viSNE and Wanderlust, two algorithms for the visualization and analysis of high-dimensional single-cell data

El-ad David Amir

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

viSNE and Wanderlust, two algorithms for the visualization and analysis of high-dimensional single-cell data

El-ad David Amir

The immune system presents a unique opportunity for studying development in mammals. White blood cells undergo differentiation and proliferation, a never-ending process throughout the life of the organism. Hematopoiesis, the development of cells in the immune system, depends upon the interaction between many different cell types (some of which comprise less than a tenth of a percent of the population), transient regulatory decisions, genomic rearrangement events, cell proliferation, and death. To capture these events we employ mass cytometry, a novel technology that measures fifty proteins simultaneously in single cells. Mass cytometry results in large quantities of high-dimensional data which challenges existing computational techniques. To address these challenges, we developed two dimensionality reduction algorithms for analyzing mass cytometry and other single-cell data. The first, viSNE, transforms high-dimensional data into an intuitive two-dimensional map, making it accessible to visual exploration. The second algorithm, Wanderlust, receives as input a static snapshot (where cells occupy different stages of their development) and constructs their developmental ordering: the developmental trajectory.

viSNE maps healthy bone marrow into a canonical shape that separates cell subtypes. In leukemia, however, the shape is malformed: the maps of cancer samples are distinct from the healthy map and from each other. The algorithm highlights structure in the heterogeneity of

surface phenotype expression in cancer, traverses the progression from diagnosis to relapse, and identifies a rare leukemia population in minimal residual disease settings.

Wanderlust was applied to healthy B lineage cells, where the trajectory follows known marker expression trends and genetic recombination events. Using the Wanderlust trajectory we identified CD24 as an early marker of B cell development. The trajectory captures the coordination between several regulatory mechanisms (surface marker expression, signaling, proliferation and apoptosis) during crucial development checkpoints.

As new technologies raise the number of simultaneously measured parameters in each cell to the hundreds, viSNE and Wanderlust will become a mainstay in analyzing and interpreting such experiments.

Tables of Contents

List of Figures

List of Tables

Acknowledgements

Exactly five years ago, in May 2008, a person was sitting in front of a computer in Beersheba, Israel. He was writing the acknowledgements of his Master's thesis, the first sentence of which read (loosely translated from Hebrew):

"I always naively believed that the solution was right around the corner."

Although I share a name with that person, the five years that separate us are an insurmountable chasm: I am not him, although there is a high pro-Bability (given a reasonable definition of the multiverse) that one day he will become me. This time around I knew that the solution would not be waiting right around the corner. Instead, I painstakingly gathered evidence for its existence, and then, in a moment of inspiration (or luck?), I seized it. HA!

Many have accompanied me throughout this journey and deserve gratitude and recognition. First and foremost, I would like to thank Prof. Dana Pe'er, my adviser and mentor for the past five years. Dana possesses a unique, piercing view of biology and statistics, and I have learned many valuable lessons from her approach to science and the academic world. I am indebted to the members of my thesis committee: Prof. Boris Reizis, Prof. Ulf Klein, Prof. Christina Leslie and Prof. Peter Sims. Their comments and suggestions have been crucial for the maturation and success of the work described here. I am also grateful to the Howard Hughes Medical Institute, who generously provided the funding for my work at Columbia University; to Prof. Garry Nolan, Dr. Sean Bendall, Dr. Kara Davis and Dr. Erin Simonds, my collaborators, with whom I held many fruitful and fascinating discussions; and to Daniel, Anat, Bo-Juen, Oren and Michelle, my colleagues and friends, whose advice and support made this research possible.

מוקדש להוריי, דבורה ורן, לאחיותיי, הדר וליאת אביבה,

ולשלושת מערכות הכוכבים הזוהרות בשמי הלילה של ניו יורק:

עינת,

רועי יונתן

ודניאל אפריל.

Chapter 1 Introduction

1.1 Single-cell analysis of the immune system

The immune system is one of the organism's central lines of defense against pathogens [1]. It is a complex system, composed of many different cell subtypes, most of which belong to one of two major lineages [2]: the myeloid lineage (that includes monocytes, macrophages, erythrocytes and others) and the lymphoid lineage (T cells, B cells and NK cells). Each such subset can be further divided into multiple high-specialized cell subtypes. All immune system cells share a common ancestor, the hematopoietic stem cell (HSC) [2]. The HSC gives rise to different progenitors that in turn develop into the final, differentiated cells.

Both the healthy and the diseased immune system depend upon the interaction between many different cell types (some of which comprise less than a tenth of a percent of the population), transient regulatory decisions, genomic rearrangement, cell proliferation, and death. Due to this high level of heterogeneity, single-cell methods are an ideal choice for exploring the immune system. Flow cytometry has become a technology of choice for cellular analysis in the immune system. It has paved the way to understanding many different cellular processes in immunology and in other fields, such as profiling phosphor-protein networks in cancer cells [3], elucidating causal influences in protein-signaling networks [4], identifying the hierarchy between three different types of stem cells in human [5], and exploring the role of stem cells in healthy systems [6] and in cancer [7].

In the context of the immune system, a typical flow cytometry experiment will assay four to eight markers. Markers are labeled with antibodies that target a specific protein or a specific

form of a protein (such as phosphorylated state). They belong to one of several categories. Surface markers, which reside on the cell's membrane, are mostly receptors (though they fill other functions as well, such as enzymes in the case of CD45). They are most often used for classification of cell type. Signaling markers are present in the cell's cytoplasm and transduce signals from outside and within the cell. They allow us to assay the cell's current activity and its regulation. Additional markers might also be used, including cell cycle markers (Ki67 and cParp in this work), genomic recombination markers (TdT and Rag1/2) or transcription factors.

From a computational point of view, the output of the flow cytometer is a matrix with a column for each marker and a row for each cell. Before actual analysis can commence, several pre-processing steps need to be done. First, the instrument accumulates debris, such as cell fragments, that needs to be filtered out (gated), usually based on their size or lack of DNA. Cell types which are not relevant to the current research should also be removed- for example, T cells in a B-cell experiment. Next, the measurements of each marker need to be compensated due to fluorescence spectral overlap, a phenomenon where the light spectrum of one fluorochrome overlaps another. Finally, during the analysis itself, antibody "stickiness" should be considered: while antibodies have the highest affinity to their target, they also bind to other molecules (including the cell's membrane), and form a source of noise that needs to be acknowledged.

Improvements in optics and hardware and the discovery of new fluorochromes have steadily increased the number of concurrent parameters that can be measured in each cell [8] and recent work has reached 15-color experiments [9]. However, due to fluorescence spectral overlap, compensation becomes more challenging as the number of colors increases, resulting in difficult staining panel design [10] and potential artifacts [11] and imposing a physical bound on the

number of parameters that flow cytometry can measure. A different technology is required in order to capture the full complexity of the immune system.

Mass cytometry [8, 12, 13] offers a variation on flow cytometry by attaching lanthanide isotopes to the antibodies targeting the proteins of interest (instead of fluorochromes as in flow cytometry). The isotope levels in each cell are measured using an atomic mass spectrometer. Since there is no spectral overlap, no compensation is required. The method can currently measure approximately 50 parameters per cell, with a theoretical limit of 100 parameters per cell [14, 15]. Mass cytometry is joined by other recent technological advances that have enabled the study of a large number of parameters in single cells. For example, high-resolution microscopy [16, 17] and single-cell RNA quantification [18-21] allow analysis of 100 parameters in hundreds and soon thousands of individual cells. These innovations promise to transform the way we research, study and think about development, differentiation, and disease [8, 13, 22].

Mass cytometry and similar groundbreaking single-cell technologies raise several computational challenges. One, merely visualizing such high-dimensional data (millions of cells over a hundred dimensions) in an intuitive, accessible manner is daunting. Innovative methods will be required to give us an initial window to understanding the complexities therein. Two, the regulatory pathways in biological systems are complex and involve feedback and crosstalk [23, 24]. Therefore, analysis cannot assume simple pairwise or linear relationships, but should adopt a systematic approach and integrate information from all of the parameters available. Three, single-cell technologies often aim to measure miniscule amounts of each target protein, leading to high noise levels which should be acknowledged and addressed. The combination of these issues calls for the development of novel computational approaches for the analysis of high-dimensional single-cell data.

1.2 Dimensionality reduction

In the context of the current work, the challenge of analyzing high-dimensional single-cell data is answered using dimensionality reduction [25]. As its name implies, dimensionality reduction is a class of algorithms that transform high-dimensional data into a compact lower-dimensional representation. The number of dimensions in the reduced representation depends on its purpose. When the goal is to reduce the memory or computational requirements for further analysis the number of dimensions will be dictated by the computational complexity of that analysis or the size of available memory [26-28]. Oftentimes, however, the purpose of dimensionality reduction is to enable exploration and classification of the data by creating effective visualization [29]. In such cases, the number of dimensions in the reduced representation is set to two or three, facilitating its visualization and making it accessible to human researchers.

Dimensionality reduction is defined as follows: given a data set X, (represented as a matrix of size $N \times D$, denoting N observations in D dimensions), we are looking for a projection of X into a lower-dimensional matrix Y of size $N \times d$, where $d << D$. The projection should maintain the geometry of X. The minimal d that is required to preserve the geometry of X is called the intrinsic dimensionality of the data.

This definition portrays the two central challenges to dimensionality reduction. One, the geometry of X is often unknown or ill-defined. Therefore, it is rarely possible to provide an accurate, quantitative metric that describes the similarity between the geometries of X and Y. The initial assumption of any dimensionality reduction algorithm is the definition of the geometry to be preserved. Two, we do not know the value of d, the intrinsic dimensionality of X. While several techniques exist for estimating it [30-32], they are themselves approximations, they

might mismatch or contradict the algorithm's geometry assumption, and the intrinsic dimensionality might be higher than afforded by our goal (for example, no matter the value of *d*, we cannot visualize more than three dimensions). A robust dimensionality reduction algorithm should address these concerns.

