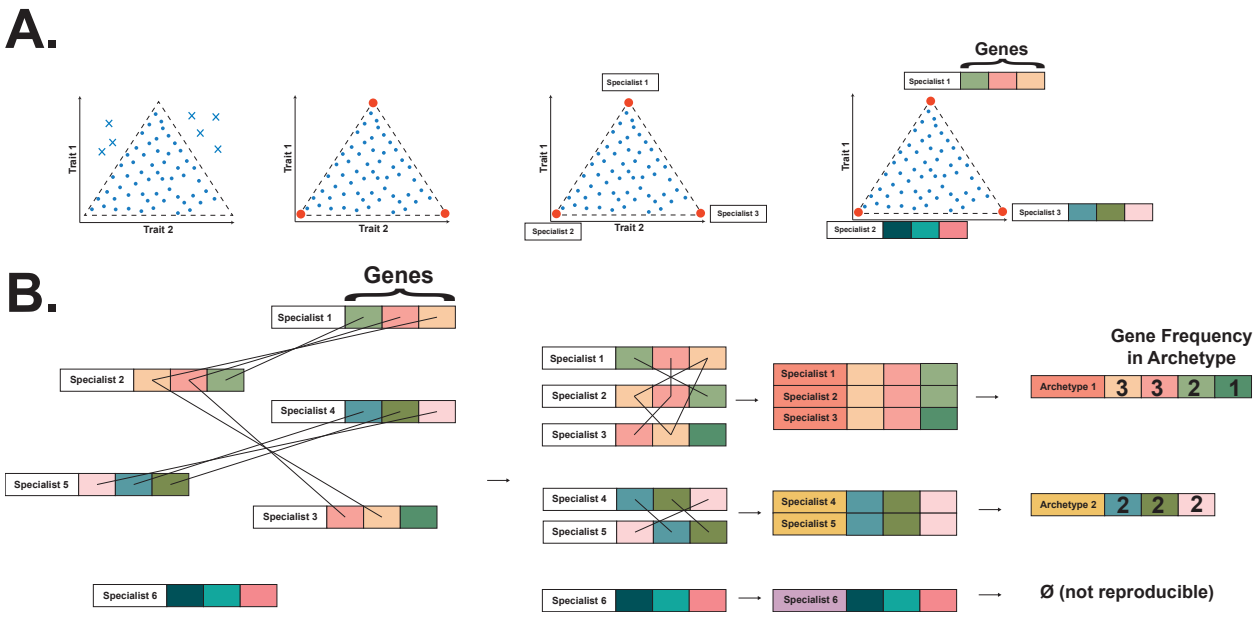

Figure 1. Process Overview. A. Pareto Optimality Theory. Pareto optimality places limits on allowable phenotypes. In trait space, this constrains phenotypes to be bound by a polytope. Specialist phenotypes lie at the vertices of the polytope and are enriched in traits relevant to their specialist functions. **B. Archetype Alignment.** Specialist phenotypes are aligned on a per-cell-type basis, yielding functional archetypes that are reproducible across donors and tissues.

