Comparing phenotypic manifolds with Kompot: Detecting differential abundance and gene expression at single-cell resolution

Dominik J. Otto[1,2,3], Erica Arriaga-Gomez[4], Elana Thieme[1,2,3,4,5], Ruijin Yang[1,2,3], Stanley C. Lee[4]+, Manu Setty[1,2,3]+

[1] Basic Sciences Division, Fred Hutchinson Cancer Center, Seattle WA

[2] Computational Biology Program, Public Health Sciences Division, Seattle WA

[3] Translational Data Science IRC, Fred Hutchinson Cancer Center, Seattle WA

[4] Translational Science and Therapeutics Division, Fred Hutchinson Cancer Center, Seattle WA

[5] Molecular and Cellular Biology Program, University of Washington, Seattle WA

+ Corresponding authors:

Manu Setty (msetty@fredhutch.org)

Stanley Lee (sclee3@fredhutch.org)

Abstract

Kompot is a statistical framework for holistic comparison of multi-condition single-cell datasets, supporting both differential abundance and differential expression. Differential abundance captures changes in how cells populate the phenotypic manifold across conditions, while differential expression identifies condition-specific changes in gene regulation that may be localized to particular regions of that manifold. Kompot models the distribution of cells and gene expression as continuous functions over a low-dimensional representation of cell states, enabling single-cell resolution inference with calibrated uncertainty estimates. Applying Kompot to aging murine bone marrow, we identified a continuum of shifts in hematopoietic stem cell and mature cell states, transcriptional remodeling of monocytes independent of compositional changes, and divergent regulation of oxidative stress response genes across cell types. By capturing both global and cell-state–specific effects of perturbation, Kompot reveals how aging reshapes cellular identity and regulatory programs across the hematopoietic landscape. This framework is broadly applicable to dissecting condition-specific effects in complex single-cell landscapes.

Introduction

Multi-condition single-cell datasets offer new opportunities to dissect how mutations or other perturbations affect both differentiation trajectories and disease phenotypes[1-4]. Systematic analyses of such datasets require characterization of two key aspects of condition-induced variability: differential abundance and differential expression.

Differential abundance refers to differences in the composition of cell states between two conditions (**Fig. 1A, B**). Even in the absence of overt phenotypic shifts such as loss or gain of broad cell types, changes in the frequency of specific cell states can provide early indicators of disease[5, 6] and response to treatment[7, 8]. These changes are often nuanced and cell-type specific, necessitating statistical approaches that rigorously assess significance while accounting for underlying biological variability.

Perturbations can also reshape the phenotypic landscape itself through differential expression, where gene expression changes that would be unlikely or impossible in unperturbed conditions can emerge[9] (**Fig. 1A, C**). These transcriptional responses reveal condition-specific molecular programs and regulatory networks that often manifest as subtle yet systematic changes across multiple states[9] and can frequently be masked by the inherent variability within biological systems .

Several computational approaches have been developed for comparative analysis of multi-condition single-cell data with differential abundance receiving greater attention[10-15]. A number of these methods rely on discretizations such as clustering cells into discrete groups to stabilize variability estimates. While facilitating interpretation, such discretizations lead to loss in resolution when changes are gradual or occur along continuous trajectories (**Supp. Fig. 1A**). Additionally, many existing methods lack principled significance measures, making it difficult to distinguish genuine biological signals from technical noise.

Here, we introduce Kompot, a novel computational framework that provides a continuous and statistically rigorous approach for differential abundance and differential expression testing using multi-condition single-cell data. Kompot leverages Gaussian Process (GP) modeling to estimate continuous functions across the cell-state space. This approach enables quantification of differential abundance and differential expression at single-cell resolution, while incorporating uncertainty estimates to provide statistically calibrated comparisons. By addressing both aspects of multi-condition analysis within a unified statistical framework, Kompot provides a comprehensive toolkit for extracting biological insights from increasingly complex single-cell datasets across a range of biological systems.

Results

Overview of the Kompot modeling framework

A single-cell RNA-seq dataset provides a snapshot of cellular states in a biological system, with the molecules measured representing a sample of the transcriptome of the cell. To characterize the system and compare it to similar systems under different conditions (such as mutations or other perturbations), we need to understand what this data tells us about the true composition of transcribed genes in a cell and the true distribution of cell states in the system. With Kompot we leverage high-dimensional, continuous cell-state representations and Bayesian Inference to learn both, the composition of the system and transcriptomes of each cell, enabling a holistic comparison of biological systems at single-cell resolution with uncertainty quantification. This comparison is divided into differential cell-state abundance that reflect changes in cell-state composition, and differential-expression that reflect changes in the transcriptome of cells in equivalent states.

The input to Kompot is a latent representation of co-embedded multi-condition single-cell data. This latent representation can incorporate batch effect correction[16, 17] to ensure that cell states considered equivalent share similar locations in that state space (**Supp. Fig. 1B**). From this foundation, Kompot employs

diffusion maps[18] to derive a cell-state representation with biologically meaningful structure (**Fig. 1A**). Diffusion maps are particularly effective for single-cell analysis as they capture the intrinsic geometry of the data while providing biologically meaningful measures of cell-to-cell distance[19, 20]. This distance metric supports both accurate local abundance estimation and principled information sharing when mapping from cell-state space to gene expression, enabling the statistical comparisons at the core of Kompot's differential testing framework.

Kompot differential abundance testing

To quantify differences in cell-state composition between conditions, Kompot first constructs condition-specific density functions on the diffusion map representation. For each condition, Kompot computes a separate continuous density function (**Fig. 1B, Supp. Fig. 2A-B**) using our Mellon framework[21]. Mellon leverages nearest-neighbor distances within a Gaussian Process model to estimate continuous density functions that capture the distribution of cells across the state space[21]. The continuous nature of these density estimates enables evaluation at any position within the cell-state space, including states predominantly observed in the alternate condition (**Supp. Fig. 2B**). By computing the ratio of density estimates at corresponding positions in the cell-state space, Kompot derives a measure of differential abundance that quantifies the fold-change in cell density between conditions at single-cell resolution (**Supp. Fig. 2C**).

While the ratio of density estimates provides a measure of fold-change, rigorous comparison between conditions requires statistical assessment of significance. Kompot addresses this challenge by computing uncertainty estimates for each density function. These uncertainty estimates account for two critical factors: limited data availability in regions of the cell-state space and inherent uncertainty through the sparse function representation. By incorporating these uncertainty estimates into the statistical framework, Kompot calculates posterior tail probabilities (PTPs), an alternative for p-values, for differential abundance at each cell state. This enables the generation of volcano plots for differential abundance testing where each point represents a single cell-state (**Fig. 1B**), allowing for threshold-based significance assessments. Thus, Kompot can identify subtle cell-state transitions and rare subpopulations that would likely be undetected by clustering-based approaches, providing insights into the continuous spectrum of cellular responses to perturbations (**Fig. 1D**).

Kompot differential expression testing

For differential expression testing between conditions, Kompot constructs condition-specific mappings from cell-states to gene expression space (**Fig. 1A, C**). These mappings are computed using Gaussian Processes (GPs)[22], by predicting log-transformed gene expression as a continuous function of the diffusion map coordinates. The non-linear relationship between diffusion maps and gene expression is effectively captured by GPs[22], essentially serving as a rigorous approach to gene expression imputation. The continuous nature of these mapping functions allows Kompot to estimate gene expression in one condition for cell states observed in another, producing counterfactual expression profiles that reflect how those same cell states would behave under different conditions (**Fig. 1C**). By computing the ratio between the actual expression in a cell and its counterfactual expression, Kompot quantifies gene expression-fold changes for every gene at single-cell resolution across the entire phenotypic landscape (**Fig. 1C**).

Given the large number of genes evaluated, we require a robust significance measure that allows us to identify the most relevant genes, accounting for both uncertainty in the gene expression function and statistical interdependence of imputed values. Kompot therefore employs the Mahalanobis distance, which measures the distance of a point from a multivariate normal distribution, considering both variance and covariance[23]. It effectively measures how many standard deviations a gene expression shift is away

from expected variation, while accounting for the correlation structure between neighboring cell states, making it an ideal significance measure for differential expression. Kompot therefore uses Mahalanobis distances as the score in a quasi-volcano plot (**Fig. 1C**) to identify significantly differentially expressed genes either in the full population of cells or in specific subsets (**Fig. 1D**).