Principal component analysis (PCA) [29] and factor analysis (FA) [33] are two leading and well-studied methods of dimensionality reduction. Briefly, PCA is a linear transformation of the data into the principal eigenvectors (called components) of the covariance matrix of the zero-mean centered *X*. The components are orthogonal and maximize the amount of variance in the data. The first component has the largest variance and therefore accounts for the most information in the data. The variance decreases as the components increase (second, third, etc.). FA, a similar method, searches for latent variables, called factors. The data is modeled as a combination of the factors and is fit using linear regression. The factors are then explained as real-world phenomena that give rise to the data. While PCA and FA are related, they follow opposite conceptual approaches: PCA extracts the component while FA estimates the factors.

PCA has been successfully applied in many biological contexts, for example: de-convoluting in vivo signaling data [34], providing a visual validation for a gene-based noise modelling [35], and in previous work relating to detecting a small subset of the B-lineage trajectory [12]. Despite its success, at the basis of PCA (and FA) lies the assumption that the relationship between variables in the data is linear. However, nonlinearity is abundant across many real-life systems, including in fields as diverse as ecology [36], machine vision [37], engineering [38], and biology [4, 12, 39, 40]. In all of these, nonlinearity is the norm rather than the exception. While PCA and FA might offer rough outlines of the data's true structure, more often than not they will lead to imprecise or even outright fallacious conclusions. The situation is further complicated by the

challenge of quantifying the accuracy of the resulting model: initial observation might suggest that a linear approximation is correct, missing more subtle nuances of the data.

Recent years have seen a resurgence of nonlinear dimensionality reduction algorithms. Broadly, these can be classified into one of two categories: projective methods and manifold modeling [41]. Projective methods aim to preserve global properties of the data. They define a metric that considers all of the data points and optimize that metric in the low-dimensional representation. Manifold modeling, on the other hand, is based on preserving local properties of the data. A manifold is a topological space that resembles Euclidean space in the close vicinity of each point [42]. Manifold modeling methods approximate the manifold by examining a small neighborhood around each data point and match the neighborhood's structure between the low- and high-dimensional representations. Another important classification that crosses both categories is spectral methods [43], which derive the low-dimensional matrix from the eigenvectors of a matrix constructed from *X*. Based on these definitions, PCA is spectral while FA is not, and both methods are linear and projective.

Isomap is a characteristic projective method [44]. Isomap's goal is to preserve the intrinsic geometry of the data, based on the definition of the geodesic distance: the distance between a pair of points across the manifold. The algorithm has three steps. First, the data is transformed into a nearest-neighbors graph, where a nearest-neighbor is defined either as a point in a given radius, or one of the k nearest neighbors. Then, Isomap estimates the geodesic distance between each pair of points as the shortest-path distance across the graph. Finally, Isomap applies multidimensional scaling [45], an extension of PCA to non-Euclidean distances, to the matrix of graph distances. Multidimensional scaling preserves the geodesic distances in the low-dimensional representation. Since the reduction space of Isomap is based on multidimensional

scaling, it is computationally efficient, has a global optimum, and is guaranteed to asymptotically converge, three favorable algorithmic characteristics.

Two other popular projective methods are maximum variance unfolding (MVU) [46] and Kernel PCA [47]. MVU extends Isomap by trying to maximize the Euclidean distance between data points while preserving the geodesic distances, with the goal of unfolding the manifold [48]. This additional constraint necessitates optimization through semidefinite programming, increasing run time and losing the asymptotic convergence guarantee. Kernel PCA extends PCA into a nonlinear method through the application of the so-called kernel trick [49]. A kernel function maps the data space into a higher-dimensional feature space and PCA is run over the feature space. The method is fast and guaranteed to converge. However, kernel function selection is critical and there are no computational methods for choosing the correct one for the data at hand.

Locally linear embedding (LLE) [50] is one of the earliest and best-studied manifold modeling methods. Following a conceptual approach opposite to that of Isomap, LLE concerns itself only with optimizing tight, local neighborhoods. As with the previous algorithms, LLE begins by finding the nearest neighbors for each data point. Next, for each point, it computes a series of weights that best reconstruct the point from its neighbors; appropriately, these are referred to as reconstruction weights. This calculation is formed as a least-squares pro-Blem. The combination of all of the reconstruction weights forms a sparse matrix *W*. On its last step, LLE returns the low-dimensional representation that best preserves these weights by solving the spectral pro-Blem of finding the *d* smallest eigenvectors of *W*. The embedding is nonlinear since the matrix is based solely on local neighborhoods. Thanks to *W*'s sparsity, LLE is computationally fast and has low memory requirements.

Two notable additional manifold modeling methods are Hessian eigenmaps [51] and stochastic neighbors embedding (SNE) [52]. Hessian eigenmaps waives an assumption shared by both Isomap and LLE, namely, that the high-dimensional space is convex. When embedding the data in the low-dimensional space, the algorithm first minimizes the local curvature of the high-dimensional space using a Hessian estimator, which is calculated using PCA. The estimators form a matrix *H*, and its *d* smallest eigenvectors are the low-dimensional representation. Hessian eigenmaps is also referred to as Hessien LLE due to similarity of this final spectral step to the aforementioned algorithm. SNE replaces LLE's "hard" neighborhoods with a "soft", pro-Babilistic neighborhood definition. The Kullback-Leibler divergence between the set of pro-Babilities in the high- and low-dimensional spaces is minimized using gradient descent. SNE forms the basis of t-SNE, which in turn is the heart of viSNE, one of the novel methods defined in this work, and is therefore described in detail later.

1.3 viSNE and Wanderlust

Here we introduce two novel dimensionality reduction methods for exploring single-cell data, and our findings from applying these methods to questions about the structure, shape and development of the immune system.

In chapter 2 we discuss viSNE, a dimensionality reduction algorithm that maps high-dimensional data into two dimensions. The viSNE map is reminiscent of a scatter plot, a central tool in immunology, and therefore forms a familiar visual. However, it preserves pairwise relationships from the high-dimensional space and therefore encapsulates information that would not be accessible otherwise. In the immune system, for example, the viSNE map faithfully captures the classification of cells into subsets, separates between healthy and cancer bone marrow samples,

identifies the progression of cancer from diagnosis to relapse, and detects a tiny cancer population in an otherwise healthy sample. The viSNE map validates expected notions about the system while serving as a hypothesis-generation tool, revealing unfamiliar behaviors that can be further researched.

In chapter 3 we present Wanderlust, a graph-based trajectory detection algorithm. Wanderlust receives a snapshot of a developing system and outputs its underlying trajectory: the temporal ordering of the cells. The trajectory detection is done *de novo* and requires no prior knowledge of the system other than a single starting cell. The trajectory serves as scaffolding: we preserve the resolution of the single-cells, but instead of examining each cell, we instead explore the ordering of developmental events. This allows us to move from constituent cells underlying development, to ordering the developmental events themselves. When applied to mass cytometry data of B lineage cells, the Wanderlust trajectory follows known trends of B cell development. We validated the ordering of cells identified by the algorithm by confirming that the DNA rearrangement status of immunoglobulin across the trajectory follows the expected chronology. Through the trajectory we show how different regulatory mechanisms are coordinated in developmental checkpoints. The bone marrow contains cells from all differentiation steps, allowing Wanderlust to extract these results from a single time point. Applications of the algorithm can potentially extend beyond the immune system to any system where acquisition of time-series data is difficult or impossible or when the developmental process is only partly known.

The abundance of novel single-cell technologies is an invaluable addition to our arsenal of research methods. Biological systems involve high-order relationships, heterogeneity and stochasticity, and incorporating these attributes is crucial for fully understanding the pro-Blem

we are examining. We are certain that viSNE and Wanderlust, two powerful, complementary methods for the exploration and analysis of high-dimensional data, will be imperative in utilizing single-cell technologies to the fullest.

The rest of this work is organized as follows. Chapter 2 discusses the viSNE algorithm, its implementation, the viSNE maps of healthy and cancer bone marrow samples, and the algorithm's application in minimal residual disease settings. Chapter 3 introduces the intuition behind Wanderlust and the technical details of the algorithm, and showcases its robustness when applied to synthetic data. Chapter 4 revolves around application of Wanderlust to the development of B-lineage cells in the healthy immune system. Chapter 5 presents the broader implications of the algorithms presented for current biological experimentation and for the future of the field.

Chapter 2 viSNE enables visualization of high-dimensional single-cell data and reveals phenotypic heterogeneity of leukemia

The following text is a reprint of Amir et al. [53].

2.1 Introduction

Emerging single-cell technologies have revealed an extensive degree of heterogeneity between and within tissues [8]. Analysis of single-cell data has shed light on many different cellular processes [3-7, 54] and recent technological advances have enabled the study of a large number of parameters in single cells at unparalleled resolution. For example, mass cytometry [14] can measure up to 45 parameters simultaneously in tens of thousands of individual cells. High-resolution microscopy [16, 17] and single-cell RNA quantification [18-21] allow analysis of 100 parameters in dozens and soon hundreds of individual cells. These innovations promise to transform the way we think about development, differentiation, and disease [8, 13, 22].

However, it is difficult to visualize such high numbers of dimensions in a meaningful manner. Single-cell data is often examined in two dimensions at a time in the form of a scatter plot [55]. Yet, as the number of parameters increases, the number of pairs becomes overwhelming. A typical mass cytometry dataset allows several hundred pairwise combinations. In addition, a pairwise viewpoint could miss biologically meaningful multivariate relationships that cannot be discerned in two dimensions. Several computational tools, such as SPADE [56], have been developed to address these pro-Blems [57, 58]. However, these approaches typically cluster cells

and examine the average of each cluster, resulting in the loss of single-cell resolution of the data. Principal component analysis (PCA) [29], another computational tool, has been applied to mass cytometry datasets [12] and can be used to project data into two dimensions while maintaining single-cell resolution. However, PCA is a linear transformation that cannot faithfully capture the nonlinear relationships that are a hallmark of many single-cell datasets. Therefore, we need new tools to visualize and interpret high-dimensional single-cell data such as those produced by mass cytometry. An ideal tool would enable visualization at single-cell resolution, preserve the geometry and nonlinearity of the data, represent both abundant and rare populations, and provide a robust, interpretable view of the data.