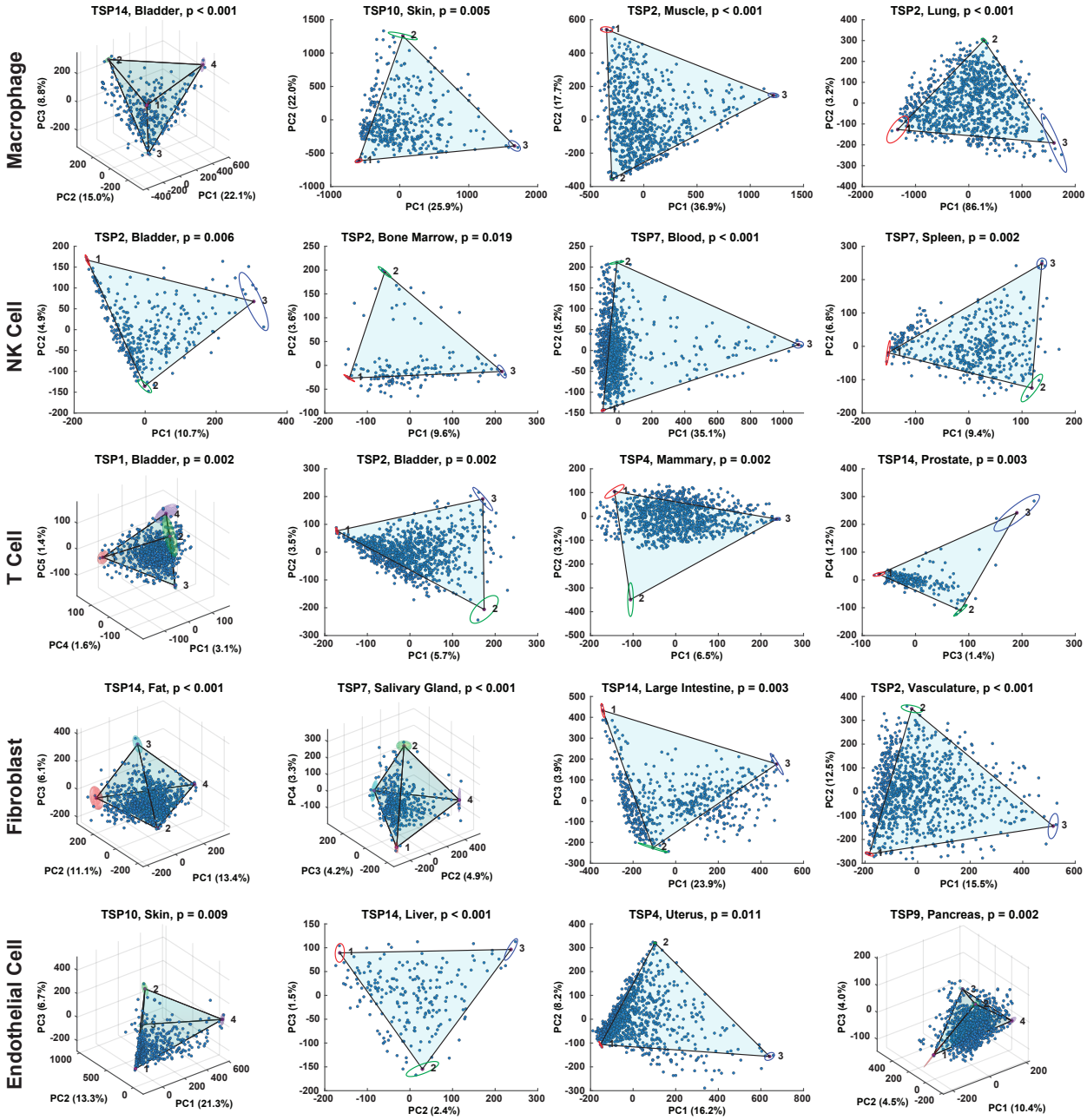

Figure 2. Reproducible Geometry of Donor-Tissue-Cell Types and Fits. Principal components analysis and Pareto Task Inference analysis performed independently on macrophages, NK cells, T cells, fibroblasts, and endothelial cells across 4 unique donor-tissues each demonstrates this geometry. Confidence bounds on vertex position are shown by ellipses and ellipsoids.

Using the ParTI package, we fit a polytope to each donor-tissue-cell type in PCA space, **Figure 2**.(*11, 12*). For this study, we set α=0.05 and maximum false discovery rate (FDR) =0.10. Three-quarters of cell types (84/113, 74.3%) were significantly fitted by polytopes (defined by ≥ 50% of donors having significant tissues), **Figure 3A, B, C**. The associated false discovery rate is 0.0843, **Figure 3D**. Independently, a majority (247/440) of donor-tissue-cell type strata were

significantly fitted by polytopes. Furthermore, when removing singlet and doublet cell types (i.e. cell types with only one or two donor-tissues), 90% of cell types (37/41) were significantly fitted. We observed no substantial relationships between quality control metrics and significance of fits, **Figure S2.**