Differential Abundance in Aging Bone Marrow

Aging is a compelling biological context for evaluating differential abundance methods, as it induces progressive and often subtle shifts in cellular composition rather than dramatic phenotypic transformations. Hematopoiesis has emerged as a particularly informative system for studying aging[24], characterized by declining stem cell function[25], reduced lymphocyte production[26], and increased susceptibility to myeloid malignancies[27]. To investigate these age-related changes, we generated a novel scRNA-seq dataset of murine bone marrow at 10, 63, and 103 weeks of age, representing young, middle-aged, and older timepoints (**Supp. Fig. 3A**). After quality control verification (**Supp. Fig. 3B, C**), we successfully identified most hematopoietic cell types across all ages (**Fig. 2A, B, Supp. Fig. 4**), providing us with a rich dataset to characterize cell-state changes and gene expression changes in aging hematopoiesis using Kompot.

We applied Kompot differential abundance testing on our aging hematopoiesis dataset to determine age-dependent composition changes (**Fig. 2C-E**). We first derived a unified cell-state representation for all cells using diffusion maps. We then calculated age-specific density functions using the subsets of cells from each timepoint (**Supp. Fig. 5A-B**). We evaluated these individual density functions at all cell-states in the dataset to derive condition-specific densities and associated uncertainties (**Supp. Fig. 5C-D**). This approach enabled pairwise comparisons that quantify age-related changes in cell-type composition (**Fig. 2C-D**). A key advantage of our framework is the ability to set thresholds based on both effect size and statistical significance at single-cell resolution, allowing for identification of cell states with significant differential abundance rather than relying solely on fold-change cutoffs (**Fig. 2C-E**).

Our analysis revealed significant expansions and contractions in distinct cell populations over time. Comparing young to old mice, we observed an expansion of the hematopoietic stem cell (HSC) compartment and a marked shift from naïve to memory cells in both B-cell and T-cell compartments (**Fig. 2F-G)**. The depletion of naïve B-cells and an enrichment of memory B-cells in older mice is a well-characterized phenomenon of hematopoietic aging[26]. This shift reflects fundamental changes in differentiation dynamics: naive B-cells rely on a continuous influx from hematopoietic progenitors, whereas memory B-cells accumulate over time due to antigen exposure. The observed decline in naïve B-cells reflects the reduced lymphoid differentiation potential of aged HSCs[28], while the expansion of memory B-cells, including age-associated B-cells, is driven by persistent immune activation[29]. Similarly, the T-cell compartment displayed a significant reduction in naïve T-cells with a concurrent increase in memory T-cells, a hallmark of immunosenescence that contributes to diminished adaptive immune responses in aging[30].

An increased accumulation of hematopoietic stem cells (HSCs) is the most significant change we observed in the progenitor cells (**Fig. 2F-G**). Progressive enrichment of HSCs with age is a well-documented phenomenon[31]. HSCs reside at the top of the hematopoietic hierarchy, maintaining blood production through self-renewal and differentiation into mature immune and blood cells. With aging, HSC function declines, leading to accumulation in the bone marrow while their ability to differentiate into different lineages becomes impaired[31]. While the accumulation of aged HSCs has been previously observed, it has traditionally been quantified using bulk measurements or clustering-based approaches. We took advantage of the single-cell resolution of Kompot to better characterize the age-induced change in HSCs. By comparing the differential abundance fold-change along pseudotime ordering[19], we observed that the expansion of HSC population is in fact a subtle shift in cell-state, rather than a global

168 increase in a predefined population (**Fig. 2H**). Specifically, HSCs in older mice progressively shift towards
169 a less differentiation-primed state, creating a continuum of altered stem cell states rather than a uniform
170 expansion (**Fig. 2H**). This subtle but consequential repositioning along the differentiation trajectory
171 provides a possible explanation for the functional decline in aged HSCs, revealing how changes in cell
172 abundance are accompanied by fundamental alterations in cellular function and developmental potential.

173 Comparing differential abundance results across ages, we observed that changes in mature cell
174 populations predominantly occurred between young and middle-aged mice (**Fig. 2D, Supp. Fig. 6A**),
175 while alterations in the HSC compartment were more pronounced between middle-aged and older mice
176 (**Fig. 2E, Supp. Fig. 6B**). This temporal separation suggests a sequential progression of aging, where
177 compositional shifts in differentiated cells precede fundamental changes in stem cell states. The early
178 alterations in mature immune cells may represent initial adaptive responses to aging, while the later HSC
179 changes potentially indicate more persistent reprogramming of the hematopoietic system.

180 Our results demonstrate the ability of Kompot to identify subtle but biologically meaningful abundance
181 changes between conditions. By combining a continuous representation of cell states with robust
182 statistical measures, Kompot enables precise detection of population-level changes while maintaining
183 single-cell resolution.

184

185 Differential Gene Expression in Aging Hematopoiesis

186 We applied Kompot differential expression testing on our mouse aging hematopoiesis dataset to
187 characterize age-related transcriptional changes (**Fig. 3**). We first inferred age-specific mapping functions
188 from diffusion map coordinates to log-normalized gene expression values, using the subset of cells from
189 each timepoint. We then used these mappings to impute gene expression and associated uncertainty at
190 all cell-states across the dataset. This enabled us to compute gene-level fold changes by comparing
191 observed expression values to their counterfactual predictions under alternative age conditions. To
192 assess significance, we calculated Mahalanobis distances for each gene across specified groups of cells,
193 quantifying deviations from expected variation while accounting for the covariate structure among
194 neighboring cell states. We excluded T and NK cells for this analysis since do not differentiate in the bone
195 marrow (**Fig. 3H**).

196 We first applied our approach to characterize gene expression changes in HSCs, a population known to
197 undergo functional and transcriptional changes with age[31], by comparing HSCs from middle-aged to old-
198 aged mice (**Fig. 3A-B**). Aging in HSCs has been extensively studied, and the Hematopoietic Aging Atlas
199 has synthesized results from multiple datasets to identify genes consistently up- or downregulated with
200 age[32]. For each gene, this atlas assigns a consistency score that reflects the reproducibility of age-
201 associated expression changes across studies[32]. Kompot's differential expression results strongly
202 prioritized genes with the highest consistency scores in the Hematopoietic Aging Atlas, with genes
203 showing large fold changes and Mahalanobis distances highly enriched for reproducible aging signatures
204 (**Fig. 3C**, **Supp. Fig. 7**).

205 Our differential abundance analysis revealed that HSCs in older mice shift towards a less differentiated
206 state along the cell-state continuum (**Fig. 2H**). We next asked whether the genes differentially expressed
207 in aged HSCs reflected this shift. Indeed, we found that gene expression fold changes were structured
208 along the same continuum: genes exhibited larger fold changes in cell states that were more differentially
209 abundant (**Fig. 3D**). These results highlight the transcriptional changes in HSCs that accompany and
210 may underlie the shift toward a less differentiated state with age,

211 We next investigated age-related gene expression changes in mature cell types. We focused on changes
212 from young to old mice since the abundance changes in mature cell types were most pronounced in this
213 comparison (**Fig. 2D-F**). To distinguish expression changes from abundance-driven effects, we restricted

our analysis to cell types that *did not show significant compositional changes*. Monocytes exhibited clear transcriptional changes despite stable abundance (**Fig. 3E-F)**. Gene ontology analysis revealed that downregulated genes were enriched in core monocyte/macrophage functions such as defense response (**Fig. 3G**), consistent with their reduced functional capacity with age[33]. In contrast, upregulated genes were enriched for antigen processing and presentation pathways, driven by increased expression of MHC class II genes (**Fig. 3G**), which have been implicated in age-related immune remodeling[34]. Notably, our results recapitulate findings from a previous study that identified H2-Aa, H2-Ab1, H2-Eb1, Cd74, and Aw112010 as the top dysregulated genes in aging monocytes across both mice and humans, with only Igkc surpassing H2-Eb1 in our ranking[34]. The near-perfect agreement with experimentally validated gene sets underscores the power of our differential expression testing approach to detect biologically meaningful transcriptional shifts. Gene expression differences in other mature cell types were not as pronounced but did show a tendency for upregulation of MHC Class II genes similar to monocytes (**Supp. Fig. 8**).

