scientific reports

OPEN

Single cell profiling of capillary blood enables out of clinic human immunity studies

Tatyana Dobreva[1,3]✉, David Brown[2,3], Jong Hwee Park[2] & Matt Thomson[2]✉

An individual's immune system is driven by both genetic and environmental factors that vary over time. To better understand the temporal and inter-individual variability of gene expression within distinct immune cell types, we developed a platform that leverages multiplexed single-cell sequencing and out-of-clinic capillary blood extraction to enable simplified, cost-effective profiling of the human immune system across people and time at single-cell resolution. Using the platform, we detect widespread differences in cell type-specific gene expression between subjects that are stable over multiple days.

Increasing evidence implicates the immune system in an overwhelming number of diseases, and distinct cell types play specific roles in their pathogenesis[1,44]. Studies of peripheral blood have uncovered a wealth of associations between gene expression, environmental factors, disease risk, and therapeutic efficacy[45,46,48]. For example, in rheumatoid arthritis, multiple mechanistic paths have been found that lead to disease, and gene expression of specific immune cell types can be used as a predictor of therapeutic non-response[46]. Furthermore, vaccines, drugs, and chemotherapy have been shown to yield different efficacy based on time of administration, and such findings have been linked to the time-dependence of gene expression in downstream pathways[28,36,47]. However, human immune studies of gene expression between individuals and across time remain limited to a few cell types or time points per subject, constraining our understanding of how networks of heterogeneous cells making up each individual's immune system respond to adverse events and change over time. The advent of single-cell RNA sequencing (scRNA-seq) has enabled the interrogation of heterogeneous cell populations in blood without cell type isolation and has already been employed in the study of myriad immune-related diseases[1–4]. Recent studies employing scRNA-seq to study the role of immune cell subpopulations between healthy and ill patients, such as those for Crohn's disease[5], Tuberculosis[6], and COVID-19[7], have identified cell type-specific disease relevant signatures in peripheral blood immune cells; however, these types of studies have been limited to large volume venous blood draws which can tax already ill patients, reduce the scope of studies to populations amenable to blood draws, and often require larger research teams to handle the patient logistics and sample processing costs and labor. In particular, getting repeated venous blood draws within a single day and/or multiple days at the subject's home has been a challenge for older people with frail skin and those on low dosage Acetylsalicylic acid[8]. This dependence on venous blood dramatically impacts our ability to understand the high temporal dynamics of health and disease.

Capillary blood sampling is being increasingly used in point-of-care testing and has been advised for obese, elderly, and other patients with fragile or inaccessible veins[9–12]. The reduction of patient burden via capillary blood sampling could enable researchers to perform studies on otherwise difficult or inaccessible populations, and at greater temporal resolution. Additionally, capillary blood is being shown to be comparable to traditional venous blood draws for a variety of applications. For example, Catala et al. have shown that 39 out of 45 clinically relevant metabolites had overlapping ranges between capillary blood vs traditional venous blood draws[13], and Toma et al. have shown strong correlation (Spearman correlation coefficient ≥ 0.95) between bulk RNA sequencing data between capillary and venous blood from the same donor[14]. However, to date, scRNA-seq of human capillary blood has not yet been validated nor applied to study the immune system. In order to make small volumes of capillary blood (100 ul) amenable to scRNA-seq we have developed a platform which consists of a painless vacuum-based blood collection device, sample de-multiplexing leveraging commercial genotype data, and an analysis pipeline used to identify time-of-day and subject specific genes. The potential of our platform is rooted in enabling large scale studies of immune state variation in health and disease across people. The

[1]Andrew and Peggy Cherng Department of Medical Engineering, California Institute of Technology, Pasadena, CA, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [3]These authors contributed equally: Tatyana Dobreva and David Brown. ✉email: tdobreva@caltech.edu; mthomson@caltech.edu