We developed viSNE for this purpose. viSNE allows visualization of high-dimensional single-cell data and is based on the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm [12]. viSNE finds the two dimensional representation of single-cell data that best preserves their local and global geometry. The resulting viSNE map provides a visual representation of the single-cell data that is similar to a biaxial plot, but the positions of cells reflect their proximity in high-dimensional rather than two dimensional space. We utilize color as a third dimension to interactively visualize features of these cells. Here we apply viSNE to interpret mass cytometry data derived from healthy and leukemic human bone marrow.

2.2 Results

2.2.1 Preserving high-dimensional relationships in single-cell data

In viSNE, each cell is represented as a point in high-dimensional space. Each dimension is one parameter (i.e, the expression level of one protein). An optimization algorithm searches for a projection of the points from the high-dimensional space into two or three dimensions such that

pairwise distances between the points are best conserved between the high- and low-dimensional spaces (see Materials and Methods). The resulting low-dimensional projection, which we call the viSNE map, is visualized as a scatter plot, where a cell's location in the plot represents information from all of the original dimensions.

We also developed *cyt*, an interactive tool for visualization of viSNE maps. *cyt* has multiple features, including plotting the maps, coloring cells by marker expression, sample or subtype, and gating. Figure 2-1A demonstrates how viSNE works on a synthetic example; the optimization algorithm identifies the global structure of the data (a 1D line, along with its curvature, embedded in 3D space) and the local structure (pairwise distances between points along the line are conserved).

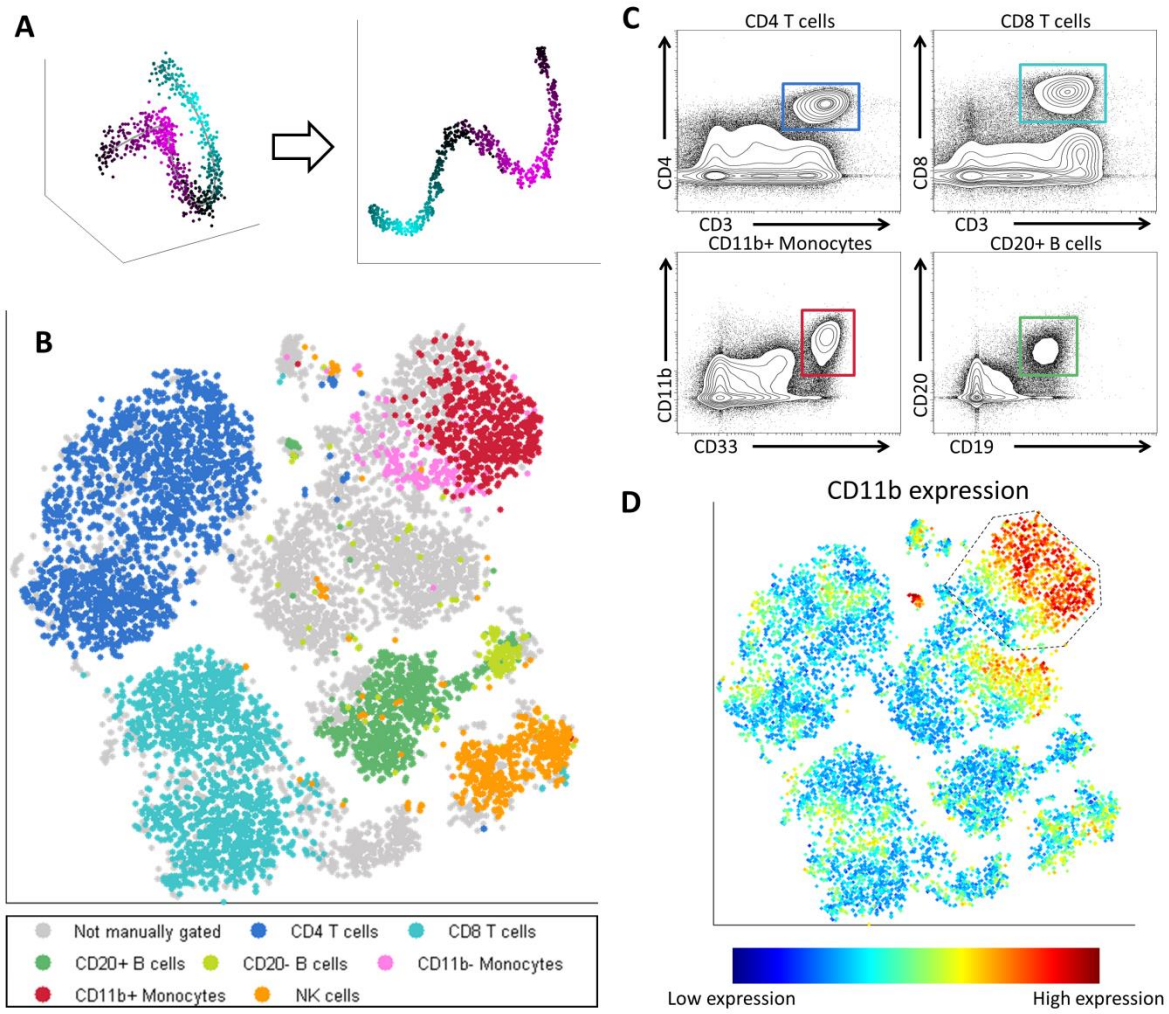


Figure 2-1. viSNE map of healthy human bone marrow.
(a) In a synthetic toy example, viSNE projects a one-dimensional curve embedded in three dimensions (left) onto two dimensions (right). The color gradient shows that points that are in close proximity in three dimensions remain in close proximity in two dimensions. (b) Application of viSNE to a healthy human bone marrow sample, stained with 13 markers and measured with mass cytometry [12] automatically separates cells into spatially distinct subsets based on the combination of markers that they express. Each point in the viSNE map represents an individual cell and its color represents its immune cell subset as designated by independent expert manual gating (manual gates are defined at the bottom). Gray points were not classified by manual gating. The axes are in arbitrary units. (c) Biaxial plots represent the same data shown in panel b, and show the gates drawn manually by expert operators. The colors of the squares match the colors in panel b. The actual manual gating used here is more complex and uses a series of biaxial plots to gate each population [12]. Note, that unlike in panel b, no single biaxial plot spatially separates all subsets. (d) The same viSNE map shown in panel b is colored according to intensity of CD11b expression. Many of the cells within the dotted line gate were not classified as monocytes by manual gating (grey cells panel b).

2.2.2 viSNE map of healthy human bone marrow

First, we evaluated viSNE's ability to map the well-characterized system of human bone marrow hematopoiesis [2]. We analyzed previously generated data of healthy human bone marrow (Marrow 1) stained with elemental isotope-conjugated antibodies specific for 13 surface markers [12]. When applied to this dataset, viSNE generated a map that clearly separated different cell subsets in space (Figure 2-1b). To validate and label the map, we used an independently derived classification of the cells based on expert manual gating of a series of biaxial plots (Figure 2-1c; see Materials and Methods). Although viSNE was not provided with this classification or with any knowledge of immune subsets, it successfully grouped cells of the same subset together and cells of different subsets separately (Figure 2-1b-c). viSNE accurately distinguished $CD4^+$ and $CD8^+$ T cells, mature and immature B cells, mature and immature monocytes and natural killer (NK) cells. Notably, NK cells formed a distinct subset even though CD56, the canonical marker associated with this lineage, was not included in the antibody staining panel.

To further compare the expert manual gating and the viSNE map, we used the *cyt* feature to gate subsets directly from the viSNE map (Figure 2-2a). In all cases, the viSNE gate included cells that were not classified by the expert manually gated biaxial plots; these cells are labeled in grey in the viSNE map. Examination of the marker expression of these cells reveals that they are typically just beyond the threshold of one marker, but the viSNE classification is strongly supported based on the expression of all other markers. For example, in Figure 2-1d, wherein cells are colored for CD11b marker expression, the cells in the gated region express the canonical monocyte marker CD33 (Figure 2-2b). However, only 47% of these cells were classified as monocytes by the manual gating (Figure 2-1b). In addition, the marker intensity distributions between the CD11b- monocytes in the viSNE map monocyte gate and in the

manually set monocyte gate (Figure 2-2c) are similar, supporting the notion that the cells gated in the viSNE map are indeed CD11b- monocytes.