Our use of Mahalanobis distance as a significance measure enables Kompot to identify genes that exhibit coordinated but opposing changes across subpopulations. For example, consider a gene that is upregulated in one cell type and downregulated in another. While the net fold change may be close to zero, the Mahalanobis distance remains high, reflecting a statistically significant deviation from expected variation if unaffected. This occurs because such a pattern is unlikely under the local covariance structure of the cell-state space, where neighboring states are typically expected to exhibit similar expression patterns. As a result, Kompot can detect structured differential expression that would be missed by conventional fold-change-based approaches.

To test this capability, we applied Kompot differential expression testing between young and old cells across all cell-types that differentiate in the bone marrow (**Fig. 3H-I**). The top genes with both high positive fold change and high Mahalanobis distance were strongly enriched for MHC class II genes, consistent with our monocyte-specific results and indicating that MHC class II upregulation is a general feature of hematopoietic aging (**Fig. 3J**). Likewise, genes with high Mahalanobis distance and strong negative fold change were consistently downregulated across multiple cell types (**Fig. 3J**). In contrast, genes ranked highly by Mahalanobis distance but with low absolute fold change exhibited divergent expression patterns across cell types, with upregulation in some and downregulation in others (**Fig. 3J**). To illustrate these distinctions, we visualized expression levels of representative genes from each category across the cell-state manifold (**Fig. 3K**). Genes selected by fold change showed broad, monotonic shifts in expression, whereas Mahalanobis-selected genes revealed distinct, cell-type-specific regulation (**Fig. 3K**).

We next performed gene ontology analysis to assess the biological relevance of genes with high Mahalanobis distance but low absolute fold change. Antioxidant activity emerged as the most significantly enriched term (**Fig. 3L**), a pathway known to be impacted by aging[35]. Notably, our results indicate that while antioxidant activity is broadly impacted across hematopoietic cell types, the direction of expression changes of the pathway genes varies across genes and cell types, suggesting a context-specific adaptation to aging (**Fig. 3M**). Oxidative stress is a known driver of HSC dysfunction, promoting DNA damage, apoptosis, and biased differentiation toward myeloid lineages[35]. The differential expression of antioxidant pathway genes observed in our study likely represents a compensatory mechanism to mitigate reactive oxygen species (ROS) accumulation in aged HSCs[35].

These results demonstrate that Mahalanobis distance captures biologically meaningful heterogeneity in gene expression and provides a powerful approach for aggregating structured expression differences across cell states without requiring prior clustering, making it especially suited for complex, continuous single-cell landscapes.

Discussion

Kompot provides a flexible statistical framework for quantifying condition-specific effects in single-cell data, modeling both cell abundance and gene expression as continuous functions over the cell-state manifold. By estimating condition-specific densities and expression profiles using Gaussian processes, Kompot performs differential analysis at single-cell resolution while accounting for uncertainty and spatial structure in the data. This approach supports biologically grounded interpretations of perturbation effects, whether they manifest as global shifts in population structure, localized transcriptional responses, or context-specific coordination between abundance and expression changes.

Applying Kompot to aging murine bone marrow revealed both well-established and previously unrecognized features of hematopoietic aging. Kompot recapitulated known expansions in memory lymphocytes and age-associated B cells, while resolving more subtle shifts in hematopoietic stem cell states that do not correspond to discrete population changes. In the transcriptomic landscape, Kompot uncovered gene expression programs that are either conserved or cell-type specific, including divergent regulation of oxidative stress pathways. These findings demonstrate how modeling continuous phenotypic structure supports more nuanced detection of biological change. Beyond aging, Kompot is broadly applicable to studies of development, disease progression, and response to perturbation, where condition-induced effects often unfold along continuous, high-dimensional trajectories.

Methods

Kompot framework

Kompot is a tool for comparing single-cell phenotypic manifolds across conditions. It provides frameworks for both differential abundance testing and differential gene expression testing at single-cell resolution using multi-condition datasets. Differential abundance testing enables the detection of abundance shifts of cell-states across the phenotypic manifold, without requiring predefined cell-type boundaries. This is achieved by estimating condition-specific cell-state densities using the Bayesian inference scheme implemented in Mellon. In parallel, through differential gene expression testing, Kompot identifies gene expression changes across the same manifold, capturing subtle but structured transcriptional shifts that may vary across cell states and may not follow consistent directionality To do so, Kompot leverages the Gaussian process inference implemented in Mellon to estimate condition-specific gene expression functions for thousands of genes within minutes, while explicitly sharing information across neighboring cell states to address sparsity and enable statistically meaningful comparisons

Differential Abundance

This section explains the methods of Kompot's differential abundance (DA) analysis framework. Our density quantification relies on the Gaussian Process (**GP**) estimation implemented in Mellon. This estimation process is a Bayesian inference that provides us with both a "best estimate" function that maps each point in the cell-state space to a density value $\mu(x_i)$, and a quantification of uncertainty $\sigma(x_i)$. Since we infer these quantities independently for each condition, the posterior distribution for the density change from condition a to condition b at any cell state x_i is given by

$$\delta(x_i) \sim \mathcal{N}\left(\mu_b(x_i) - \mu_a(x_i), \sigma_a^2(x_i) + \sigma_b^2(x_i)\right).$$

Therefore, our density log-fold change is defined by $\mu_b(x_i) - \mu_a(x_i)$, and our uncertainty quantification by $\sigma_a^2(x_i) + \sigma_b^2(x_i)$. Notably, using the correct intrinsic dimensionality, Mellon reports density in the unit log number of cells per volume.

To make density values comparable across samples with different cell counts, we can transform this unit into log fraction of cells per volume by subtracting the log total number of cells. Therefore, we normalize density in Kompot using

$$\mu'_a(x_i) = \mu_a(x_i) - \log N_a$$

where N_a represents the total number of cells measured for condition a. For this normalization to be valid, it is essential, that we assume the correct intrinsic dimensionality of the phenotypic manifold. To that end we use Mellon dimensionality estimation that is assumed to be constant across the dataset.

Finally, we can define the Kompot's local density log-fold change statistic as

$$\Delta(x_i) = \mu_b(x_i) - \mu_a(x_i) - \log {}^{N_a}/_{N_b}\,.$$

Additionally, we provide a p-value-like significance value through the posterior tail probability (**PTP**). For example, if the density change $\Delta(x_i)$ is positive, then we compute $\mathbb{P}(\delta(x_i) < 0)$, and $\mathbb{P}(\delta(x_i) > 0)$ when $\Delta(x_i)$ is negative. The PTP quantifies the probability that the true fold change goes into the opposite direction of our best estimate, given our data and modeling assumptions. Since our posterior distribution is a normal distribution, the local Kompot PTP is given by:

$$PTP(x_i) = \Phi\left(\frac{-|\Delta(x_i)|}{\sqrt{\sigma_a^2(x_i) + \sigma_b^2(x_i)}}\right)$$

were Φ is the cumulative distribution function of the standard normal distribution.

Differential Expression

Similar to the density differences, we also leverage Mellon's GP implementation (**Supp. Note 1**) for differential expression (DE) analysis. In contrast to the density inference algorithm, this DE method directly conditions the gene-expression functions on the measured data, as classically done in GPs. For each condition, the goal is to compute a gene-expression function that predicts an expression value for each gene at any location in the cell-state space, specific to that condition. These functions provide a smooth and continuous representation of gene expression across the manifold and allow for quantifying differential expression between conditions at any point in state space, effectively integrating all data points without requiring predefined cluster boundaries.

The resulting gene-expression functions and uncertainty estimation can be used equivalently to the cell-state density functions to compute local gene-expression differences, featuring local fold change and PTP. However, apart from the question in which cell-state a gene may be differentially expressed, we want to answer if a gene asserts any significant DE within a cell population or across the entire cell-state space, effectively aggregating measurements across multiple cells.