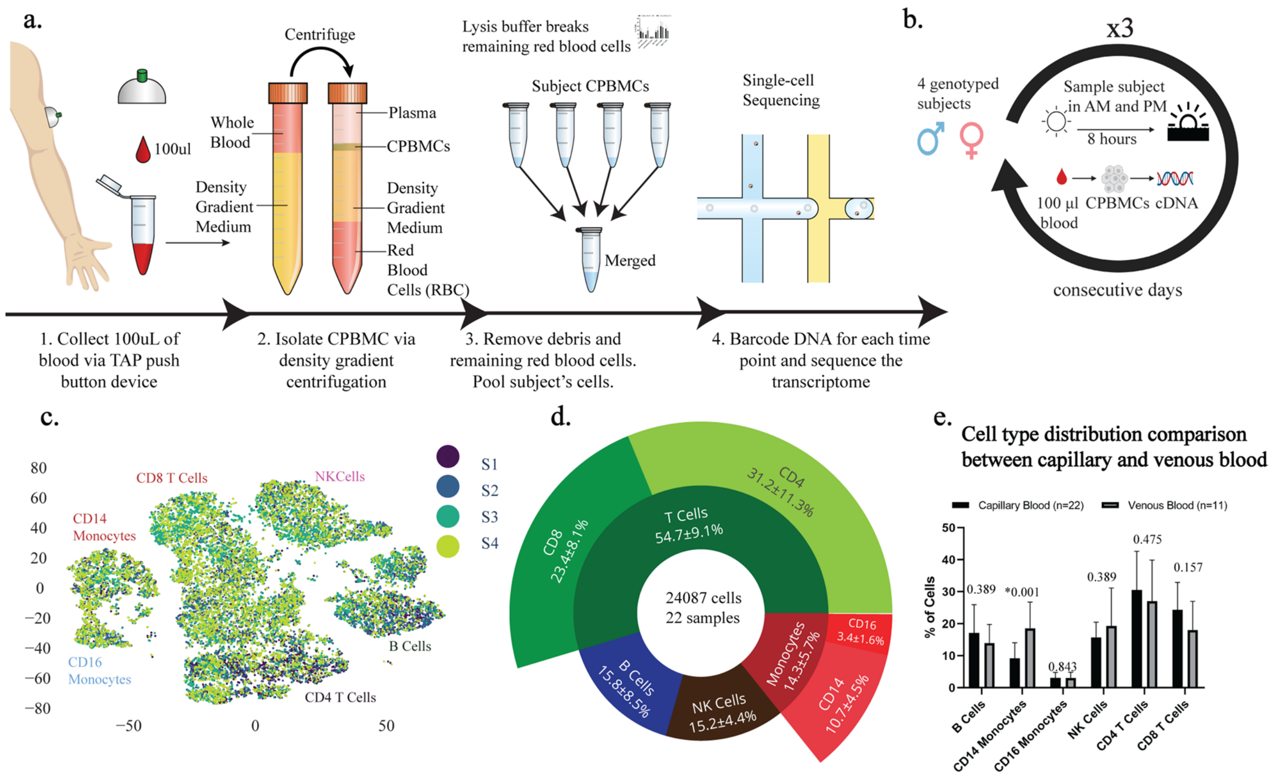

Figure 1. Experimental workflow and consistency of capillary blood sampling. (**a**) Experimental workflow for capillary blood immune profiling. 1. Blood is collected using the TAP device from the deltoid. 2. Capillary peripheral blood mononuclear cells (CPBMCs) are separated via centrifugation. 3. Red blood cells are lysed and removed, and samples from different subjects are pooled together. 4. Cell transcriptomes are sequenced using single-cell sequencing. (**b**) Time-course study design. CPBMCs are collected and profiled from 4 subjects (2 male, 2 female) each morning (AM) and afternoon (PM) for 3 consecutive days. (**c**) 2-dimensional t-SNE projection of the transcriptomes of all cells in all samples. Cells appear to cluster by major cell type (Fig. S6) (**d**) Immune cell type percentages across all samples shows stable cell type abundances (includes cells without subject labels). (**e**) Cell type ratios between capillary blood from this study, and venous blood from 3 other studies were the same, with the exception of CD14$^+$ Monocytes, which are more abundant in venous blood (FDR < 0.05, 2-sided student t-test, multiple comparison corrected) The q-values are displayed for each cell type comparison.

high-dimensional temporal transcriptome data could be paired with computational approaches to predict and understand emergence of pathological immune states. Most importantly, our platform makes collection and profiling of human immune cells less invasive, less expensive and as such more scalable than traditional methods rooted in large venous blood draws.

Results

Platform for low-cost interrogation of single-cell immune gene expression profiles.

Our platform is comprised of a protocol for isolating capillary peripheral blood mononuclear cells (CPBMCs) using a touch activated phlebotomy device (TAP)[9], pooling samples to reduce per-sample cost using genome-based demultiplexing[15], and a computational package that leverages repeated sampling to identify genes that are differentially expressed in individuals or between time points, within subpopulations of cells (Fig. 1a). Using a painless vacuum-based blood collection device such as the commercial FDA-approved TAP to collect capillary blood makes it convenient to perform at-home self-collected sampling and removes the need for a trained phlebotomist, increasing the ease of acquiring more samples. The isolation of CPBMCs is done using gradient centrifugation and red blood cells are further removed via a red blood cell lysis buffer. The cells from the different subjects are pooled, sequenced via scRNA-seq using a single reagent kit, and demultiplexed[15] via each subject's single-nucleotide polymorphisms (SNPs), reducing the per-sample processing cost. By pooling the data across all 6 time points, and using a genotype-free demultiplexing software (popscle), we were able to identify which cells belonged to which subject across time points, removing the need for a separate genotyping assay to link subjects together across batches.

Single-cell RNA sequencing (scRNA-seq) of low volume capillary blood recovers distinct immune cell populations stably across time.

As a proof-of-concept, we leveraged our scRNA-seq of capillary blood platform to identify genes that exhibit diurnal behavior in subpopulations of cells and find subject-specific immune relevant gene signatures. We performed a three-day study in which we processed capillary