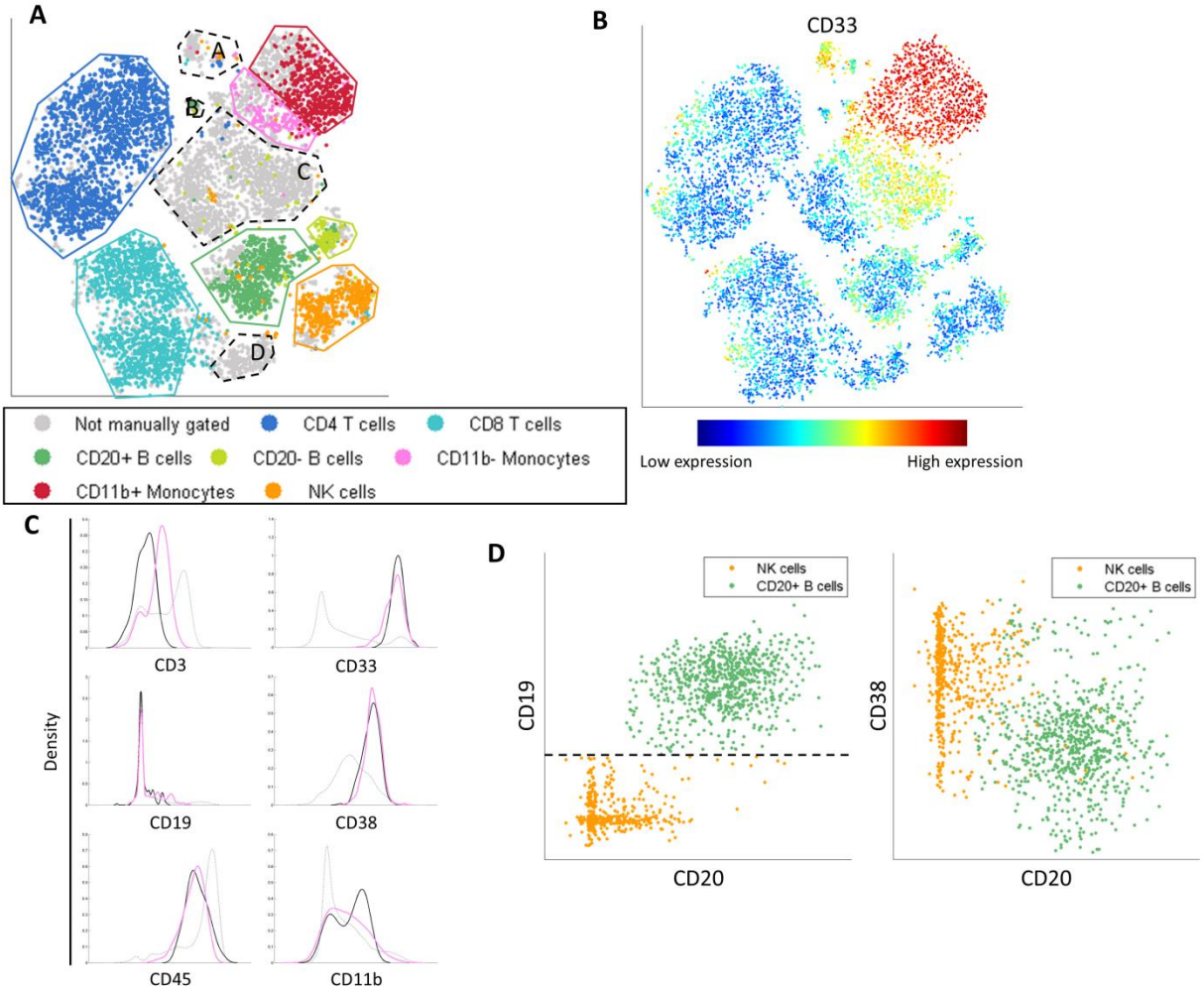


Figure 2-2. A viSNE map can classify cells that were labeled incorrectly by manual gating.
(a) viSNE map is identical to Figure 2-1b. Each of the cell subtypes is surrounded by a gate corresponding to that subtype's color as designated by manual gating. Grey points inside the viSNE map gate were not classified by manual gating, although further examination reveals that they belong to the relevant subtype. In addition, there are four regions that do not conform to known subtypes. Region A has cells positive for CD45 only, we suspect these are missing a marker needed to classify them. Region B includes doublets: pairs of cells that were read together by the machine as if a single cell, and are therefore positive for suspect marker combinations (for example, CD19+ CD11b+). The cells in Region C are negative for all channels and are pro-Bably debris. Region D has cells positive for CD45 and CD3 only, these might be NKT cells whose canonical marker is missing. (b) The same map as in A, coded by CD33 (myeloid marker) expression. The monocyte cell population is clearly visible to the top right. (c) Marker expression level densities for the entire population (grey), the manually gated CD11b- monocyte cells (black) and the viSNE gated CD11b- monocytes cells (pink, as shown in panel A). The two CD11b- monocyte populations have almost identical marker distributions, except the pink population shows some CD3 staining due to reagent "stickiness", but this level of

CD3 expression is significantly lower than that on bona fide CD3+ cells. Rather than exclude these cells based on a hard threshold on a single marker, viSNE groups the cells together based on their tight similarity in all other markers. Taken together these data support viSNE's identification of CD11b- monocytes. (d) Gating using the viSNE map is typically more accurate than manual gating, since 2D views and hard thresholds can be misleading. There are a number of cells labeled as CD20+ B-cells by viSNE (dark green outline in A) and labeled NK cells based on biaxial gating (orange dots within the B-cell region). These cells just miss the hard CD19 threshold in one of the gates (black dashed line) and therefore their CD20 level is never examined during the gating (the presented biaxial plot is not part of the gating scheme used). Their high CD20 levels and borderline CD19 levels support their labeling as B-cells, rather than NK-cells. Their CD38 levels further support their B-cell label.

Traditional gating relies on hard thresholds to classify cells into subsets. Thus cells whose marker values are slightly below or above the threshold might not be classified correctly, or classified at all (Figure 2-2d). When dealing with the hematopoietic continuum, this may result in the inability to accurately capture transitional cell types. For example, using *cyt* to color cells based on marker intensity revealed that viSNE organized monocytes based on a gradient or smooth increase in expression of CD11b, a marker of monocyte maturity (Figure 2-1d). This finding highlights the continuous and gradual nature of CD11b expression during monocyte maturation and better represents the continuum of normal differentiation [59]. viSNE takes into account all phenotypic markers concurrently instead of relying on hard thresholds and, as a result, classifies more cells and captures a more accurate view of the variability within each subset when compared to biaxial gating. The single cell resolution of the viSNE map provides fine detail of each subset, going beyond clustering and enabling investigation of the variation, structure and transitions within each subset.

2.2.3 Robust subset classification even without canonical markers

We performed a number of analyses to evaluate the robustness of viSNE. The viSNE map in Figure 2-1b includes 10,000 cells that were subsampled from the complete data set of Marrow1. We independently subsampled multiple subsets of the data and ran viSNE on each. Reassuringly, these separate analyses resulted in similar viSNE maps that conserved the spatial separation

between subsets. Thus the viSNE map consistently and reliably represents real structure in the data.

To test viSNE's reliance on specific markers for classification of immune cell subsets, we generated multiple viSNE maps but excluded some of the markers when generating each map. The viSNE map remained consistent in terms of spatial separation of subsets even after removal of any single marker (Figure 2-3a). Remarkably, even after excluding the canonical markers of B cells, T cells and myeloid cells (CD19/CD20, CD3 and CD33, respectively), the viSNE map remained consistent with the map constructed using all thirteen markers (Figure 2-3a). These findings imply that non-canonical markers, when analyzed together, contain the information needed to separate distinct immune cell subsets. This speaks to a previously unappreciated level of organization where specialized immune subtypes have tightly coordinated surface marker expression beyond their canonical identifiers. The different subtypes reside in distinct well-separated subset shapes in high-dimensional space.

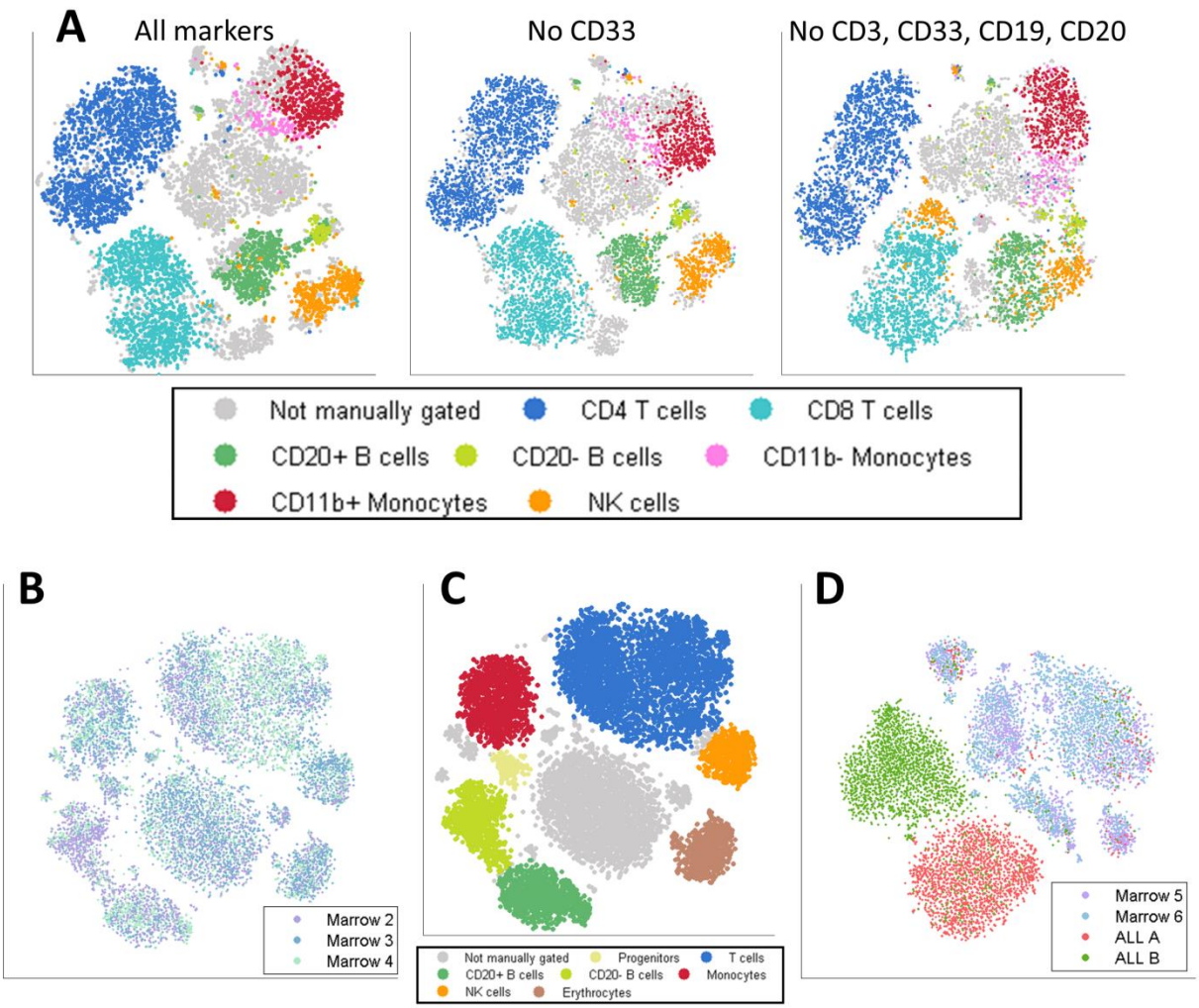


Figure 2-3. viSNE is robust, consistent, does not require canonical markers.
(a) The left map is the same as in Figure 2-1b, and was generated by considering all 13 markers. Middle: viSNE map of the same cells, projected after removing CD33. Right: viSNE map of same cells, projected after removing CD33, CD3, CD19 and CD20. Despite removing four canonical markers, viSNE separates most major subtypes using the remaining nine channels. (b) Bone marrow samples from three healthy donors (Marrow2-4) were mapped using viSNE. Each point represents a single cell, and different colors represent different samples. (c) The map is the same map as in panel b, but is color coded by cell subset as identified by analyzing gradients of expression of individual markers. Subsets are indicated below the map. (D) Bone marrow samples from two healthy donors (Marrow5-6) and two pediatric B-cell ALL patients (ALL A-B) were mapped using viSNE. Samples are color-coded as indicated in key.