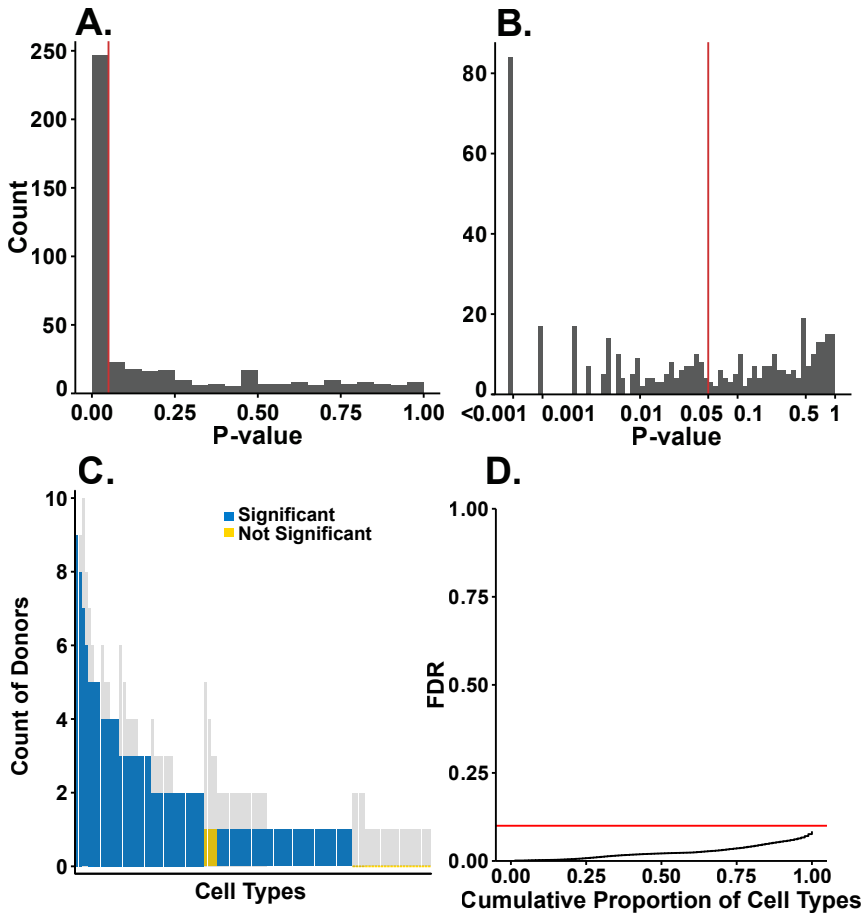

Figure 3. Significance of fits of each donor-tissue-cell type in Tabula Sapiens (both **A.** standard and **B.** log-scaled axes shown). **C**. Counts of donors with significant tissue(s) present in each cell type, plotted against a background of all donors available for that cell type. Cell types with significance in at least 50% of available donors are considered significant. **D.** Cumulative distribution function of false discovery rate (FDR) as a function of proportion of significant cell types retained. Three quarters of available cell types are significantly fit at α=0.05 and maximum FDR=0.10. Vertical red lines are drawn at p=0.05 (A. and B.) and FDR=0.10 (C.).

A detailed analysis of the frequency of significance of cell types shows that each donor and tissue has significantly fitted cell types (except for TSP3—only a single donor-tissue-cell type available—and kidney—only 2 donor-tissue-cell types available), **Figure S3**. The tissues with the highest number of donor-cell type pairs available were blood, spleen, and lung, **Figure S4**. The donors with the highest number of tissue-cell type pairs available were TSP14, TSP2, and TSP7, **Figure S5**. Finally, the cell types with the highest number of donor-tissue pairs available were macrophages, natural killer (NK) cells, T cells, fibroblasts, and endothelial cells, **Figures S4, S5**. We therefore focused on these cell types across tissues to understand their properties.

We considered vertices of each significant polytope to be specialist phenotypes, representing cells in each donor-tissue-cell type that perform a specialized set of tasks, **Figure 1B**. Genes that were significantly differentially expressed in the cells closest to each vertex were used to infer the archetype represented by the vertex's putative biological function.(*11*) Restricting our functional inference analysis to the 84 significant cell types (defined above) left 854 vertices with significantly enriched genes across 243 significantly fitted donor-tissue-cell types. To define the expressional program of each specialist phenotype, we consider up to the top 10 significantly enriched genes, sorted by expression (some specialist phenotypes have fewer than 10 significant genes). When searching for enriched genes, we used the set of all protein-coding genes, including mitochondrial and artefact genes, to allow the detection of artefactual specialist phenotypes.

For each cell type, we aligned specialist phenotypes based on their shared genes, **Figure 1B**. When a specialist phenotype was reproducibly found in at least two-thirds of donors, we considered the specialist phenotype to represent a functional archetype that performed an essential, characteristic function of the given cell type. Note that this donor-level filter limits the following analysis to the 37 significant, non-singlet, non-doublet cell types.

For functional inference, we considered the frequency-ranked union of genes among an aligned set of specialist phenotypes, and ignored genes that were only present in a single specialist phenotype. To generate functional insight for each archetype, we used a large language model to annotate all gene lists. Similar LLM approaches have been used recently and showed satisfactory agreement with gene ontology.(*13*) We used Gene Ontology-Biological Process, GPT-4, GPT-4o, and Claude 3 Opus; Claude 3 Opus generally gave the most interpretable annotations.

Macrophage archetypes reflect their immune roles.

Macrophages are multitasking cells found in all tissues. Current approaches subdivide macrophages into discrete categories such as M1 and M2 states, although data shows that macrophage expression is continuous in gene expression space. We find that macrophages are well described by triangles (three archetypes) or tetrahedra (four archetypes) across tissues.

The first macrophage archetype—**antigen presentation and complement activation**—describes macrophages' role in antigen presentation via the Major Histocompatibility Complex (MHC) class II pathway (with enrichment in *CD74*, *HLA-DRA*, *HLA-DPB1*, *HLA-DPA1*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DRB5*), **Figure 4A**. This archetype also shows macrophage roles in activation and amplification of the complement system (with enrichment in *C1QA*, *C1QB*, *C1QC*, *C3*).