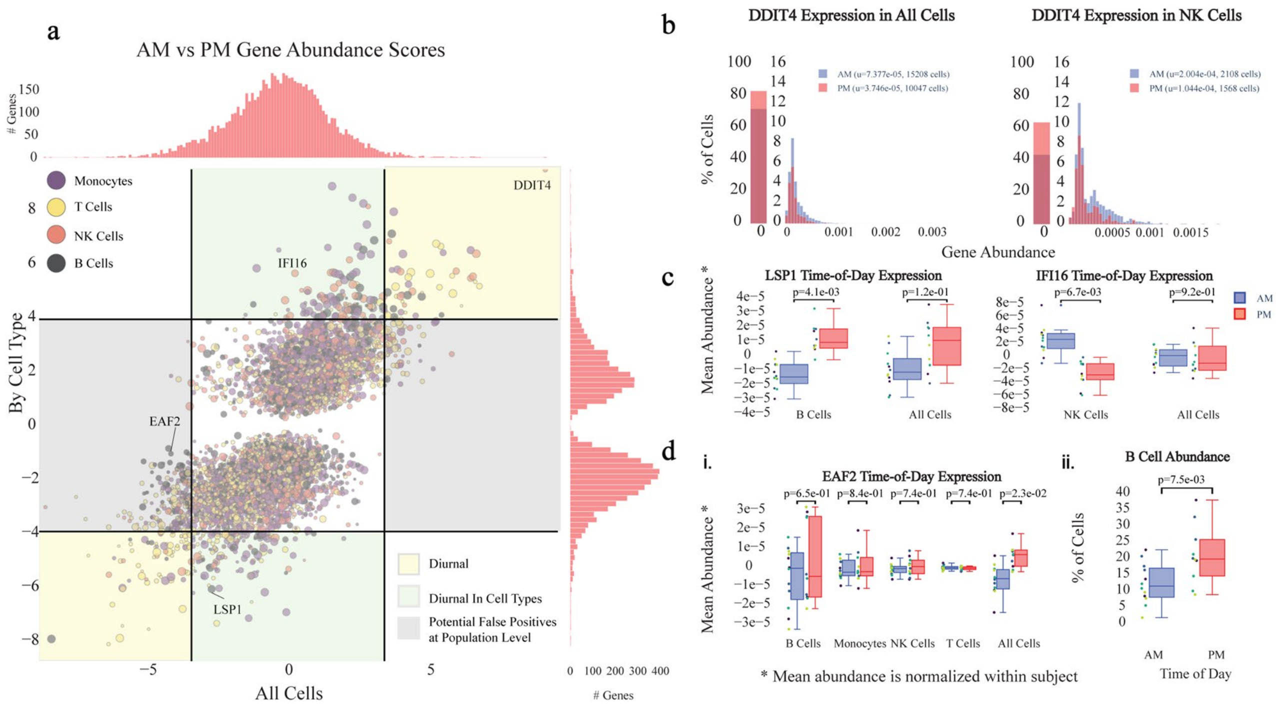

Figure 2. Diurnal variability in subpopulations of capillary blood (**a**) Magnitude (Z-score) of the difference in AM vs PM gene expression across the whole population of cells (x) vs the cell type with the largest magnitude Z-score (y). Points above or below the significance lines (FDR < 0.05, multiple comparison correction) display different degrees of diurnality. The size of each marker indicates the abundance of the gene (the largest percent of cells in a subpopulation that express this gene). (**b**) Distribution of expression of DDIT4, a previously identified circadian rhythm gene[22], shows diurnal signal across all cells, as well as individual cell types, such as natural killer (NK) cells. u indicates the mean fraction of transcripts per cell (gene abundance). (**c**) Example of newly identified diurnal genes, LSP1 and IFI16 that could be missed if analyzed at the population level (**d**) Example of a gene, EAF2, that could be falsely classified as diurnal (i) without considering cell type subpopulations due to a diurnal B cell abundance shift (ii).

blood from four subjects in the morning and afternoon, totaling 24,087 cells across 22 samples (Fig. 1b). Major immune cell types such as T cells (CD4$^+$, CD8$^+$), Natural Killer cells, Monocytes (CD14$^+$, CD16$^+$), and B cells are present in all subjects and time points with stable expression of key marker genes (Fig. 1d, Fig. S1), demonstrating that these signals are robust to technical and biological variability of CPBMC sampling (Fig. 1c). In order to compare cell type distributions derived from our method with venous blood draws, we used data from 11 healthy subjects provided by three independent studies[7,16,17] (Table S4). CD14$^+$ Monocytes make up a higher percentage of PBMCs in venous blood (n = 11) versus capillary blood (n = 22) (FDR < 0.05, 2-sided student t-test, multiple comparison corrected), while other cell types do not have a significant difference in distributions (Fig. 1e).

High frequency scRNA-seq unveils new diurnal cell type-specific genes. Genes driven by time-of-day expression, such as those involved in leukocyte recruitment[18] and regulation of oxidative stress[19], have been determined to play an important role in both innate and adaptive immune cells[20]. Medical conditions such as atherosclerosis, parasite infection, sepsis, and allergies display distinct time-of-day immune responses in leukocytes[21], suggesting the presence of diurnally expressing genes that could be candidates for optimizing therapeutic efficacy via time-of-day dependent administration. However, studies examining diurnal gene expression in human blood have been limited to whole blood gene panels via qPCR, or bulk RNA-seq[22–24].

Leveraging our platform, which enables single-cell studies of temporal human immune gene expression, we detected 395 genes (FDR < 0.05, multiple comparison corrected) exhibiting diurnal activity within at least one cell subpopulation (Fig. 2a). Among the 20 top diurnally classified genes, we found that 35% of those genes were previously correlated with circadian behavior (Table S1), such as DDIT4[22] (Fig. 2b), SMAP2[25], and PCPB1[26]. However, only 119/395 (30.1%) of these genes are detected as diurnal at the whole population level (FDR < 0.05, multiple comparison corrected), suggesting there may be many more diurnally-varying genes than previously discovered. For example, IFI16 and LSP1 (Fig. 2c) have diurnal expression only in NK cells and B cells, respectively, and display previously unreported transcriptional diurnal patterns. In particular, LSP1 has been implicated in numerous leukemias and lymphomas of B cell origin[27]. Given previous evidence of increased efficacy of time-dependent chemotherapy administration[28,29] and tumor cells exhibiting out-of-sync behavior compared to normal cells[30], understanding LSP1's diurnal expression pattern can potentially guide timely administration of candidate therapeutics. Out of the identified 395 diurnally-varying genes, 114 (29%) are considered druggable under the drug gene interaction database (https://www.dgidb.org/).

scRNA-seq profiling distinguishes diurnal gene expression from cell type abundance changes. We also detected 406 genes (FDR < 0.05, multiple comparison corrected) exhibiting diurnal behavior when analyzed at the population level, such as EAF2, that do not display diurnal variation in any of our major cell types (Fig. 2d.i). Such false positives may come from diurnal shifts in cell type abundance rather than up- or down-regulation of genes. In the case of EAF2, which is most abundant in B cells, we hypothesized that the diurnality detected at the population level was a result of an increase of B cell abundance in the afternoon, and verified this in our data ($p = 7.5 \times 10^{-3}$, one-sided student-t test) (Fig. 2d.ii). This finding highlights the importance of looking at expression within multiple cell types to avoid potentially misleading mechanistic hypotheses.