2.2.4 Consistent and reproducible healthy bone marrow map

Having demonstrated viSNE's robustness when applied to a single healthy bone marrow sample, we examined its robustness across bone marrow samples from multiple healthy individuals.

Three healthy bone marrow samples (Marrow2-4) were assayed by mass cytometry using a panel of 31 phenotypic surface markers. The resulting viSNE map grouped cells into distinct subpopulations, and cells from all three individuals overlapped within each subpopulation (Figure 2-3b). We used the Jenson-Shannon (JS) divergence to quantify the similarity between the viSNE maps of the three individuals (see Materials and Methods). The JS divergence between each pair of healthy individuals was 0.04, confirming that there is almost no divergence between the viSNE maps of healthy samples. Using *cyt* to visualize expression gradients of individual markers, we gated specific immune cell subsets in viSNE maps (Figure 2-3c).

To further evaluate the robustness of viSNE's map of healthy bone marrow, we applied viSNE to an additional bone marrow sample collected using conventional fluorescence-based flow cytometry. The resulting viSNE map is similar to the map generated by mass cytometry (Figure 2-4), demonstrating not only consistency in the map between healthy samples, but also that viSNE is well-suited for the analysis of fluorescence-based cytometry data. The cellular subtypes comprising the human immune system are reproducibly represented by viSNE and the fidelity of this structure is maintained across multiple cytometry platforms, marker panels, and most importantly, individual patients.

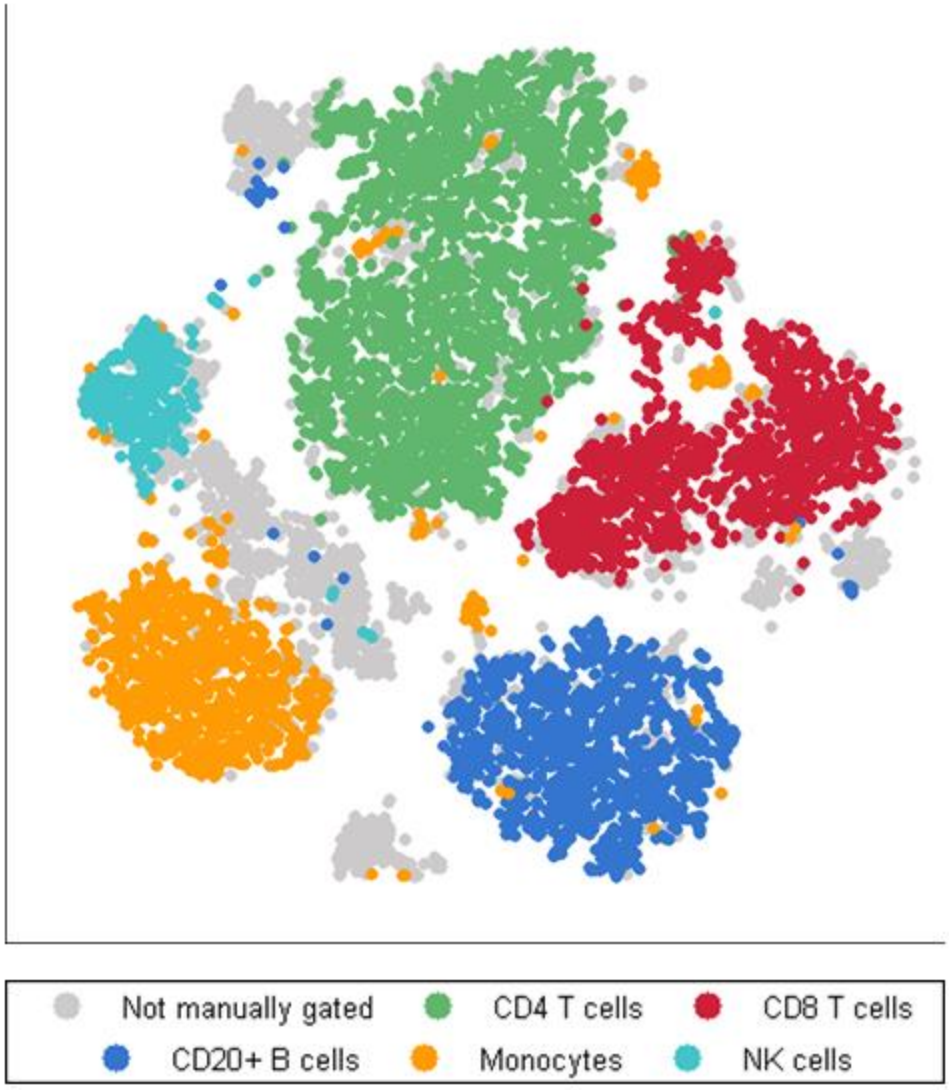


Figure 2-4. viSNE can be used for the analysis of flow cytometry data.
viSNE map of flow cytometry data of healthy bone marrow. The flow cytometry panel includes eight markers (CD45, CD45RA, CD3, CD4, CD8, CD33, CD20, CD7) and therefore identifies fewer subsets.

2.2.5 Deformed shapes of leukemic bone marrow maps

Encouraged by the consistency and robustness of viSNE maps of healthy bone marrow samples, we used viSNE to analyze leukemic bone marrow. We stained two bone marrow samples donated from healthy individuals and two from pediatric acute B-cell lymphoblastic leukemia (B-cell ALL) patients with a panel of 29 antibodies optimized for the analysis of B-cell ALL (Table 2-1). For example, recognition of erythrocyte and progenitor subsets was aided by the

addition of anti-CD61 and anti-CD117 to the panel, whereas the CD4 and CD8 T cell populations were merged because those markers were omitted.

Figure	Section	Dataset	Cells subsampled	Markers used
1	B/D	Marrow1	10,000 cells from Marrow1	CD11b, CD123, CD19, CD20, CD3, CD33, CD34, CD38, CD4, CD45, CD45RA, CD8, CD90
2	A left	The exact same cells as Figure 1B/D		CD11b, CD123, CD19, CD20, CD3, CD33, CD34, CD38, CD4, CD45, CD45RA, CD8, CD90
	A middle			CD11b, CD123, CD19, CD20, CD3, CD34, CD38, CD4, CD45, CD45RA, CD8, CD90
	A right			CD11b, CD123, CD34, CD38, CD4, CD45, CD45RA, CD8, CD90
	B/C	Marrow2	Combination of 4,000 cells from each of these samples	CD10, CD117, CD11a, CD123, CD127, CD179b, CD19, CD2, CD20, CD22, CD235, CD3, CD33, CD34, CD38, CD43, CD45, CD45RA, CD45RO, CD47, CD49d, CD5, CD61, CD7, CD79b, CXCR4, HLADR, IgD, IgM
		Marrow3		
		Marrow4		
	D	Marrow5	Combination of 2,500 cells from each of these samples	CD10, CD117, CD11b, CD127, CD133, CD179a, CD179b, CD19, CD20, CD22, CD24, CD3, CD33, CD34, CD38, CD43, CD45, CD45RA, CD47, CD49d, CD61, CD7, CD72, CD79b, CXCR4, Flt3, HLA-DR, IgM, pre-BCR
		Marrow6		
		ALL A		
		ALL B		
3	A	ALL A	10,000 cells from ALL A	CD10, CD117, CD11b, CD127, CD133, CD179a, CD179b, CD19, CD20, CD22, CD24, CD3, CD33, CD34, CD38, CD43, CD45, CD45RA, CD47, CD49d, CD61, CD7, CD72, CD79b, CXCR4, Flt3, HLA-DR, IgM, pre-BCR
		ALL B	10,000 cells from ALL B	
	B	AML A	10,000 cells from AML A	CD114, CD117, CD123, CD133, CD14, CD15, CD16, CD19, CD2, CD20, CD22, CD3, CD33, CD34, CD38, CD44, CD45, CD45RA, CD47, CD49d, CD5, CD56, CD64, CD7, CD79b, CD90, CXCR4, Flt3, HLADR, IgD, TIM3
		AML B (Diagnosis)	10,000 cells from AML B	
	C	The exact same cells as Figure 4A, ALL a		
4		AML B (Diagnosis)	Combination of 5,000 cells from each of these samples	
		AML B (Relapse)		
5		The exact same cells as Figure 4, Relapse		CD114, CD117, CD123, CD133, CD14, CD15, CD16, CD19, CD2, CD20, CD22, CD3, CD33, CD34, CD38, CD44, CD45, CD45RA, CD47, CD49d, CD5, CD56, CD64, CD7, CD79b, CD90, CXCR4, Flt3, HLADR,
6		MRD	10,000 cells, biased subsampling from each of these samples (see text)	CD10, CD15, CD20, CD3, CD34, CD38, CD45, CD7
		Control		

Table 2-1. Experiment details, per figure.
Figure and section refers to the location of the figure in the text. In Dataset, Marrow stands for healthy bone marrow, ALL for acute lymphoid leukemia, AML for acute myeloid leukemia and MRD for the *in vitro* MRD

experiment. “Cells subsampled” is the number of cells subsampled from the whole dataset. “Markers used” is the list of surface markers measured in the dataset.