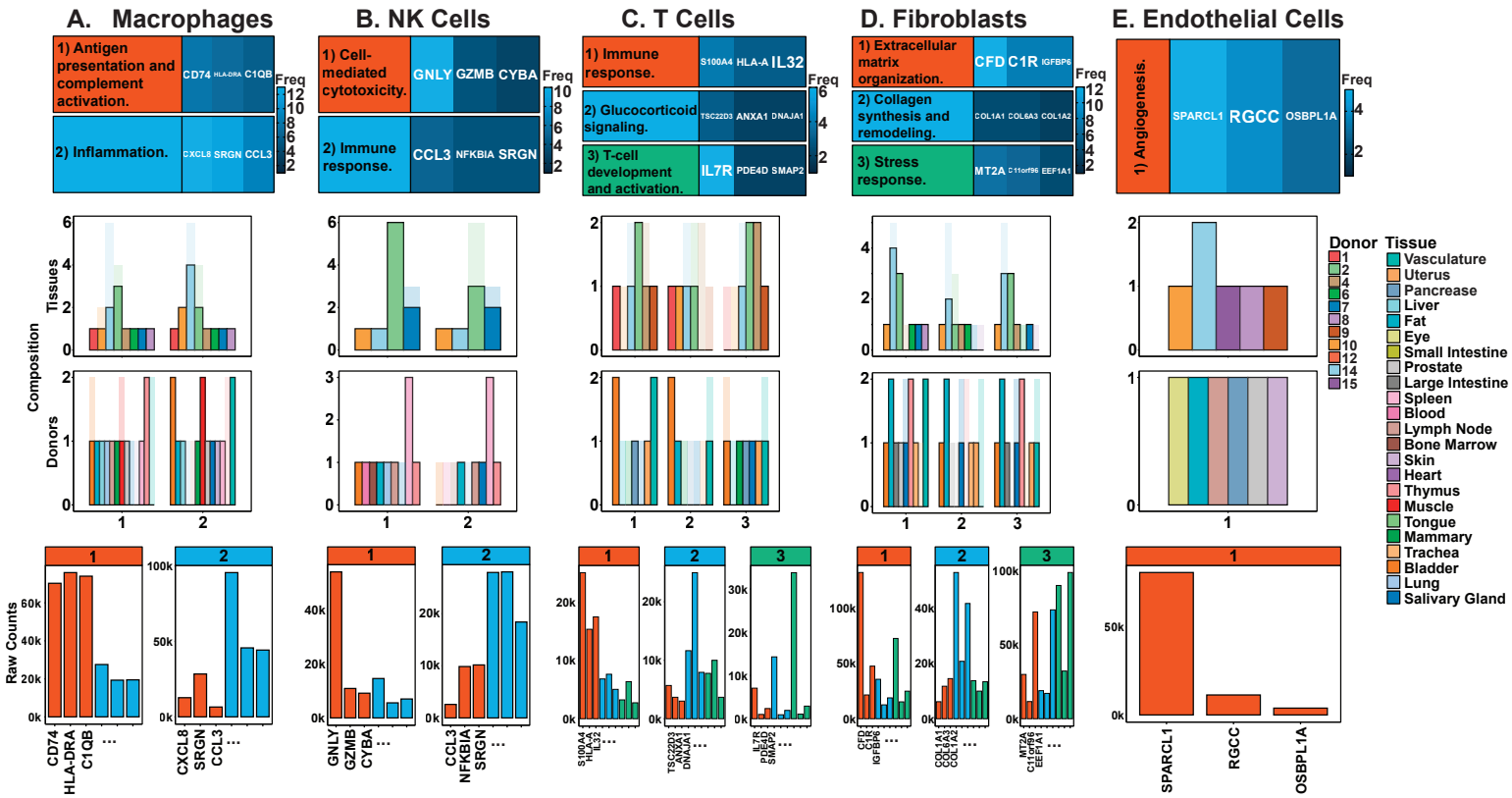

Figure 4. Archetypes of Ubiquitous Cell Types. For 5 ubiquitous cell types (**A. Macrophages**; **B. NK Cells**; **C. T Cells**; **D. Fibroblasts**; and **E. Endothelial Cells**), we show: 1st row: Top 3 most frequent archetype-defining genes in archetypes containing two-thirds or more of available donors, annotated via a commercial large language model; 2nd row: Counts of tissues present in each donor for each archetype, plotted against a background of all tissues available for that donor-cell type; 3rd row: Counts of donors present in each tissue for each archetype, plotted against a background of all donors available for that tissue-cell type; 5th row: Gene expression (raw counts) of top 3 genes of each archetype, visualized across all archetypes, using the 5% of cells closest to each vertex on a donor-tissue-cell type basis, with bars labeled at the top by the number of cells used to represent each archetype.