This comes with an important caveat: The GP based gene-expression functions are an imputation that already accumulates measurements of multiple cells, and the resulting gene-expression values $f(x_i)$ at different cell states x_i are not independent measurements that can be aggregated directly. This caveat can be addressed very directly through Mellon's uncertainty quantification: It covers not only the variance for a function value $f(x_i)$ but also the covariance to other function values $f(x_j)$ based on the similarity of x_i and x_j. This covariance can be understood as a statistical connection between the expressions of two cell states that comes from our assumption that genes change expression smoothly throughout cell-state space. Formally, under our modelling assumption, the true gene expression values $f(x)$ at cell states $x = \{x_i\}_{i=1}^m$ are distributed according to a posterior

$$f(x) \sim \mathcal{N}\big(\mu(x), \Sigma(\mathrm{x})\big).$$

Therefore, the significance for an expressional change between two conditions k and l can be quantified through the Mahalanobis distance (**Supp. Note 2**):

$$D(a,b) \quad = \sqrt{(\mu_a - \mu_b)^T\,(\Sigma_a + \Sigma_b)^{-1}\,(\mu_a - \mu_b)}$$

Notably, this adequately accounts for redundancy, as adding a nearly identical cell state contribute nearly nothing to the value of this statistic, while more distinct cell-states have a larger contribution. We leverage this to our advantage. By producing a fixed set of landmarks, using k-means clustering, we limit the evaluation of the gene-expression function and the size of the covariance matrix to a fixed size. By using a fixed number of steps and k++ initialization, both the clustering and the GP only grow linearly in run-time with the number of cells. Notably, k-means grows with $O(t \times k \times d \times n)$ where t is the number of iterations, k the number of landmarks, d the dimensionality of the space, and n the number of cells.

Data analysis

Detailed descriptions of data generation and analysis described in the manuscript will be included in a subsequent version of the manuscript.

Data Availability

Raw single-cell data for the murine aging hematopoiesis will be deposited to GEO. Processed anndata object with annotations is available to download from Zenodo[36].

Code Availability

Kompot is available as a Python module at https://github.com/settylab/Kompot. Jupyter notebooks detailing the usage of Kompot differential expression and differential abundance testing are available at https://github.com/settylab/kompot/tree/main/examples and documentation at https://kompot.readthedocs.io/en/latest/.

Acknowledgements

This study was supported by National Institute of General Medical Studies grant no. R35 GM147125 to M.S.; National Cancer Institute grant no. R01CA292932, Mark Foundation for Cancer Research Endeavor Award and Edward P. Evans Foundation Discovery Research Grant to S.C.L; Brotman Baty Institute Pilot Award and Translational Data Science IRC New Collaboration Award to M.S. and S.C.L; National Institute of General Medicine Studies grant no. T32GM136534 to E.T.; National Institutes of Health grant no. ORIP S10OD028685 to support high-performance computing at the Fred Hutchinson Cancer Center.

Competing Interests

The authors declare no competing interests.

Figures

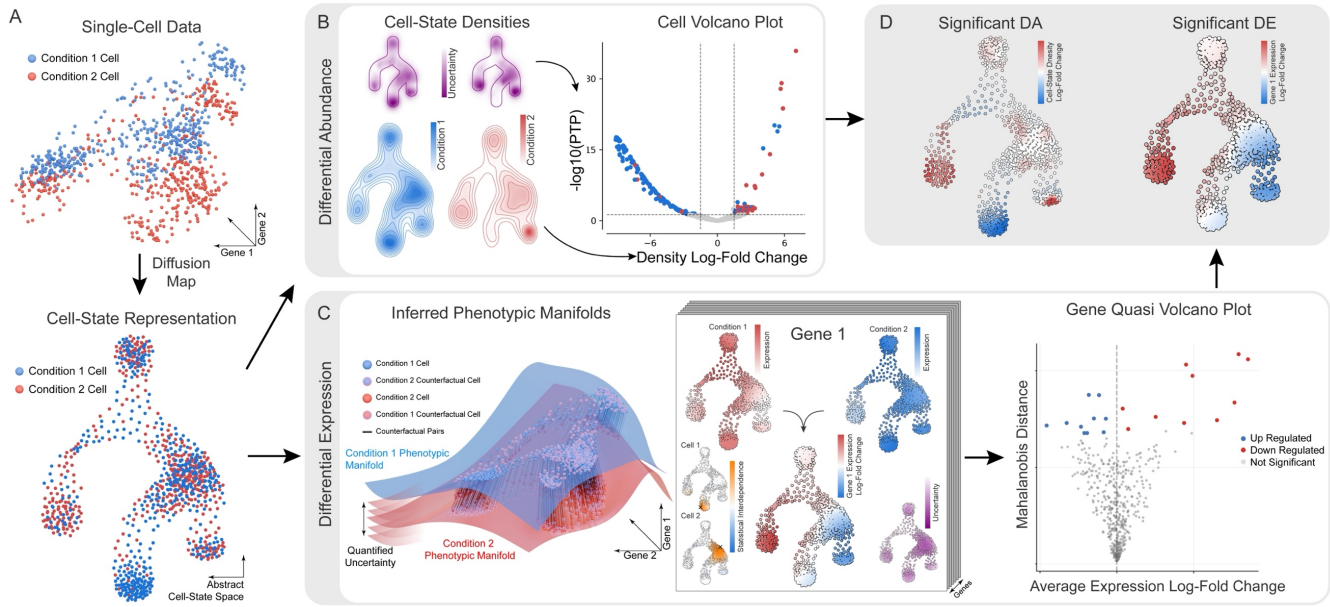


Figure 1: Overview of Kompot's framework for differential abundance and expression analysis.

(A) Illustration of high-dimensional single-cell transcriptomic data from two conditions, projected into a shared low-dimensional cell-state space using diffusion maps.

(B) Differential abundance analysis in Kompot. Condition-specific cell-state densities with associated uncertainty estimates are computed and compared to yield density log-fold changes. Significance is quantified using posterior tail probability (PTP), visualized in a quasi-volcano plot.

(C) Differential expression analysis. Left: schematic of inferred phenotypic manifolds for both conditions with counterfactual cell pairs illustrating expression comparison. Middle: gene-specific differences are illustrated for one example gene, showing condition-specific expression levels, statistical interdependence, and uncertainty. Right: A quasi-volcano plot summarizes gene-level results with the average log-fold change on the x-axis and Mahalanobis distance (accounting for variance and covariance) as a multivariate significance score on the y-axis.

(D) Visualization of differential abundance (left) and differential expression (right) across the cell-state space, colored by log-fold change magnitude.