The maps of the two healthy bone marrow samples (Marrow5-6) overlap (JS divergence 0.04) (Figure 2-3d). In contrast, the two ALL samples occupy a completely separate region within the viSNE map (JS divergence 0.45), and each forms a distinct population separate from the other ALL sample (JS divergence 0.42). Some cells from the ALL samples (~5%) overlap with cells from the healthy samples. Inspection of these cells revealed marker combinations that correspond to healthy immune cells, supporting their placement with the other healthy cells.

When we applied viSNE separately to each ALL sample, each sample mapped into a large deformed shape (Figure 2-5a) and several smaller shapes; the latter corresponded to healthy immune cell populations, indicating that the malignant cells were related to each other but sufficiently distinct from healthy cells. We also applied viSNE to two acute myeloid leukemia (AML) patient samples. The viSNE maps of these AML samples also displayed a single large deformed shape (Figure 2-5b), in contrast to the separated and distinct subpopulations of healthy samples. We noted a considerable population structure within each cancer, as discerned by multiple peaks and saddle points in the contour map. Moreover, each cancer sample formed a unique viSNE map, in which healthy subpopulations were consistently separated from the abnormal leukemic subpopulations.

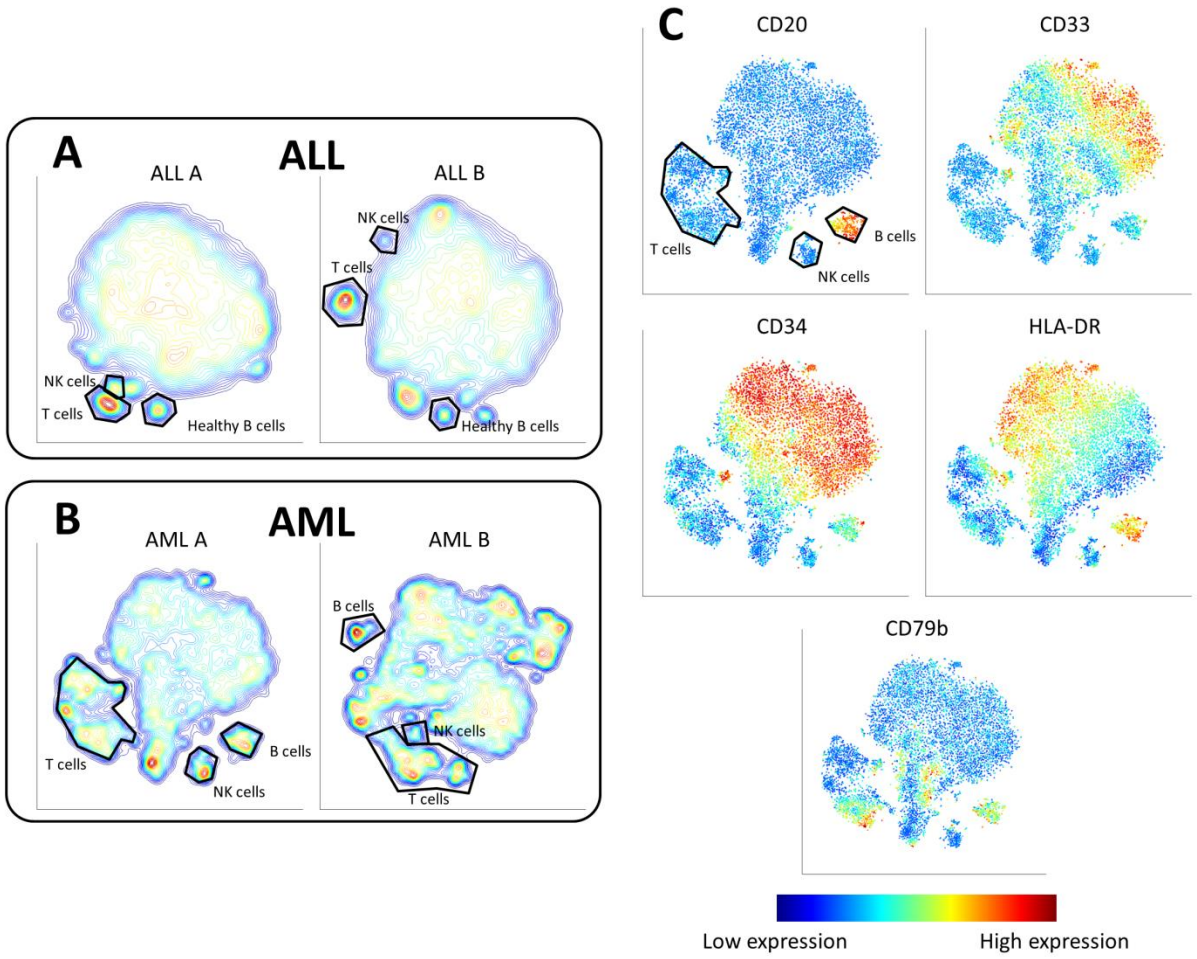


Figure 2-5. Cancer samples form contiguous but heterogeneous shapes.
(a, b) Contour plots of the viSNE maps of two different ALL (a) and AML (b) samples. The contours represent cell density in each region of the map. Small gated populations represent indicated healthy immune subtypes as revealed by examination of their marker expression. Since the structure of each tumor dramatically changes between samples, viSNE places the healthy regions in different locations in each map (as each healthy subtype is positioned as close as possible to the most similar cancer cells). (c) viSNE map of a diagnosis bone marrow sample from AML patient 1. Cells are colored according to intensity of expression of the indicated markers. CD20 helps identify the healthy B cell subpopulation.

2.2.6 viSNE can explore cancer heterogeneity

While healthy samples can be studied by biaxial gating based on known surface phenotypes of individual immune cell subsets, exploring cancer heterogeneity in high dimensions can be a daunting task as cancer samples frequently display abnormal combinations of surface markers and there are hundreds of possible biaxial combinations of surface markers. In current clinical

practice, hematopoietic malignancies are analyzed using at most four to eight markers simultaneously. Hematopoietic malignancy immunophenotyping results have typically been displayed using biaxial plots focused on key markers. However, by combining mass cytometry with viSNE, we are able to visualize cancer at single cell resolution in a single map that takes into account ~30 markers; this sort of analysis can reveal additional structure, abnormal marker combinations and subpopulations.

We used viSNE to comprehensively characterize a diagnostic AML bone marrow sample. Although the overall viSNE map shape of cancer is deformed compared to that of healthy bone marrow, some markers (e.g. CD33, CD34 and HLA-DR) show gradients of expression whereas others (e.g. CD79b) show clustered expression (Figure 2-5c).Within the subpopulation of cells that highly express CD34 (a marker of stem/progenitor cells) is a gradient of expression of CD33 (a marker of monocytes; Figure 2-5c). This marker combination suggests a derailed development program in cancer, because during normal healthy immune cell development, as monocytes mature, expression of CD34 (a marker of immaturity) decreases. Perhaps in this cancer, oncogene activity promoted a progenitor-like $CD34^{+}$ state, but the cells continued to differentiate aberrantly as indicated by the induction of CD33 expression. The single cell resolution of viSNE highlights cancer as a continuum of heterogeneous phenotype states, demarcated by gradients of marker expression rather than distinct subpopulations.

2.2.7 Comparing diagnosis and relapse samples

Because the viSNE map might reflect aspects of cancer progression, we used viSNE to analyze two samples from a single AML patient: one sample was taken before chemotherapy and the other was taken after disease relapse. The map was generated using a merged dataset composed of both samples. Using viSNE , we could clearly visualize a separation between the diagnosis

and relapse samples (Figure 2-6a). viSNE reveals phenotypes unique to the diagnosis sample, which are presumably eliminated by chemotherapy, as well as phenotypes that arise only at relapse. Notably, the viSNE map identifies a region of phenotypes occupied by both samples, but that is considerably rarer at diagnosis. This may suggest enrichment of a rare drug-resistant clone that maintained a consistent phenotype from diagnosis to relapse. We also note populations of healthy cells that overlap in the diagnosis and relapse sample viSNE maps; these provide an internal technical control for the similarity of staining between samples. Regarding specific markers, FLT3 expression is pervasive in the diagnosis sample, but diminished in the relapse sample. Genetic analysis of the diagnosis sample revealed an internal-tandem duplication of FLT3, a common mutation in AML [60], suggesting relapse derived from a clone lacking this mutation (FLT3 genetic status at relapse was unavailable). The clone that reemerged at relapse had an altogether different and more immature phenotype, with cells expressing both high CD34 and CD33 throughout a large fraction of the sample (Figure 2-6b). The relapse sample was highly heterogeneous, as distinct regions expressed different markers from the myeloid lineage (CD64 and CD15) and lymphoid lineage (CD7) (Figure 2-6b)