Antigen presentation and complement activation are known to be expressed in macrophages across tissues and are key to core macrophage function and immunity.(*14*) Our analysis indicates that these two functions may be performed by transcriptionally similar macrophage specialist cells. This sheds light on an open question, which is to elucidate the transcriptional programs active in these pan-tissue functions, and assess the degree to which they are shared or unique.(*14*)

While calling for more comprehensive analysis, current macrophage ontogeny has divided tissue-nonspecific macrophages into two distinct subpopulations, LYVE1^{low}MHCII^{high}, and LYVE1^{high}MHCII^{low}.(*15*) The above pan-tissue macrophage archetype is representative of MHCII^{high} macrophages. That said, a variety of expressional programs could underly a two-marker dichotomy. We argue that this two-marker dichotomy may highlight one functionally distinct, “specialist macrophage” archetype, and may have aggregated “generalist”

macrophages with functionally distinct "specialists" of another extreme. Thus, our analyses, which uses global transcriptional patterns, identified a second macrophage specialist phenotype.

The second macrophage archetype—**inflammation**—reflects macrophages role in the inflammatory response and was enriched in cytokine and chemokine production (*CXCL8*, *CCL3*, *IL1B*, *CCL4*, *CCL20*, *CCL3L1*, *CCL4L2*) and regulation (*SRGN*), as well as cell migration in response to inflammation (GPR183), **Figure 4A**.

It has been argued that the M1/M2 macrophage taxonomy should be abandoned in favor of a function-based nomenclature(*16*). Here, we applied a principled method to extract these functions and define them based on molecular expression. This analysis provides a detailed description of gene co-regulation underlying specific functional programs.

Natural killer cell archetypes represented motifs of cytotoxicity and immune cell recruitment.

NK cells are also found in all tissues and carry out cytotoxic functions. We found that NK cells are well described by a continuum of gene expression. The analysis reveals two NK cell archetypes that occur in all tissues. The first archetype—**cell-mediated cytotoxicity**—expresses markers of cytotoxic potential such as perforin and granzyme (*GNLY*, *GZMB*, *PRF1*, *CTSW*) and proliferation (*CENPF*, *STMN1*, *TUBB*, *TUBA1B*), alongside markers of active NK cells (*FCER1G*, *S1PR5*, *SELL*), **Figure 4B**. The second NK cell archetype—**immune response**—also expressed markers of active NK cells (*CD69*, *GZMK*), but had an expressional profile characterized by cytokine production (*CCL3*, *CCL4*), migration (*CXCR4*), and inflammatory regulation (*FOS*, *NFKBIA*, *SRGN*, *DUSP2*, *RGS1*, *SLA*, *TLE1*), **Figure 4B**. Thus, while both of these archetypes might be considered active NK cells, the first NK cell archetype likely focuses on direct cytotoxicity and proliferation, while the second NK cell archetype modulates the immune signaling environment.

Only recently have attempts been made to characterize NK cell populations at the transcriptomic level. A comprehensive characterization of NK cells based on CITEseq and snRNAseq identified several clusters of NK cells (NK1, NK2, NK3) and aimed to standardize NK cell ontology, which is an important and valuable effort.(*17*) However, this effort still relied on arbitrary delineation of clusters and subclusters within a transcriptional landscape that looked continuous in both state and trajectory. The NK cell archetypes (measured across a broader set of tissues) recapitulate these subsets.

The commonalities among the NK cell archetypes and NK1, NK2, and NK3 are type-specific granzyme production. NK1 matches archetype 1 on the genes (*GZMB*, *PRF1*, *FCER1G*, *S1PR5*), and NK2 has overlap with archetype 2 (*CD69*, *FOS*, *GZMK*, *CXCR4*, *DUSP2*). However, some archetype marker genes were expressed among other NK populations. *SELL* is a marker in our archetype 1 but their NK2, and *CCL3* and *CCL4* are markers for their NK1 and our archetype 2.(*17*)

Given that this same comprehensive characterization of NK cells also established continuous transcriptional trajectories between their subtypes, we argue that an archetypal description of NK cell heterogeneity is more natural because it allows the description of intermediate or combination states without forcing arbitrary subdivisions that may not represent the underlying transcriptional/phenotypical landscape. This may also help resolve the nonspecific elevation of marker genes across NK1, NK2, and NK3 (i.e. *CXCR4* and *CD69*), because taking whole-cluster averages of marker gene expression can obscure the true dynamics of expression across NK cells in the case of uniform expression gradients.(*17*)

Finally, the classical division of NK cells into active and quiescent populations is in line with our analysis. In our low-dimensional search for archetypes, we only identified active NK cell populations. This owes in part to the distinct, reproducible nature of these active expression programs. If the expressional program of quiescent NK cells is less pronounced, it could be difficult to detect in our low-dimensional archetype analysis. This is to say that, unlike T cells (discussed next), quiescent NK cells do not seem to have pronounced homeostatic activity, which we'd expect to manifest as elevation in some other expressional program. Interestingly, Rebuffet et al. did not identify a quiescent NK cell subtype either.(*17*)

Effector T cells and naïve T cells are archetypes, not cell types.