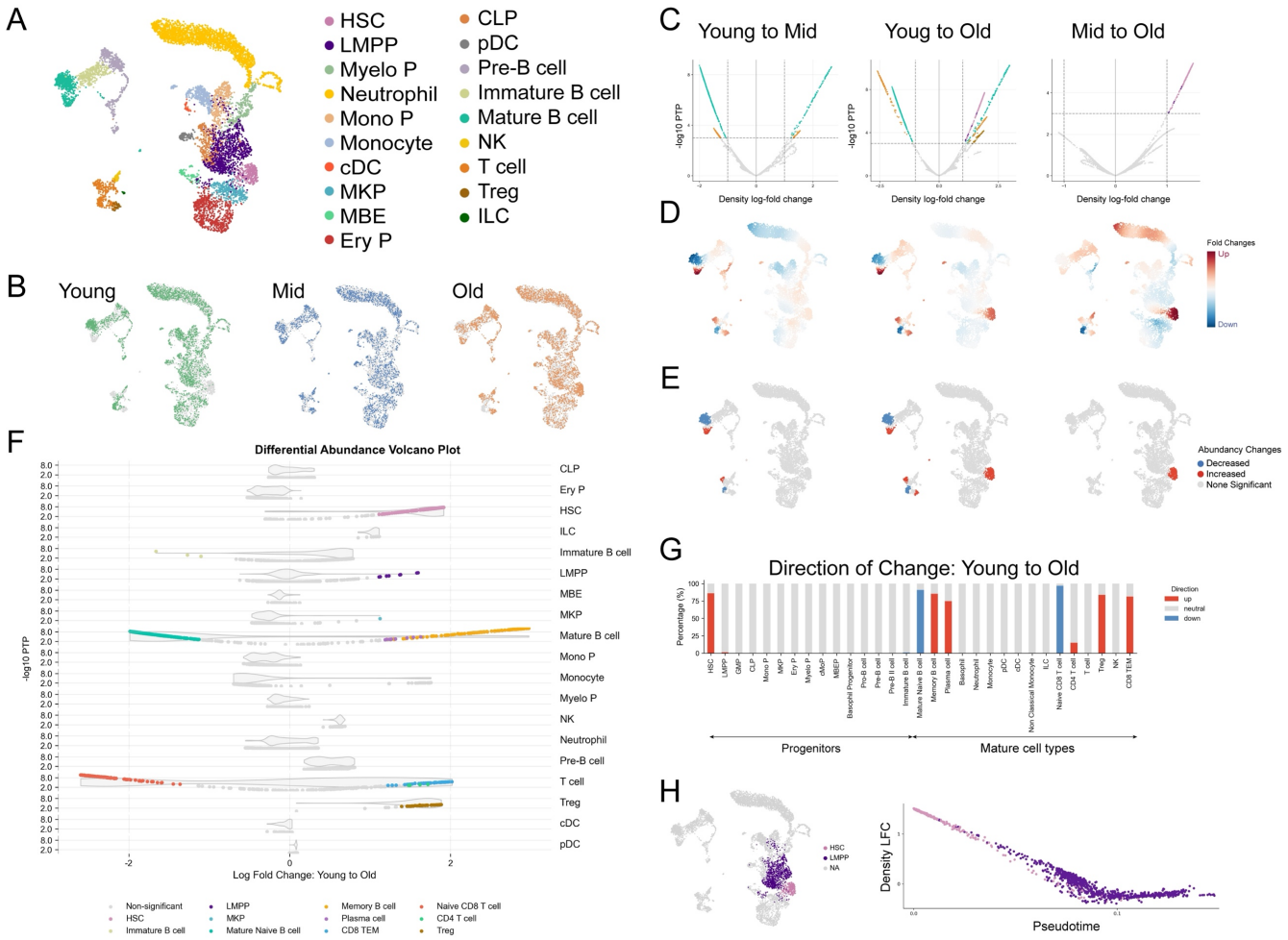


Figure 2: Kompot differential abundance analysis in aging hematopoiesis.

(A) UMAP projection of aging murine bone marrow scRNA-seq dataset, colored by medium-resolution cell types.

(B) UMAPs highlighting cells from mice grouped by age: Young (10 weeks), Mid (63 weeks), and Old (103 weeks).

(C) Differential abundance results for the Young to Mid (left), Young to Old (middle) and Mid to Old (right) comparisons. Quasi-cell volcano plot with density log-fold change on the x-axis and Posterior Tail Probabilities (PTP) on the y-axis. Significantly different cell-states (absolute log fold change > 1, PTP < 10^{-3}) are highlighted by their cell type annotations from (A).

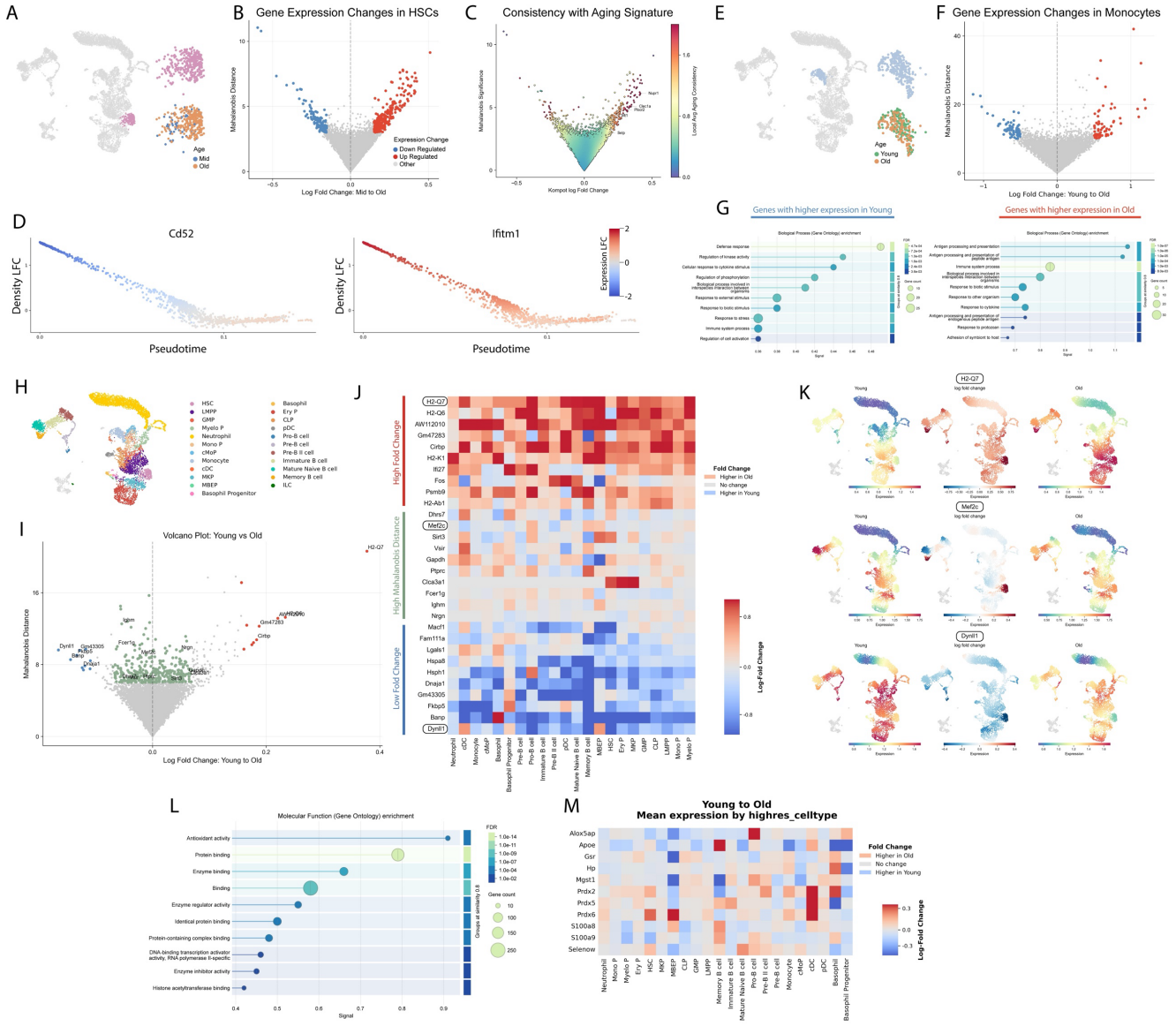
(D) UMAPs displaying the estimated local density log-fold changes across the full joint cell-state space derived from all samples (Young, Mid, and Old) for comparisons in C.

(E) UMAPs highlighting the subset of cell states with statistically significant abundance changes, as defined by the thresholds shown in the volcano plots in C.

(F) Differential abundance volcano plots showing density log-fold changes across cells, stratified by major cell type and colored by subtype in the Young to Old comparison. y-axis indicates posterior tail probability. Violin plots in the background indicate abundance of cells at the respective density log-fold change in this group.

(G) Direction and magnitude of abundance changes per cell type in the Young to Old comparison. Bars show the fraction of significantly changing cell-states within each annotated cell type, color-coded by the direction of change. Cell types are ordered along a progenitor-to-mature axis.

(H) Kompot results for early hematopoietic progenitors. Left: UMAP highlighting HSC, and LMPP subsets. Right: Scatterplot of cells colored by density log-fold change along pseudotime (x-axis), showing dynamic abundance shifts during early differentiation.



429

430 Figure 3: Kompot differential expression analysis in aging hematopoiesis.

431 (A) UMAP from Fig. 2A, with HSCs highlighted.

432 (B) Differential expression results for Mid to Old HSCs. Quasi-cell volcano plot with expression log-fold
433 change on the x-axis and Mahalanobis distance on the y-axis. Significantly different genes (absolute log
434 fold change > 0.15, Mahalanobis distance > 3) are highlighted.

435 (C) like (B) colored by weighted average of reported “Aging Consistency” using standard Gaussian Kernel
436 after scaling x- and y-axis to standard deviation 1.

437 (D) Same plot as Fig. 2H, colored by Mid to Old HSC gene expression log fold change for a gene
438 downregulated with age (Cd52, left) and an upregulated gene (Ifitm1, right).

439 (E) UMAP from Fig. 2A, with monocytes highlighted.

440 (F) Same as (B), for Young to Old monocytes.

441 (G) Gene ontology analysis from STRING[37] for downregulated genes (left) and upregulated genes (right)
442 from Fig. 2F.

443 (H) UMAP for Fig. 2A, highlighting all cell types except T and NK cells

(I) Differential expression results for all cells in Fig. 2H for Young to Old comparisons. Top 10 genes with positive fold change and high Mahalanobis distance are shown in red, bottom 10 genes with negative fold change and high Mahalanobis distance are shown in blue, and genes with log absolute fold change but high Mahalanobis distance are shown in green (absolute log fold change < 0.07, Mahalanobis distance > 6). 20 genes are labeled.