Individuals exhibit robust cell type-specific differences in genes and pathways relevant to immune function. Gene expression studies of isolated cell subpopulations across large cohorts of people have revealed a high degree of variability between individuals that cannot be accounted for by genetics alone, with environmental effects that vary over time likely playing a critical role[31,32]. Furthermore, these transcriptomic differences have been linked to a wide range of therapeutic responses, such as drug-induced cardiotoxicity[33]. However, while immune system composition and expression has been shown to be stable over long time periods within an individual, acute immune responses generate dramatic immune system changes, meaning that large single time point population studies are unable to establish whether variability between individuals is stable or the result of dynamic response to stimuli[34].

To probe the stability of individual gene expression signatures at the single-cell level, we used our pipeline to identify genes whose variation in gene expression is most likely caused by intrinsic intersubject differences rather than high frequency immune system variability. We compared the mean gene expressions of all time points between subjects in all cell types and identified 1284 genes (FDR < 0.05, multiple comparison corrected) that are differentially expressed in at least one subpopulation of cells. Like Whitney et al., we found MHC class II genes, such as HLA-DRB1 and HLA-DRA (Fig. 3a) to be among the largest sources of variation between subjects[35]. Additionally, we found that DDX17, which was classified by Whitney et al. as a gene with high intersubject variability, but low intrasubject variability via repeat sampling over longer time scales, may be a new class of temporally varying gene that varies by day of week, having consistently increasing expression each subsequent sampling day. This stresses the importance of high frequency sampling for identifying genes with the most intrinsic interindividual variability.

Numerous subject-specific genes are revealed in specific immune cell types. Within the 1284 genes with intrinsic interindividual variability, we found myriad disease-relevant genes for all subjects and cell types, which can be explored at our interactive online portal (https://capblood-seq.caltech.edu). As just one example, subject S1's monocytes have a consistent downregulation ($p = 9.1 \times 10^{-7}$, two-sided student t-test) of LIPA, a gene that is implicated in Lysosomal Acid Lipase Deficiency (Fig. 3c). Given the low abundance of monocytes in blood samples, such findings would typically only be discovered from a targeted blood test or RNA sequencing of isolated monocytes, either of which would only be performed if the disease was already suspected; this showcases how automated discovery in heterogeneous cell populations can be leveraged for personalized, preventative care.

Immune function and disease pathways are enriched in subject-specific genes. Given that genes do not act alone, we also found cell type-specific pathway differences among subjects. In particular, Subject 2's S100A8, S100A9, and S100A12 genes, calcium-binding proteins that play an important role in macrophage inflammation, are significantly downregulated in monocytes ($p_{S100A8} = 1.3 \times 10^{-5}$, $p_{S100A9} = 9.0 \times 10^{-5}$, $p_{S100A12} = 3.0 \times 10^{-4}$, two-sided student t-test) compared to other subjects (Fig. S2). We further explored our findings by inspecting the pathways that are most enriched in individual and time-varying genes, and found that genes that are implicated in immune system function ($p = 0.085$) and immune diseases ($p = 0.029$) are more present in subject-specific genes (Fig. 3b). This stands in contrast to pathways of core cellular functions such as genetic information processing ($p = 0.029$) and metabolism ($p = 0.095$), which are less present in subject-specific genes.

Discussion. Genome and transcriptome sequencing projects have unveiled millions of genetic variants and associated gene expression traits in humans[37,38]. However, large-scale studies of their functional effects performed through venous blood draws require tremendous effort to undertake, and this is exacerbated by the cost and complexity of single-cell transcriptome sequencing. Efforts such as the Immune Cell Census[39] are already underway to perform single-cell profiling of large cohorts, but reliance on venous blood draws of PBMCs will likely limit the diversity and temporal resolution of their sample pool. Our platform gives researchers direct, scalable access to high resolution immune system transcriptome information of human subjects, lowering the barrier of entry for myriad new research avenues. Examples of such studies include: 1. tracking vulnerable populations over time, such as monitoring clonal expansion of $CD8^+$ T cells in Alzheimer's disease progression[1], 2. profiling of individuals who are under home care to track disease progression and therapeutic response, such as transplant patients and people under quarantine, and 3. tracking how stress, diet, and environmental conditions impact the immune system at short and long time scales, particularly in underrepresented populations who do not have easy access to hospitals or research institutions, such as people in rural or underdeveloped areas. Larger, more diverse subject pools coupled with time course studies of cell type gene expression in health and disease will have a dramatic impact on our ability to understand the baseline and variability of immune function.