We identified three pan-tissue T cell archetypes. The first—**immune response**—represents effector T cells (*IL32*, *CD52*) involved in antigen presentation via MHC class I molecules (*HLA-A*, *HLA-B*, *HLA-C*, *B2M*), as well as high migratory and motile potential through cytoskeletal modulation (*ACTB*, *PFN1*, *SH3BGRL3*, *TMSB4X*, *TMSB10*, *S100A4*), **Figure 4C**.

The expression of *TMSB4X* and *SH3BGRL3* in an active state is consistent with their expression in a model of IL-21-stimulated T cells(*18*), and the expression of *S100A4* is consistent with a tissue-nonspecific, tissue-resident activated T cell signature.(*19*) Moreover, a module score derived from genes in archetype 1 recapitulates the T cell "effectorness" axis identified in Cano-Gamez et al,(*20*) and a gene module score derived only from MHC class I molecules in archetype 1 (*HLA-A*, *HLA-B*, *HLA-C*, *B2M*) is expressed increasingly along the "effectorness" axis, **Figure S6.** In this context, the observation of MHC class I expression in effector T cells is novel. This activity may be indicative of some as yet uncovered T cell-T cell interaction—potentially a capacity for cross-presentation (exogenous antigen presentation through MHC class I) to CD8+ T cells. This is a known function in dendritic cells.(*21*) This enrichment also may be due to an inflammatory environment or serve some autoimmune recognition purpose, or play a role in infection of T cells (as in HIV infection).

The second archetype—**glucocorticoid signaling**—is indicative of a glucocorticoid-mediated, anti-inflammatory (*TSC22D3*, *ANXA1*) stress response (*DNAJA1*, *JUN*, *RGCC*), **Figure 4C**. Genes in this archetype have been observed to be differentially expressed across subsets of stimulated T cells(*18*), and this archetype may represent a homeostatic mechanism to prevent excess inflammation.

The third archetype—**T-cell development and activation**—likely represents naïve T cells (*IL7R*) (*22*) with regulated intracellular signaling (*PDE4D*, *SMAP2*, *ZFP36L2*) to maintain a state primed for activation, or may be an intermediate state of activation, **Figure 4C**. Furthermore, phosphodiesterases, of which PDE4 is the most well-studied in T cells, regulate intracellular cAMP during T cell activation and differentiation. PDE4 is the target of current pharmacotherapy in several diseases.(*23*) Given the gene expression pattern in this archetype, it is also possible that it comprises memory T cells in part or in full (*20*), but this distinction does not prevent us from identifying this archetype as T cells that are primed for cytokine activation. In this case, archetype analysis may have identified a function that is shared across multiple T cell subtypes, which highlights both the difficulty inherent in subtyping cells, and the power of archetype analysis to overcome this discretization.

The identification of effector and naïve T cells as archetypes is interesting in terms of our understanding of T cell biology. Cell typing implies discrete states, whereas archetyping naturally accommodates a transcriptional continuum. Given the repeated identification of continuous naïve-to-effector T cell trajectories, describing these states as extrema of a continuum may be a more accurate description of the underlying biology.

Fibroblast archetypes highlight distinct roles in ECM maintenance.

Fibroblasts are also found in all tissues, and are increasingly recognized as having regulatory roles beyond ECM building. We find three pan tissue fibroblast archetypes.
The first fibroblast archetype—**extracellular matrix organization**—expressed markers that include an ECM-productive and myofibroblast-like phenotype. These include markers of production of a variety of ECM components, including microfibrils (*MFAP5*, *MFAP4*), fibulin (*FBLN1*), decorin (*DCN*), gelsolin (*GSN*), vimentin (*VIM*), and others (*MGP*, *TAGLN*). This archetype also has features of complement and growth factor interaction (*CFD*, *C1R*, *C3*, *C7*, *IGFBP5*, *IGFBP6*, *IGFBP7*) and factors that inhibit ECM degradation (*CST3*, *SERPINF1*), **Figure 4D**.

The second archetype—**collagen synthesis and remodeling**—seems to be a collagen production specialist phenotype (*COL1A1*, *COL6A3*, *COL1A2*, *COL3A1*, *COL6A2*), and may display ECM degradative features (*MMP2*), **Figure 4D**.

The extracellular matrix is composed of a wide variety of components, and it would seem unlikely that all components share a single optimal biochemical environment for their production. It could make sense that the production of these components be split among several specialist phenotypes, each with its own set of optimal biochemical conditions that are suited for producing specific components of the ECM. In our analysis, collagen production appears to be handled by a specialist fibroblastic phenotype, separate from those that handle the production of other components mentioned here (i.e. microfibrils, fibulin, decorin, gelsolin, vimentin).