(J) Heatmap showing Young to Old gene expression log fold changes across different cell types for genes highlighted in (I).

(K) Top: Imputed expression of gene H2-Q7 using young cells (left), using old cells (right) and fold change from young to old (middle). Middle, Bottom: Same as the top row for Mef2c and Dynll1. These genes serve as examples of genes with high positive fold change, high Mahalanobis but low absolute fold change, and high negative fold change respectively.

(L) Gene ontology analysis using genes with high Mahalanobis distance and low absolute fold change from (I).

(M) Heatmap showing Young to Old gene expression log fold changes across different cell types for genes involved in antioxidant activity.

References

1. Tabula Muris C. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature. 2020;583(7817):590-5, PMC8240505.

2. Konturek-Ciesla A, Dhapola P, Zhang Q, Sawen P, Wan H, Karlsson G, et al. Temporal multimodal single-cell profiling of native hematopoiesis illuminates altered differentiation trajectories with age. Cell Rep. 2023;42(4):112304.

3. Herault L, Poplineau M, Mazuel A, Platet N, Remy E, Duprez E. Single-cell RNA-seq reveals a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells. BMC Biol. 2021;19(1):19, PMC7851934.

4. Kucinski I, Campos J, Barile M, Severi F, Bohin N, Moreira PN, et al. A time- and single-cell-resolved model of murine bone marrow hematopoiesis. Cell Stem Cell. 2023.

5. Gazestani V, Kamath T, Nadaf NM, Dougalis A, Burris SJ, Rooney B, et al. Early Alzheimer's disease pathology in human cortex involves transient cell states. Cell. 2023;186(20):4438-53 e23, PMC11107481.

6. Mayr CH, Simon LM, Leuschner G, Ansari M, Schniering J, Geyer PE, et al. Integrative analysis of cell state changes in lung fibrosis with peripheral protein biomarkers. EMBO Mol Med. 2021;13(4):e12871, PMC8033531.

7. Rehman SK, Haynes J, Collignon E, Brown KR, Wang Y, Nixon AML, et al. Colorectal Cancer Cells Enter a Diapause-like DTP State to Survive Chemotherapy. Cell. 2021;184(1):226-42 e21, PMC8437243.

8. van der Leun AM, Thommen DS, Schumacher TN. CD8(+) T cell states in human cancer: insights from single-cell analysis. Nat Rev Cancer. 2020;20(4):218-32, PMC7115982.

9. Replogle JM, Saunders RA, Pogson AN, Hussmann JA, Lenail A, Guna A, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell. 2022;185(14):2559-75 e28, PMC9380471.

10. Burkhardt DB, Stanley JS, 3rd, Tong A, Perdigoto AL, Gigante SA, Herold KC, et al. Quantifying the effect of experimental perturbations at single-cell resolution. Nat Biotechnol. 2021;39(5):619-29, PMC8122059.

11. Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat Biotechnol. 2022;40(2):245-53.

12. Yi H, Plotkin A, Stanley N. Benchmarking differential abundance methods for finding condition-specific prototypical cells in multi-sample single-cell datasets. Genome Biology. 2024;25(1):9.

13. Missarova A, Dann E, Rosen L, Satija R, Marioni J. Leveraging neighborhood representations of single-cell data to achieve sensitive DE testing with miloDE. Genome Biol. 2024;25(1):189, PMC11256449.

14. Boyeau P, Hong J, Gayoso A, Kim M, McFaline-Figueroa JL, Jordan MI, et al. Deep generative modeling of sample-level heterogeneity in single-cell genomics. bioRxiv. 2024:2022.10.04.510898.

15. Ahlmann-Eltze C, Huber W. Analysis of multi-condition single-cell data with latent embedding multivariate regression. Nat Genet. 2025;57(3):659-67, PMC11906359.

16. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289-96, PMC6884693.

17. Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41-50, PMC8748196.

18. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci U S A. 2005;102(21):7426-31, PMC1140422.

19. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D. Characterization of cell fate probabilities in single-cell data with Palantir. Nat Biotechnol. 2019;37(4):451-60.

503 20. Haghverdi L, Buttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs
504 lineage branching. Nat Methods. 2016;13(10):845-8.

505 21. Otto D, Jordan C, Dury B, Dien C, Setty M. Quantifying Cell-State Densities in Single-Cell Phenotypic
506 Landscapes using Mellon. Nature Methods. 2024, PMC10369887.

507 22. Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning2006.

508 23. PC M. Reprint of: Mahalanobis, P.C. (1936) "On the Generalised Distance in Statistics.". Sankhya A.
509 2018;80(1):1-7.

510 24. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. Science. 2019;366(6465),
511 PMC8050831.

512 25. Chambers SM, Shaw CA, Gatza C, Fisk CJ, Donehower LA, Goodell MA. Aging hematopoietic stem cells
513 decline in function and exhibit epigenetic dysregulation. PLoS Biol. 2007;5(8):e201, PMC1925137.

514 26. Li H, Cote P, Kuoch M, Ezike J, Frenis K, Afanassiev A, et al. The dynamics of hematopoiesis over the
515 human lifespan. Nat Methods. 2025;22(2):422-34, PMC11908799.

516 27. Appelbaum FR, Gundacker H, Head DR, Slovak ML, Willman CL, Godwin JE, et al. Age and acute
517 myeloid leukemia. Blood. 2006;107(9):3481-5, PMC1895766.

518 28. Florian MC, Klose M, Sacma M, Jablanovic J, Knudson L, Nattamai KJ, et al. Aging alters the epigenetic
519 asymmetry of HSC division. PLoS Biol. 2018;16(9):e2003389, PMC6168157.

520 29. Bulati M, Caruso C, Colonna-Romano G. From lymphopoiesis to plasma cells differentiation, the age-
521 related modifications of B cell compartment are influenced by "inflamm-ageing". Ageing Res Rev. 2017;36:125-
522 36.

523 30. Goronzy JJ, Weyand CM. Immune aging and autoimmunity. Cell Mol Life Sci. 2012;69(10):1615-23,
524 PMC4277694.

525 31. de Haan G, Lazare SS. Aging of hematopoietic stem cells. Blood. 2018;131(5):479-87.

526 32. Flohr Svendsen A, Yang D, Kim K, Lazare S, Skinder N, Zwart E, et al. A comprehensive transcriptome
527 signature of murine hematopoietic stem cell aging. Blood. 2021;138(6):439-51.

528 33. De Maeyer RPH, Chambers ES. The impact of ageing on monocytes and macrophages. Immunol Lett.
529 2021;230:1-10.