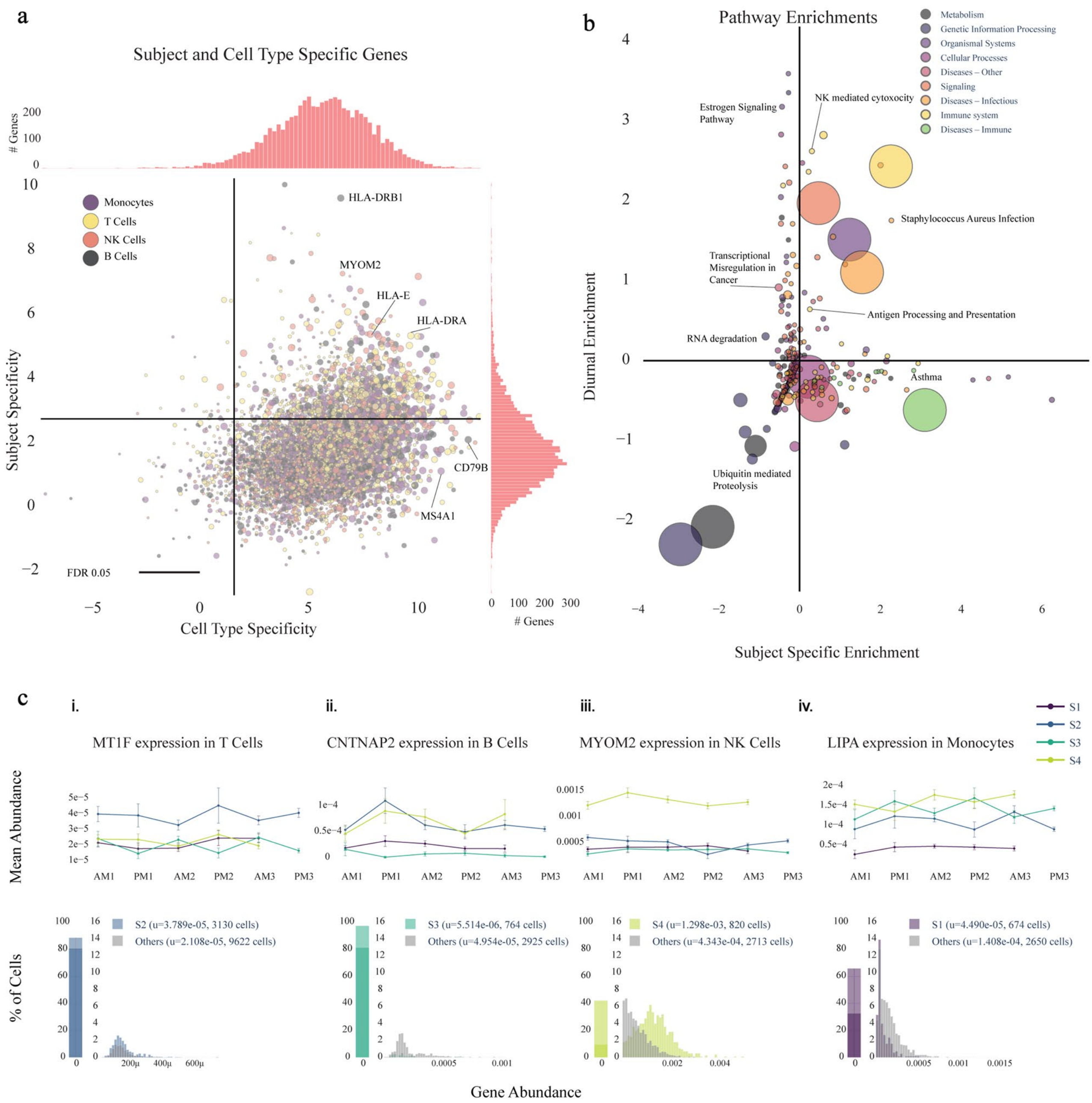

Figure 3. Subject variability in immune and disease-relevant genes and pathways. (**a**) Magnitude ($\log_2$ F statistic) of the variability in expression of genes between different cell types (x) and between subjects (y). 1284/7034 (18.3%) of genes are above the subject specificity significance line (FDR < 0.05, multiple comparison correction) and are classified as subject-specific. Several MHC class II genes (HLA-X) are strongly subject-specific, consistent with previous findings[35]. (**b**) KEGG pathways grouped into categories and their enrichment (Z-score from 2-proportion Z-test) among the top 250 diurnally and subject-varying genes vs all genes. Immune system and disease pathways are significantly enriched (p = 0.029), supportive of the conclusion that immune and disease-related genes are highly subject dependent. The large circles indicate the enrichment of the category overall, and the sizes of the smaller pathway points indicate the number of genes associated with the pathway. (**c**) Subject and cell type specific gene examples for each subject and cell type with the upper row displaying the trace of mean gene expression across time-points and the bottom row showing gene abundance shifts for the subjects of interest.

Online content

Online web portal is available to explore data presented in the main figures for study summary, diurnal and subject specific genes via https://capblood-seq.caltech.edu.

Methods

Human study cohort. This study was conducted at Caltech. Four healthy adults (2 male, 2 female) were recruited (Table S3). All participants provided written informed consent. The study was approved by the Institutional Review Board (IRB) at Caltech and all methods were performed in compliance with relevant guidelines and regulations. The blood collection took place in a non-BSL room to make sure the subjects were not exposed to pathogens. Subject blood was collected roughly 8 h apart over three consecutive days.