A recent analysis of murine colonic and human lung fibroblasts found five fibroblastic functional archetypes that broadly reflect the above-listed.(*24*) However, the archetype alignment process across multiple tissues and donors provides more interpretable marker genes. Thus, our

analysis provides further insight into gene and pathway coregulation involved in fibroblasts' immune modulatory roles. Specifically, where Adler et al. see distinct ECM productive, myofibroblastic, and immune regulatory archetypes, we see immune regulatory functions and ECM production co-occurring in a common phenotype. Additionally, we have described novel granularity of the division of labor in production of ECM components.

The third archetype is **stress response** (*MT2A*, *MT1M*, *MT1X*, *HSP90AA1*, *JUNB*, *GADD45B*, *IER3*) with elevated metabolic activity (*EEF1A1*, *RPL13*, *RPL10*, *RPL41*, *RPS18*) and inflammation regulation (*IFITM3*, *CEBPD*), **Figure 4D**. Genes in this archetype are upregulated in oxidative and other forms of stress.

Endothelial cell archetype responds to hypoxia.

In our systematic analysis of endothelial cell function, we identified an archetype with signatures of enhanced cell motility and proliferation (*SPARCL1*, *RGCC*), **Figure 4E**. This archetype is directly involved in **angiogenesis** (*KDR*, *SPARCL1*).

The angiogenic marker (*KDR*) in this endothelial cell archetype is influenced by local hypoxia. As oxygenation varies along the length of blood vessels, this archetype may represent hypoxic extrema of a continuous trajectory of endothelial cells along blood vessels, and therefore recovers a real spatial coordinate of cellular organization.

The finding of only a single endothelial cell archetype across donors does not contradict the initial hypothesis—that a tradeoff between multiple tasks gives rise to simplices in trait space. This is because, while we were only able to identify one endothelial cell archetype across donors and tissues based on marker gene expression, each donor-tissue still exhibited multiple tasks. Simply put, the other specialist phenotypes did not make it through the filters in the archetype alignment process, **Figure 1B**. Thus, the single endothelial cell archetype discussed here is rather an artefact of conservative filtering and requirements that we have set to be called an archetype.

Archetypes have distinct expressional programs.

We calculated the average gene expression (raw counts) of the defining genes of each archetype. We then compared the gene expression of archetype-defining genes within their own archetype to the expression of these genes in other archetypes. In all cell types discussed—macrophages, NK cells, T cells, fibroblasts, and endothelial cells—archetypes were defined by unique expressional programs. Overall, we conclude that the defining genes of each archetype are highly expressed within their own archetype, and lowly expressed in other archetypes, **Figure 4A-E** (bottom row). Thus, the archetypes are phenotypically distinct from each other.

Discussion.

Our results indicate that we can answer the two fundamental questions posed at the beginning of this paper in the affirmative. Most cell types are amenable to Pareto analysis and their gene

expression patterns are well fit by low dimensional polytopes, and in some cases the vertices of these polytopes can be interpreted as corresponding to functions performed by those cell types. These results are robust as they are reproducible across donors and across different tissues with shared cell types. However, it is important to note that there are some limitations to the analysis and that not every cell type falls neatly into this paradigm.

First, there is clearly some experimental variation as for a given cell type, while a majority of donors may fit a polytope, some donors do not, and the source of this variation is unclear at this time. Also, some cell types are not well fit by polytopes and it is unclear if this reflects their true biology or whether it is related to experimental artifacts. Nonetheless, it is notable that some of the most widely shared cell types across tissue are well described by polytopes and that the vertices can often be interpreted in terms of their biological function.

One possible source of artefacts in gene expression is tissue processing.(*3*) After attempting to control for this with several conservative filters, and ensuring reproducibility of our findings across donors and tissues, we did not, in general, observe artefactual genes appearing in our gene lists of interest. An exception might be the stress response archetype in fibroblasts, but we cannot say for certain that this phenotype is *only* present due to tissue processing artefacts; fibroblasts are known to be early responders to stress in healthy physiology.

In this investigation, we fit up to six vertices for each donor-tissue-cell type. Then, we applied a filter to retain only vertices that were present across two-thirds of donors within each cell type. This left us with archetypes for a given cell type that were universal across donors, and, by indirect and unenforced consequence, found in many tissues. This conservative filter also discards tissue- or donor-specific archetypes that may be biologically relevant. In the present study, we lack the power to draw conclusions from these potential archetypes, and thus restricted our analysis. Furthermore, as is the case in several cell types discussed, more vertices were fit than ended up in archetypes. For example, in endothelial cells, we found one archetype of angiogenesis. However, we found many more vertices than were included in this archetype. These vertices could be donor- or tissue-specific archetypes that did not meet the donor abundance filter.