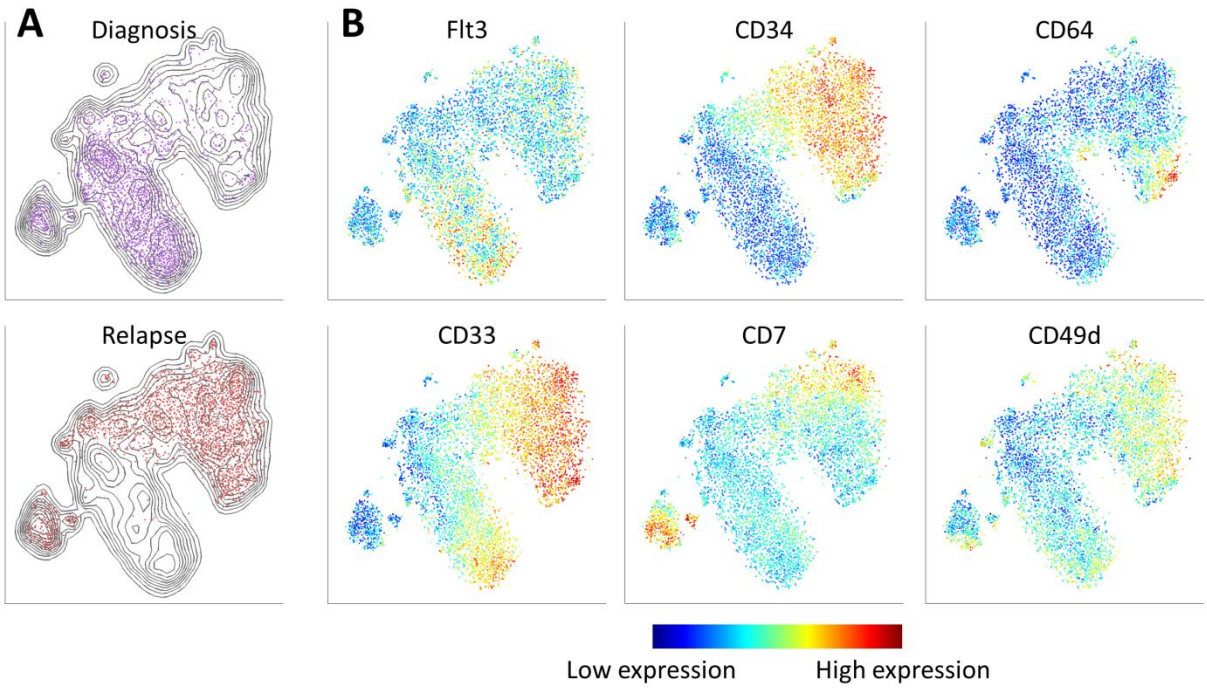


Figure 2-6. viSNE reveals the progression of cancer from diagnosis to relapse.
(a) Contour plots of the viSNE maps of diagnosis and relapse AML samples in patient AML B. The contours represent cell density in each region in the map. The map is the same in each sample. Each point represents a cell from the diagnosis (top, purple) or relapse (bottom, red) sample. (b) Cells from both diagnosis and relapse samples are shown in each map, and the map is the same as in (a). Cells are colored according to intensity of expression of the indicated markers, enabling the comparison of expression patterns before and after relapse. For example, Flt3 is expressed primarily in the diagnosis sample. CD34 emerges in the relapse sample, as do CD64 and CD7. There is a CD33 gradient in both samples. The overlapping region has cells that express high levels of CD49d.

To allow further dissection of the heterogeneity of the AML sample using experimental tools such as DNA and RNA sequencing, we used the viSNE map to devise a gating scheme that is compatible with fluorescence-activated cell sorting (FACS). We divided the AML sample into subpopulations based on expression of CD33, CD34, CD7 and CD64. We classified each marker as "on" (positive) or "off" (negative) according to a threshold that was chosen using the map (Figure 2-7a, black lines). We intersected CD34 and CD33 gates, selected the $CD34^{+}CD34^{+}$ population, and applied the CD64 and CD7 gates to it (Figure 2-7b). When examining the intersection of all eight groups (two for each marker), we identified six distinct subpopulations having at least 20 cells each which spread across the viSNE map (Figure 2-7c). The next step

would be to physically separate the relapse sample into these subpopulations using FACS and characterize them via downstream experiments.

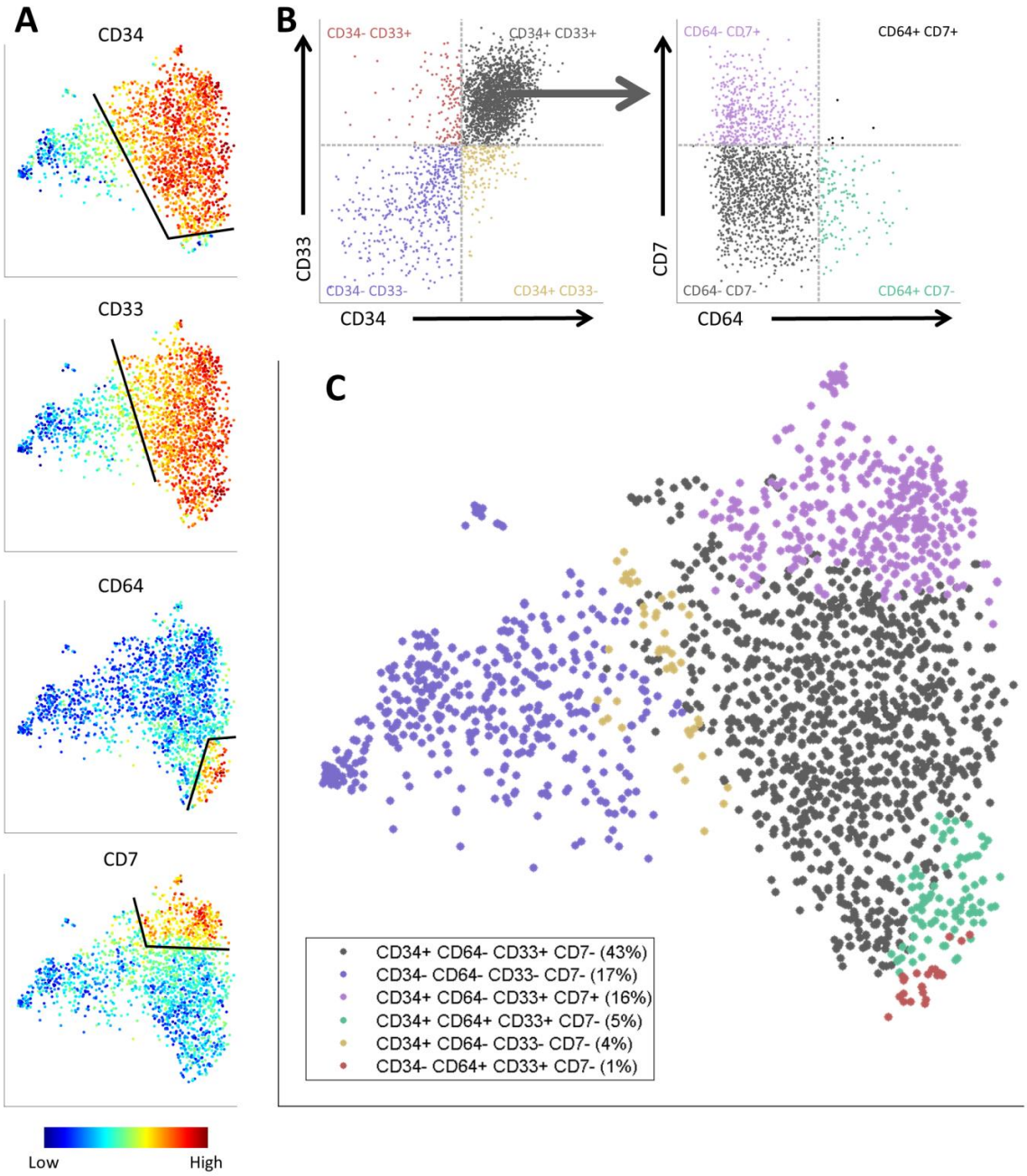


Figure 2-7. A gating scheme for fluorescence-activated cell sorting (FACS) of an AML relapse sample in patient B based on the viSNE map.
(a) The viSNE map, colored by intensity of expression of (from top to bottom) CD34, CD33, CD64 and CD7. For each marker, cells were separated into two subpopulations: "on" (positive) and "off" (negative), based on an expression threshold (black lines). (b) Left: Biaxial plot of CD34 versus CD33. Right: Biaxial plot of CD64 versus CD7 applied to the $CD34^{+}$ $CD33^{+}$ subpopulation from the upper right quadrant of the left plot. In all cases, cells are colored and labeled by the quadrants. (c) Six subpopulations (the only 6 populations having more than 20 cells each) revealed by comparisons such as that in panel b, were projected onto the viSNE map. Cells are colored by their respective subpopulation from b. The relapse sample can now be sorted into these subpopulations via fluorescence-activated cell sorting (FACS) and further studied through downstream experiments such as DNA and RNA sequencing.

2.2.8 viSNE detects minimal residual disease

The ability to detect, by flow cytometry, small numbers of cancerous cells displaying an aberrant phenotypic "fingerprint" is used to risk-stratify patients and direct treatment decisions. The presence of such minimal residual disease (MRD) can be associated with risk of relapse [61, 62]. The detection of MRD indicates a need for intensified therapy that unfortunately carries an increased risk of toxicity. Consequently, accurate detection of rare malignant populations is paramount in correctly assigning risk to an individual patient.

There are two competing manual methods for assessing MRD by flow cytometry. The first involves identifying aberrant antigen expression (leukemia-associated immunophenotype, or LAIP) on the leukemia cells at diagnosis, and then looking for that same phenotype on samples taken after chemotherapy [63]. The second method involves identifying leukemic cells based on a 'different-from-normal' phenotype by comparing to a historical bank of healthy bone marrow samples [64]. While the prognostic value of MRD measurement has been validated in several clinical trials, both of these approaches require an expert pathologist to manually inspect biaxial plots, and both approaches have shortcomings. It can be difficult to identify abnormal cells that are sufficiently phenotypically distinct from normal bone marrow. If one relies on and searches only for cells with the phenotype of the diagnostic sample, one may fail to detect other malignant populations displaying distinct yet abnormal phenotypes. Thus, a tool which automatically identifies abnormal cellular phenotypes would allow clearer identification and evaluation of remaining cancer cells.