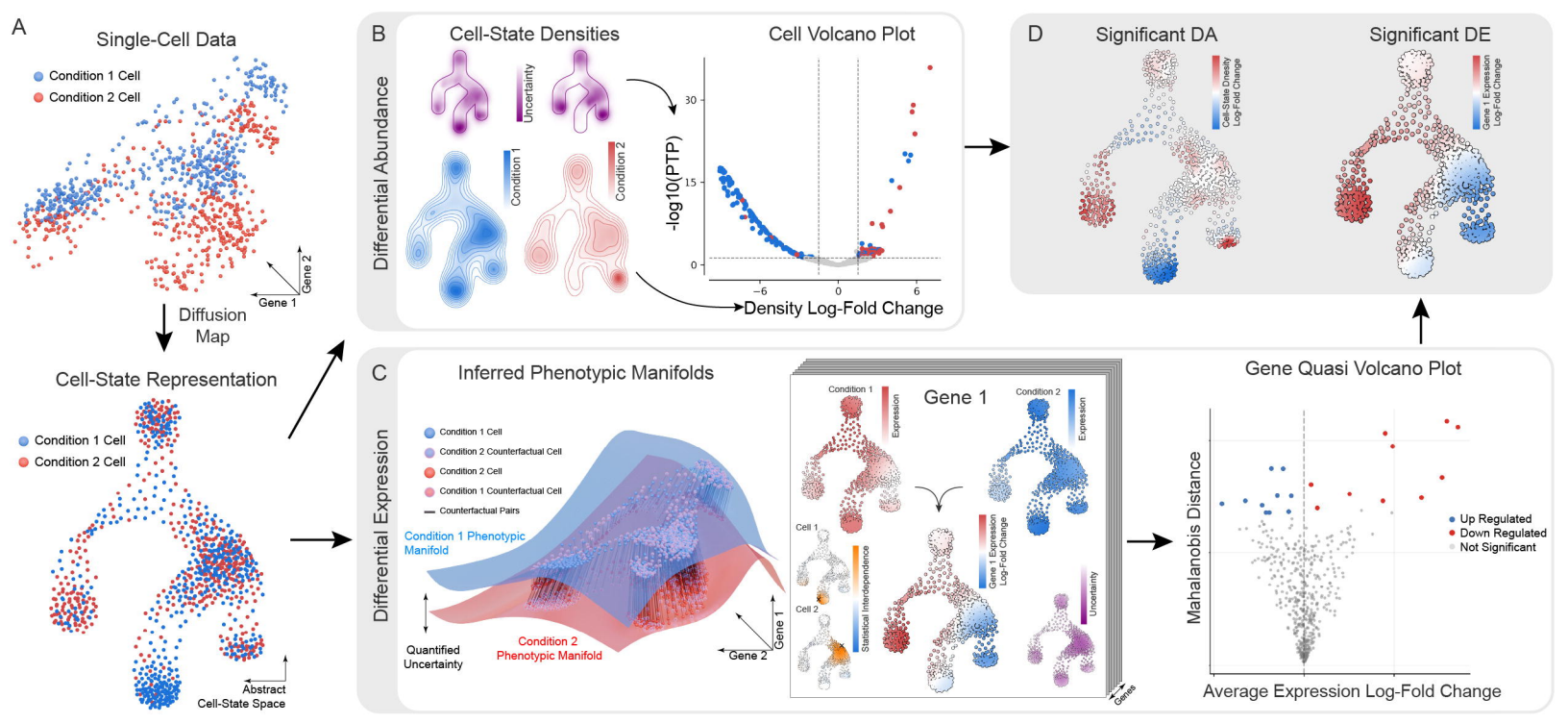
530 34. Barman PK, Shin JE, Lewis SA, Kang S, Wu D, Wang Y, et al. Production of MHCII-expressing classical
531 monocytes increases during aging in mice and humans. Aging Cell. 2022;21(10):e13701, PMC9577948.

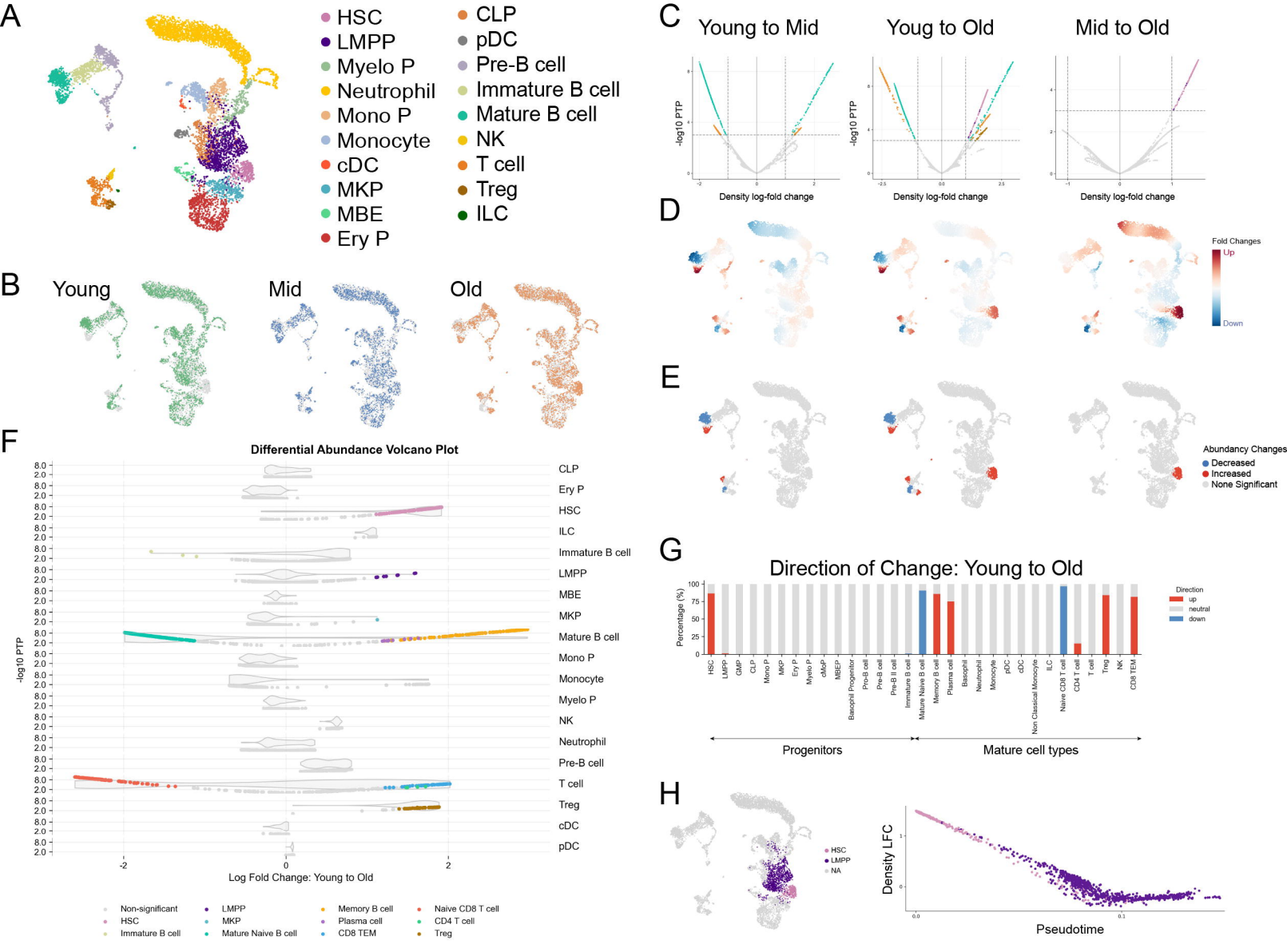
532 35. Shao L, Li H, Pazhanisamy SK, Meng A, Wang Y, Zhou D. Reactive oxygen species and hematopoietic
533 stem cell senescence. Int J Hematol. 2011;94(1):24-32, PMC3390185.

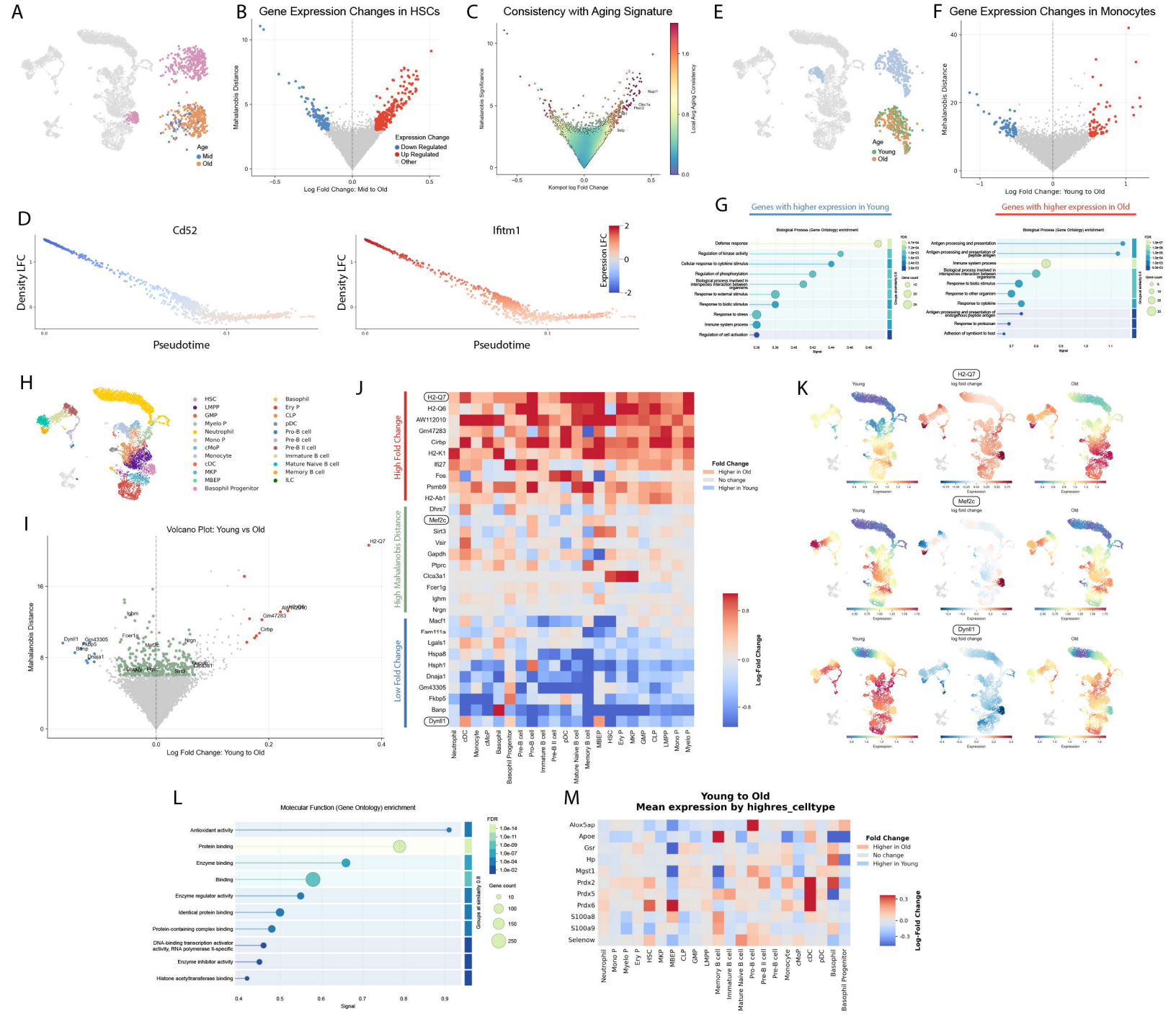
534 36. Otto D, Arriaga-Gomez E, Thieme E, Yang R, Lee S, Setty M. Single-cell CITE-seq of murine bone
535 marrow across aging (Young, Mid, Old): Zenodo; 2025 [Available from: https://doi.org/10.5281/zenodo.15587767.