Finally, we note that this approach provides a powerful way to combine two disjoint perspectives about cell type and cell state: whether they are discrete states or continuous states. The idea of having multiple discrete functions performed by a single cell type, and different cells within that type having a different balance of performing those functions that varies in a continuous fashion, provides a powerful paradigm to interpret and understand cell type and cell state distinctions.

Conclusions.

We used a principled method to identify archetypes that represent core functions performed by a cell type. For several ubiquitous cell types, we recapitulated known functions without incorporating prior biological knowledge. Placing these cellular tasks on a continuum, with cells performing some combination of specialized functions, advances our understanding of cell

typology, beyond discretization of cell populations. Compared to canonical marker gene analysis, this continuous framing allows identification of functional marker genes that is unbiased, and captures continuous cellular heterogeneity.

References and Notes:

1. H. Li *et al.*, Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science* **375**, eabk2432 (2022).
2. C. Tabula Muris, A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590-595 (2020).
3. C. Tabula Sapiens *et al.*, The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
4. A. Regev *et al.*, The Human Cell Atlas. *Elife* **6**, (2017).
5. I. Virshup *et al.*, The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat Biotechnol* **41**, 604-606 (2023).
6. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).
7. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).
8. Z. Yao *et al.*, A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598**, 103-110 (2021).
9. S. Domcke, J. Shendure, A reference cell tree will serve science better than a reference cell atlas. *Cell* **186**, 1103-1114 (2023).
10. O. Shoval *et al.*, Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* **336**, 1157-1160 (2012).
11. Y. Hart *et al.*, Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat Methods* **12**, 233-235, 233 p following 235 (2015).
12. T. M. Inc. (2022).
13. W. Hou, Z. Ji, Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods*, (2024).
14. E. Mass, F. Nimmerjahn, K. Kierdorf, A. Schlitzer, Tissue-specific macrophages: how they develop and choreograph tissue biology. *Nat Rev Immunol* **23**, 563-579 (2023).
15. S. Chakarov *et al.*, Two distinct interstitial macrophage populations coexist across tissues in specific subtissular niches. *Science* **363**, (2019).
16. M. Nahrendorf, F. K. Swirski, Abandoning M1/M2 for a Network Model of Macrophage Function. *Circ Res* **119**, 414-417 (2016).
17. L. Rebuffet *et al.*, High-dimensional single-cell analysis of human natural killer cell heterogeneity. *Nat Immunol* **25**, 1474-1488 (2024).
18. Z. Thomson *et al.*, Trimodal single-cell profiling reveals a novel pediatric CD8alphaalpha(+) T cell subset and broad age-related molecular reprogramming across the T cell compartment. *Nat Immunol* **24**, 1947-1959 (2023).
19. P. A. Szabo *et al.*, Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat Commun* **10**, 4706 (2019).

20. E. Cano-Gamez *et al.*, Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4(+) T cells to cytokines. *Nat Commun* **11**, 1801 (2020).
21. O. P. Joffre, E. Segura, A. Savina, S. Amigorena, Cross-presentation by dendritic cells. *Nat Rev Immunol* **12**, 557-569 (2012).
22. M. Andreatta *et al.*, Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nat Commun* **12**, 2965 (2021).
23. M. Bielenberg, R. Kurelic, S. Frantz, V. O. Nikolaev, A mini-review: phosphodiesterases in charge to balance intracellular cAMP during T-cell activation. *Front Immunol* **15**, 1365484 (2024).
24. M. Adler *et al.*, Emergence of division of labor in tissues through cell interactions and spatial cues. *Cell Rep* **42**, 112412 (2023).

Acknowledgements

We thank G. Ed Marti for his helpful discussion of statistical methods. We thank Madhav Mantri and Jaeyoon Lee for their discussion of the investigation. We thank Avi Mayo and Miri Adler for their discussion of the ParTI analysis package.

Some of the computing for this project was performed on the Sherlock and Bruno clusters. We would like to thank Stanford University, the Stanford Research Computing Center, and the Chan-Zuckerberg Biohub for providing computational resources and support that contributed to these research results.

Funding

Chan-Zuckerberg Biohub

Author contributions

Conceptualization: GC, UW, SRQ
Data curation: GC, UW, SRQ
Formal analysis: GC, UW, SRQ
Funding acquisition: SRQ
Investigation: GC, UW, SRQ
Methodology: GC, UW, SRQ
Project administration: GC, UW, SRQ
Resources: GC, UW, SRQ
Software: GC, UW, SRQ
Supervision: UW, SRQ
Validation: GC, UW, SRQ
Visualization: GC, UW, SRQ
Writing – original draft: GC, UW, SRQ
Writing – review & editing: GC, UW, SRQ

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

All data used in this study is publicly available from (*3*). All source code used in this study is available as **Supplemental Code**.