CPBMC isolation. 100 µl of capillary blood was collected via push-button collection device (TAP from Seventh Sense Biosystems). For each blood draw, the site of collection was disinfected with an alcohol wipe and the TAP device was placed on the deltoid of the subject per device usage instructions. The button was pushed, and then blood was collected for 2–7 min until the indicator turned red. Blood was extracted from the TAP device by gently breaking the seal foil, and mixed with PBS + 2% FBS to 1 ml. The mixture was slowly added to the side of a SepMate tube (SepMate-15 IVD, Stem Cell Technologies) containing 4.5 ml of Lymphoprep (#07811, Stem Cell Technologies) and centrifuged for 20 min at 800 RPM. Approximately 900 µl of CPBMC layer was extracted below the plasma layer. To further remove red blood cells, 100 µl of red blood cell lysis buffer (eBioscience 10× RBC Lysis Buffer, #00-4300-54) was added to the CPBMCs and incubated at RT for 15 min. The CPBMC pellet was washed twice with PBS and centrifuged at 400 rpm for 5 min. Cells were counted using trypan blue via an automated detector (Countess II Automated Cell Counter) and subjects' cells were pooled together for subsequent single-cell RNA sequencing.

Single-cell RNA sequencing. Subject pooled single-cell suspensions were loaded onto a Chromium Single Cell Chip (10X Genomics) based on manufacturer's instructions (targeted 10,000 cells per sample, 2500 cells per person per time point). Captured mRNA was barcoded during cDNA synthesis and pooled for Illumina sequencing (Chromium Single Cell 3′ solution—10X Genomics). Each time point was barcoded with a unique Illumina sample index, and then pooled together for sequencing in a single Illumina flow cell. The libraries were sequenced with an 8-base index read, 26-base read 1 containing cell-identifying barcodes and unique molecular identifiers (UMIs), and a 91-base read 2 containing transcript sequences on a NovaSeq 6000.

Single-cell dataset generation. FASTQ files from Illumina were demultiplexed and aligned using Cell Ranger v3.0 (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger) and the hg19 reference genome with all options set to their defaults.

Sample demultiplexing. FASTQ files from the single-cell sequencing Illumina libraries were aligned against the hg19 (human) reference genome using Cellranger v3.0 count function. SNPs were detected in the aligned data using freebayes (https://github.com/ekg/freebayes), which creates a combined variant call format (VCF) file, one per sample. SNPs were then grouped by cell barcode using popscle dsc-pileup (https://github.com/statgen/popscle). The SNP files for all samples were then merged into a single dsc-pileup file, and cell barcodes were disambiguated by providing a unique identifier per sample. Freemuxlet (popscle freemuxlet) was then run with default parameters to group cells into 4 subjects. This generates a probability of whether each cell barcode belongs to each subject, given the detection of single nucleotide polymorphism (SNPs) in reads associated with that cell barcode. Each cell was then assigned to the subject with the highest probability. Cells with low confidence (ambiguous cells) and high confidence in more than one subject (multiplets) were discarded, using popscle's default confidence thresholds. See the README at https://github.com/thomsonlab/capblood-seq for detailed instructions.

Debris removal. The raw cell gene matrix provided by Cell Ranger contains gene counts for all barcodes present in the data. To remove barcodes representing empty or debris-containing droplets, a debris removal step was performed. First, a UMI count threshold was determined that yielded more than the expected number of cells based on original cell counts (15,000). All barcodes below this threshold were discarded. For the remaining barcodes, principal component analysis (PCA) was performed on the log-transformed cell gene matrix, and agglomerative clustering was used to cluster the cells. The number of clusters was automatically determined by minimizing the silhouette score among a range of numbers of clusters (6 to 15). For each cluster, a barcode dropoff trace was calculated by determining the number of barcodes remaining in the cluster for all thresholds in increments of 50. These cluster traces were then clustered into two clusters using agglomerative clustering—the two clusters representing "debris" with high barcode dropoff rates and "cells" with low barcode drop-off rates. All clusters categorized as "debris" were then removed from the data.

Gene filtering. Before cell typing, genes that have a maximum count less than 3 are discarded. Furthermore, after cell typing, any genes that are not present in at least 10% of one or more cell types are discarded.

Data normalization. Gene counts were normalized by dividing the number of times a particular gene appears in a cell (gene cell count) by the total gene counts in that cell. Furthermore, for visualization only, the gene counts were multiplied by a constant factor (5000), and a constant value of 1 was added to avoid zeros and then log transformed.

Cell typing. We used single cell Variational Inference (scVI) to transform the raw cell gene expression data into a 10-dimensional variational autoencoder latent space[40]. The variational autoencoder is conditioned on sample batch, creating a latent space which is independent of any batch-specific effects. The variational autoencoder parameters: learning rate = 1e−3, number of epochs = 50.

Agglomerative clustering (sci-kit learn) was used to generate clusters from the latent cell gene expression data. These clusters were then annotated based on known cell type marker genes (Fig. S1).

In order to resolve specific cell subtypes, such as those of T cells and Monocytes, we specified 13–15 clusters as an input for agglomerative clustering. For each study, we started at 13 clusters and incremented until all 4 major cell types and 2 subtypes were separable. In cases where agglomerative clustering yielded multiple clusters of the same cell type, these clusters were merged into a single cell type for analysis.

Venous and capillary blood comparison. In order to compare venous blood cell type distributions to capillary blood, raw gene count data was downloaded from each of the respective studies, and we performed the same cell typing pipeline as for our capillary data, first projecting the data into a latent space via scVI, followed by agglomerative clustering and manual annotation based on known cell type marker genes.