536 37. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein
537 association networks with increased coverage, supporting functional discovery in genome-wide experimental
538 datasets. Nucleic Acids Res. 2019;47(D1):D607-D13, PMC6323986.

539


A
Single-Cell Data
Condition 1 Cell
Condition 2 Cell
Gene 1
Gene 2
Diffusion Map
Cell-State Representation
Condition 1 Cell
Condition 2 Cell
Abstract Cell-State Space
B
Differential Abundance
Cell-State Densities
Uncertainty
Condition 1
Condition 2
Cell Volcano Plot
-log10(PTP)
30
15
0
−6
0
6
Density Log-Fold Change
D
Significant DA
Cell-State Density Log-Fold Change
Significant DE
Gene 1 Expression Log-Fold Change
C
Differential Expression
Inferred Phenotypic Manifolds
Condition 1 Cell
Condition 2 Counterfactual Cell
Condition 2 Cell
Condition 1 Counterfactual Cell
Counterfactual Pairs
Condition 1 Phenotypic Manifold
Condition 2 Phenotypic Manifold
Quantified Uncertainty
Gene 1
Gene 2
Gene 1
Condition 1
Expression
Condition 2
Expression
Cell 1
Cell 2
Statistical Interdependence
Gene 1 Expression Log-Fold Change
Uncertainty
Genes
Gene Quasi Volcano Plot
Mahalanobis Distance
Up Regulated
Down Regulated
Not Significant
Average Expression Log-Fold Change


A
HSC
LMPP
Myelo P
Neutrophil
Mono P
Monocyte
cDC
MKP
MBE
Ery P
CLP
pDC
Pre-B cell
Immature B cell
Mature B cell
NK
T cell
Treg
ILC
B
Young
Mid
Old
C
Young to Mid
Youg to Old
Mid to Old
-log10 PTP
Density log-fold change
D
Fold Changes
Up
Down
E
Abundancy Changes
Decreased
Increased
None Significant
F
Differential Abundance Volcano Plot
-log10 PTP
Log Fold Change: Young to Old
Non-significant
HSC
Immature B cell
LMPP
MKP
Mature Naive B cell
Memory B cell
Plasma cell
CD8 TEM
Naive CD8 T cell
CD4 T cell
Treg
G
Direction of Change: Young to Old
Percentage (%)
Direction
up
neutral
down
Progenitors
Mature cell types
H
HSC
LMPP
NA
Density LFC
Pseudotime


A
Age
Mid
Old
B
Gene Expression Changes in HSCs
Mahalanobis Distance
Log Fold Change: Mid to Old
Expression Change
Down Regulated
Up Regulated
Other
C
Consistency with Aging Signature
Nahalanobis Significance
Kompot log Fold Change
Local Avg Aging Consistency
Nqr1
Clec1a
Ptcr2
Selp
E
Age
Young
Old
F
Gene Expression Changes in Monocytes
Mahalanobis Distance
Log Fold Change: Young to Old
D
Cd52
Density LFC
Pseudotime
Ifitm1
Expression LFC
G
Genes with higher expression in Young
Genes with higher expression in Old
Biological Process (Gene Ontology) enrichment
Defense response
Regulation of kinase activity
Cellular response to cytokine stimulus
Regulation of phosphorylation
Biological process involved in interspecies interaction between organisms
Response to external stimulus
Response to biotic stimulus
Response to stress
Immune system process
Regulation of cell activation
Antigen processing and presentation
Antigen processing and presentation of peptide antigen
Immune system process
Biological process involved in interspecies interaction between organisms
Response to biotic stimulus
Response to other organism
Response to cytokine
Antigen processing and presentation of endogenous peptide antigen
Response to protozoan
Adhesion of symbiont to host
Signal
H
HSC
LMPP
GMP
Myelo P
Neutrophil
Mono P
cMoP
Monocyte
cDC
MKP
MBEP
Basophil Progenitor
Basophil
Ery P
CLP
pDC
Pro-B cell
Pre-B cell
Pre-B II cell
Immature B cell
Mature Naive B cell
Memory B cell
ILC
I
Volcano Plot: Young vs Old
Mahalanobis Distance
Log Fold Change: Young to Old
H2-Q7
Gm47283
Cirbp
Igbm
Fcer1g
Mef2c
Nrgn
Dynll1
Gm43305
Fkbp5
Banp
Dnaja1
Ptprc
Sirt3
J
High Fold Change
High Mahalanobis Distance
Low Fold Change
H2-Q7
H2-Q6
AW112010
Gm47283
Cirbp
H2-K1
Ifi27
Fos
Psmb9
H2-Ab1
Dhrs7
Mef2c
Sirt3
Vsir
Gapdh
Ptprc
Clca3a1
Fcer1g
Igbm
Nrgn
Macf1
Fam111a
Lgals1
Hspa8
Hsph1
Dnaja1
Gm43305
Fkbp5
Banp
Dynll1
Fold Change
Higher in Old
No change
Higher in Young
Log-Fold Change
Neutrophil
cDC
Monocyte
cMoP
Basophil
Basophil Progenitor
Pre-B cell
Pro-B cell
Immature B cell
Pre-B II cell
pDC
Mature Naive B cell
Memory B cell
MBEP
HSC
Ery P
MKP
GMP
CLP
LMPP
Mono P
Myelo P
K
H2-Q7
Mef2c
Dynll1
Young
Old
log fold change
Expression
L
Molecular Function (Gene Ontology) enrichment
Antioxidant activity
Protein binding
Enzyme binding
Binding
Enzyme regulator activity
Identical protein binding
Protein-containing complex binding
DNA-binding transcription activator activity, RNA polymerase II-specific
Enzyme inhibitor activity
Histone acetyltransferase binding
Signal
FDR
Gene count
Groups at similarity 0.8
M
Young to Old
Mean expression by highres_celltype
Alox5ap
Apoe
Gsr
Hp
Mgst1
Prdx2
Prdx5
Prdx6
S100a8
S100a9
Selenow
Fold Change
Higher in Old
No change
Higher in Young
Log-Fold Change
Neutrophil
Mono P
Myelo P
Ery P
HSC
MKP
MBEP
CLP
GMP
LMPP
Memory B cell
Immature B cell
Mature Naive B cell
Pro-B cell
Pre-B II cell
Pre-B cell
Monocyte
cMoP
cDC
pDC
Basophil
Basophil Progenitor