Diurnal gene detection. To identify genes that exhibit diurnal variation in distinct cell types, we developed a statistical procedure that detects robust gene expression differences between morning (AM) and evening (PM) samples. Given that gene expression is different between subjects, we first normalize the mean gene expression within each subject for each cell type.

$$\mu'_{g_i,s_j,c_n,k} = \mu_{g_i,s_j,c_n,k} - \left(\frac{\sum_{k=1}^{N_{s_j}} 1_{k\in AM}\mu_{g_i,s_j,c_n,k}}{2\sum_{k=1}^{N_{s_j}} 1_{k\in AM}} + \frac{\sum_{k=1}^{N_{s_j}} 1_{k\in PM}\mu_{g_i,s_j,c_n,k}}{2\sum_{k=1}^{N_{s_j}} 1_{k\in PM}} \right) \quad (1)$$

We take the mean gene expression μ for each gene g_i in all samples k for cell type c_n and subject s_j and renormalize it into μ' by subtracting the equally weighted mean of AM and PM samples (Eq. (1)). We then split the mean gene values into an AM group and a PM group and perform a statistical test (two-tailed student-t test) to determine whether to reject the null hypothesis that gene expression in AM and PM samples come from the same distribution. We then perform Benjamini–Hochberg multiple comparison correction at an FDR of 0.05 on all gene and cell type p-values to determine where to plot the significance threshold. For plotting the genes, we choose the Z-statistic corresponding to the minimum p-value among cell types for that gene. To determine diurnality at the population level, we repeated the procedure above with all cells pooled into a single cell type.

Subject and cell type specific gene detection. To classify genes as subject specific, we detect genes with mean gene expression levels that are robustly different between subjects in at least one cell type. For each cell type c_n and gene g_i, we create subject groups containing the mean gene expression values from each sample. To determine whether the gene expression means from the different subjects do not originate from the same distribution, we perform an ANOVA one-way test to get an F-statistic and p-value for each gene. We then perform Benjamini–Hochberg multiple comparison correction at an FDR of 0.05 on all gene and cell type p-values. For plotting the genes, we chose the F-statistic corresponding to the minimum p-value among cell types for that gene.

For determining gene cell type specificity, we performed a similar procedure. In particular, for each gene g_i, we create cell type groups containing the mean gene expression values for that cell type from each sample. We then perform a one-way ANOVA, and Benjamini–Hochberg multiple comparison correction at an FDR of 0.05.

Pathway enrichment analysis. Pathways from the KEGG database (python bioservices package) were used to calculate pathway enrichment for genes that were among the top 250 most diurnal and individual specific. All remaining genes present in the data were considered background. In order to normalize for gene presence across pathways, each gene was weighted by dividing the number of pathways in which that gene appears. For each KEGG pathway[41–43], the test statistic for a two-proportion z-test (python statsmodel v0.11.1) is used to determine pathway enrichment. From the top level pathway classes, we broke out "Diseases" into "Other", "Immune Diseases", and "Infectious Diseases" and separated "Immune System" from "Organismal System" to understand diurnal and subject-specific genes in an immune relevant context.

Figure art. All drawings (Figs. 1a,b, S2a) are generated using BioRender.com. Figure 1e was generated using GraphPad Prism 8.3.1.

Data availability

Gene expression matrix and relevant metadata are available on https://data.caltech.edu/records/1407. FASTQ files are not being released to protect the identity of the subjects.

Code availability

Custom code made for diurnal and subject specific gene detection is available on https://github.com/thomsonlab/capblood-seq.

Received: 23 July 2020; Accepted: 30 October 2020
Published online: 25 November 2020

References

1. Gate, D. *et al.* Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. *Nature* **577**, 399–404 (2020).
2. Kazer, S. W. *et al.* Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection. *Nat. Med.* **26**, 511–518 (2020).
3. Uniken-Venema, W. T. *et al.* Single-cell RNA sequencing of blood and ileal T cells from patients with Crohn's disease reveals tissue-specific characteristics and drug targets. *Gastroenterology* **156**, 812–815 (2019).
4. Der, E. *et al.* Tubular cell and keratinocyte single-cell transcriptomics applied to lupus nephritis reveal type I IFN and fibrosis relevant pathways. *Nat. Immunol.* **20**, 915–927 (2019).
5. Martin, J. C. *et al.* Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. *Cell* **178**, 1493–1508 (2019).
6. Cai, Y. *et al.* Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis. *EBioMedicine* **53**, 102686 (2020).
7. Lee, J. S. *et al.* Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci. Immunol.* **5**, 1554 (2020).
8. Bennett, D. *Head of Religious Orders Study and Rush Memory and Aging Project, Personal Communicate* (2020).
9. Blicharz, T. M. *et al.* Microneedle-based device for the one-step painless collection of capillary blood samples. *Nat. Biomed. Eng.* **2**, 151–157 (2018).
10. Lenicek Krleza, J., Dorotic, A., Grzunov, A. & Maradin, M. Capillary blood sampling: national recommendations on behalf of the Croatian Society of Medical Biochemistry and Laboratory Medicine. *Biochem. Med.* https://doi.org/10.11613/BM.2015.034 (2015).
11. Tang, R. *et al.* Capillary blood for point-of-care testing. *Crit. Rev. Clin. Lab. Sci.* **54**, 294–308 (2017).
12. Robison, E. H. *et al.* Whole genome transcript profiling from fingerstick blood samples: a comparison and feasibility study. *BMC Genomics* **10**, 617 (2009).
13. Catala, A., Culp-Hill, R., Nemkov, T. & D'Alessandro, A. Quantitative metabolomics comparison of traditional blood draws and TAP capillary blood collection. *Metabolomics*. https://doi.org/10.1007/s11306-018-1395-z (2018).
14. Toma, R. *et al.* A clinically validated human capillary blood transcriptome test for global systems biology studies. *Biotechniques*. https://doi.org/10.2144/btn-2020-0088 (2020).
15. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2017).
16. Hashimoto, K. *et al.* Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 24242–24251 (2019).
17. Hu, Y. *et al.* Single-cell transcriptome mapping identifies common and cell-type specific genes affected by acute Delta9-tetrahydrocannabinol in humans. *Sci. Rep.* https://doi.org/10.1038/s41598-020-59827-1 (2020).
18. He, W. *et al.* Circadian expression of migratory factors establishes lineage-specific signatures that guide the homing of leukocyte subsets to tissues. *Immunity* **49**, 1175–1190 (2018).
19. Zhao, Y. *et al.* Uncovering the mystery of opposite circadian rhythms between mouse and human leukocytes in humanized mice. *Blood* **130**, 1995–2005 (2017).
20. Keller, M. *et al.* A circadian clock in macrophages controls inflammatory immune responses. *PNAS* **106**, 21407–21412 (2009).
21. Pick, R., He, W., Chen, C.-S. & Scheiermann, C. Time-of-day-dependent trafficking and function of leukocyte subsets. *Trends Immunol.* **40**, 524–537 (2019).
22. Braun, R. *et al.* Universal method for robust detection of circadian state from gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E9247–E9256 (2018).
23. Lech, K. *et al.* Dissecting daily and circadian expression rhythms of clock-controlled genes in human blood. *J. Biol. Rhythms* **31**, 68–81 (2015).
24. Kusanagi, H. *et al.* Expression profiles of 10 circadian clock genes in human peripheral blood mononuclear cells. *Neurosci. Res.* **61**, 136–142 (2008).
25. Foo, J. C. *et al.* Longitudinal transcriptome-wide gene expression analysis of sleep deprivation treatment shows involvement of circadian genes and immune pathways. *Transl. Psychiatry*. https://doi.org/10.1038/s41398-019-0671-7 (2019).
26. Chang, J. *et al.* Circadian control of the secretory pathway maintains collagen homeostasis. *Nat. Cell Biol.* **22**, 74–86 (2020).
27. Pulford, K., Jones, M., Banham, A. H., Haralambieva, E. & Mason, D. Y. Lymphocyte-specific protein 1: a specific marker of human leucocytes. *Immunology* **96**, 262–271 (1999).
28. Lévi, F. *et al.* Implications of circadian clocks for the rhythmic delivery of cancer therapeutics. *Adv. Drug Deliv. Rev.* **59**, 1015–1035 (2007).
29. Hermida, R. C., Ayala, D. E., Chayán, L., Mojón, A. & Fernández, J. R. Administration-time-dependent effects of olmesartan on the ambulatory blood pressure of essential hypertension patients. *Chronobiol. Int.* **26**, 61–79 (2009).
30. Ramsey, M. R. & Ellisen, L. W. Circadian function in cancer: regulating the DNA damage response. *Proc. Natl. Acad. Sci.* **108**, 10379–10380 (2011).
31. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665–1254665 (2014).
32. Thomas, D. Gene-environment-wide association studies: emerging approaches. *Nat. Rev. Genet.* **11**, 259–272 (2010).
33. Matsa, E. *et al.* Transcriptome profiling of patient-specific human iPSC-cardiomyocytes predicts individual drug safety and efficacy responses in vitro. *Cell Stem Cell* **19**, 311–325 (2016).
34. Brodin, P. & Davis, M. M. Human immune system variation. *Nat. Rev. Immunol.* **17**, 21–29 (2016).
35. Whitney, A. R. *et al.* Individuality and variation in gene expression patterns in human blood. *Proc. Natl. Acad. Sci.* **100**, 1896–1901 (2003).
36. Long, J. E. *et al.* Morning vaccination enhances antibody response over afternoon vaccination: a cluster-randomised trial. *Vaccine* **34**, 2679–2685 (2016).
37. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
38. Chen, L. *et al.* Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414 (2016).
39. The Immune Cell Census https://www.immunecensus.org/ (2020). Accessed 03 Sept 2020.
40. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
41. Kanehisa, M. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

42. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2018).
43. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).
44. Farh, K.K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2014).
45. De Jager, P. L. *et al.* ImmVar project: insights and design considerations for future studies of "healthy" immune variation. *Semin. Immunol.* **27**, 51–57 (2015).
46. Sumitomo, S. *et al.* Transcriptome analysis of peripheral blood from patients with rheumatoid arthritis: a systematic review. *Inflamm. Regener.* https://doi.org/10.1186/s41232-018-0078-5 (2018).
47. Kobayashi, M., Wood, P. A. & Hrushesky, W. J. M. Circadian chemotherapy for gynecological and genitourinary cancers. *Chronobiol. Int.* **19**, 237–251 (2002).
48. Fairfax, B. P. & Knight, J. C. Genetics of gene expression in immunity to infection. *Curr. Opin. Immunol.* **30**, 63–71 (2014).

Author contributions

T.D., D.B. and M.T. designed the study. T.D., D.B. and J.P. performed blood extraction and single-cell experiments. T.D. and D.B. performed computational analysis. T.D., D.B. and M.T. wrote the manuscript. All authors reviewed the manuscript.

Funding

The funding was provided by Shurl and Kay Curci Foundation, the Beckman Institute at Caltech, and Heritage Medical Research Institute.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-77073-3.

Correspondence and requests for materials should be addressed to T.D. or M.T.