Because viSNE revealed such a clear contrast between leukemic and healthy bone marrow, we tested whether viSNE could aid manual MRD detection. We spiked metal-barcoded [65] cells from an ALL patient sample into a healthy bone marrow sample, thereby creating a synthetic

sample with 0.25% of a MRD-like population. A single healthy bone marrow sample served as a guide for interpretation (similar to the "different-from-normal" manual approach). We used a biased subsampling method to enrich for unique non-control subpopulations and generated a viSNE map using eight markers (CD3, CD7, CD10, CD15, CD20, CD34, CD38, CD45) to emulate a MRD scenario using fluorescence-based flow cytometry (see Materials and Methods). The algorithm was blinded to the metal-barcoded channel.

A good MRD candidate region would be a distinct region of the viSNE map that contains cells from the MRD sample, but no cells from the healthy control sample. Cells from both samples were well-mixed across most of the viSNE map, except for one suspect region that was composed almost entirely from cells from the synthetic MRD sample (Figure 2-8a). We compared marker expression levels in the suspect region to the rest of the sample (Figure 2-8b) and found that the suspect cells strongly expressed CD10 and CD34, exhibited below-average expression of CD45 and also expressed CD15, a phenotypic combination often seen in B precursor ALL. Taken together, the combination of these surface markers and the absence of similar cells in the healthy control suggest that these were leukemic cells. Removal of the blinding of the metal-barcoding channel revealed that these cells were positive the metal-barcode and therefore were indeed the spiked ALL cells (Figure 2-8c). We repeated this analysis with a different set of markers and achieved similar results. As a control, we repeated the same procedure with an "MRD" sample that only included healthy cells. The two healthy samples were well mixed across the entire viSNE map; there was no region that contained only MRD cells, demonstrating that the subsampling method and viSNE do not spuriously create suspect regions. While only a synthetic example, this demonstrates viSNE's success in identifying a

minuscule cancer subpopulation in the data, suggesting that viSNE can be effectively used for MRD detection.

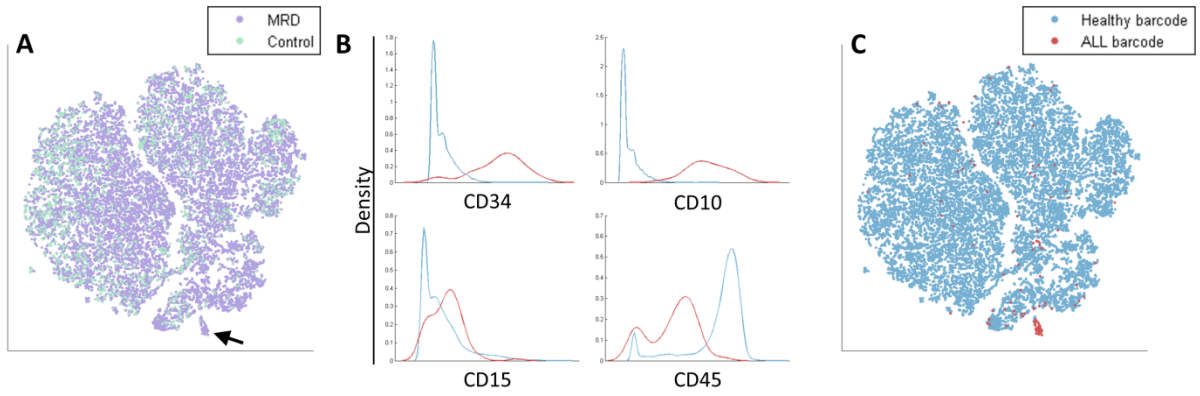


Figure 2-8. Using viSNE to identify synthetic minimal residual disease (MRD).
(a) A synthetic MRD sample was created by spiking a healthy bone marrow sample with metal-barcoded ALL cells. This synthetic MRD sample was compared to an unmanipulated healthy bone marrow sample (the viSNE algorithm was blinded to the metal-barcode channel). The viSNE map of the synthetic MRD sample (purple) and a healthy control sample (cyan) includes a suspect region (marked by an arrow) composed almost entirely of cells from the synthetic MRD sample (purple). (b) Expression of indicated markers on cells in the suspect region (red) and the non-suspect region (cyan). X-axis represents marker expression level and y-axis represents density of cells. (c) The viSNE map from panel a, after removing the blinding of the metal-barcoding channel. The map is color according to expression of the metal barcode in the spiked-in ALL cells (red). The suspect region is indeed almost entirely composed of ALL cells. ALL cells outside of the suspect region have marker expression levels conforming to healthy cells.

2.3 Comparison of viSNE to other methods

viSNE belongs to the class of nonlinear dimensionality reduction (NLDR) algorithms which, unlike principal component analysis (PCA), do not assume linear relationships between parameters. Immune subsets are nonlinear and hence PCA, unlike viSNE, fails to separate between them (Figure 2-9). We evaluated three other NLDR algorithms [66]: Isomap [44], LLE [50] and Kernel PCA [47] (Figure 2-9). Out of the three, both LLE and Kernel PCA collapsed all of the data points into a single region (in the case of LLE) or into three to five diagonals (in the case of Kernel PCA). Both of these phenomena have been described before [25] and are caused by the structure of the data being more complex than the single manifold assumed by these

methods. Isomap is the only other method that managed to separate between the different immune subtypes. However, the separation is weaker than viSNE: the two T-cell subtypes overlap, and the unclassified cells are a smudge that covers most of the low-dimensional space. Furthermore, Isomap suffers from lack of robustness, as different runs lead to qualitatively different outputs. Additionally, these methods might be confounded by the noise inherent in biological systems and measurement technologies or by the complex geometry in high-dimensional hematopoietic space.

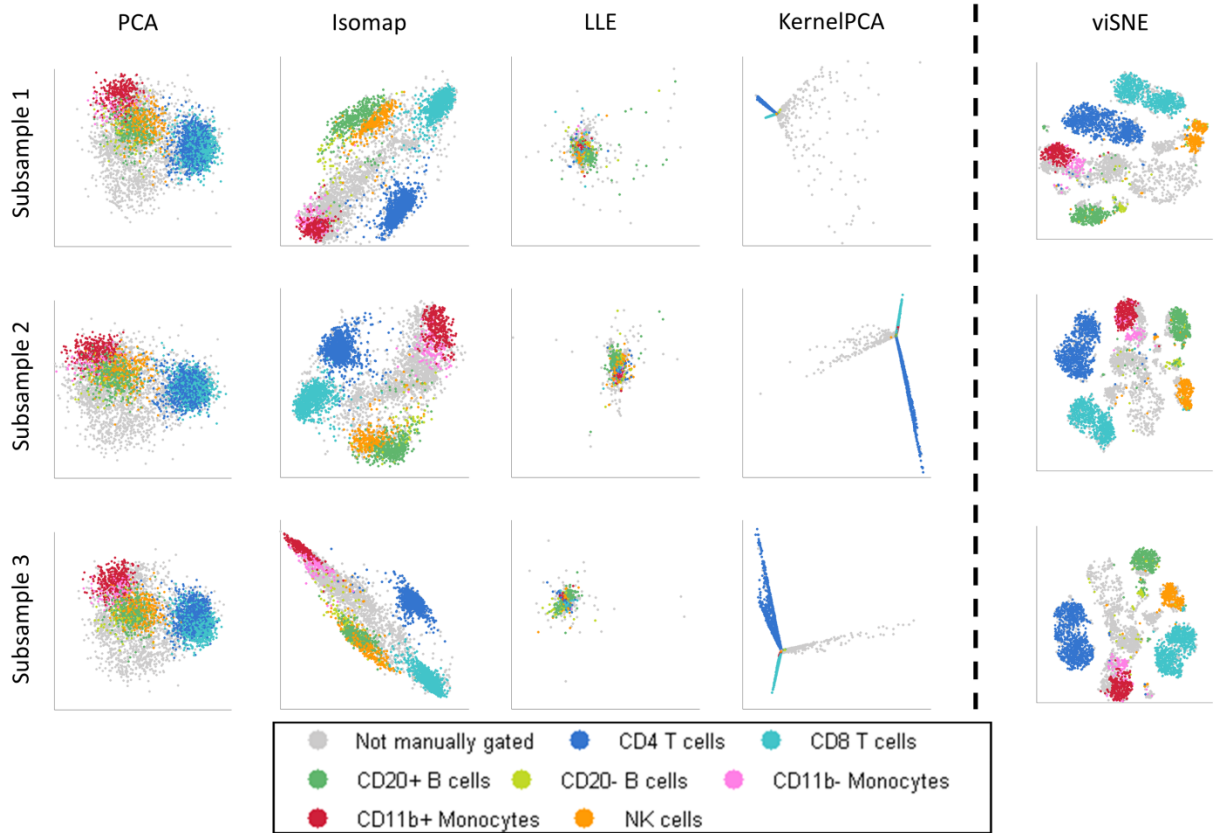


Figure 2-9. Comparison of four dimensionality reduction algorithms (PCA, Isomap, LLE and Kernel PCA) to viSNE over three subsamples of Marrow1.
Cells are color coded by immune subsets, as in Figure 2-1b. We found similar results for additional NLDR algorithms from the toolkit [25].

Another method that has been used in the context of mass cytometry is SPADE [12, 56]. SPADE is a clustering-based method for extraction of cellular hierarchy from single-cell data. However,

SPADE suffers from two flaws when compared to viSNE. One, SPADE begins by down-sampling the data using a density-based, non-uniform random process. The algorithm is not robust to this down-sampling step, and as a result multiple runs can lead to radically different outputs (Figure 2-10). Other than prior knowledge, there is no way to judge which SPADE run leads to the correct interpretation of the system, a significant disadvantage when applying the algorithm to a less-familiar system. Two, SPADE's second step is clustering the data, losing the single-cell resolution. This clustering step is not sensitive to rare cell populations; for example, when we applied SPADE to the same synthetic MRD sample as viSNE, the ALL cells were indistinguishable from healthy cells in the resulting SPADE tree (Figure 2-11), rendering SPADE inappropriate for MRD.

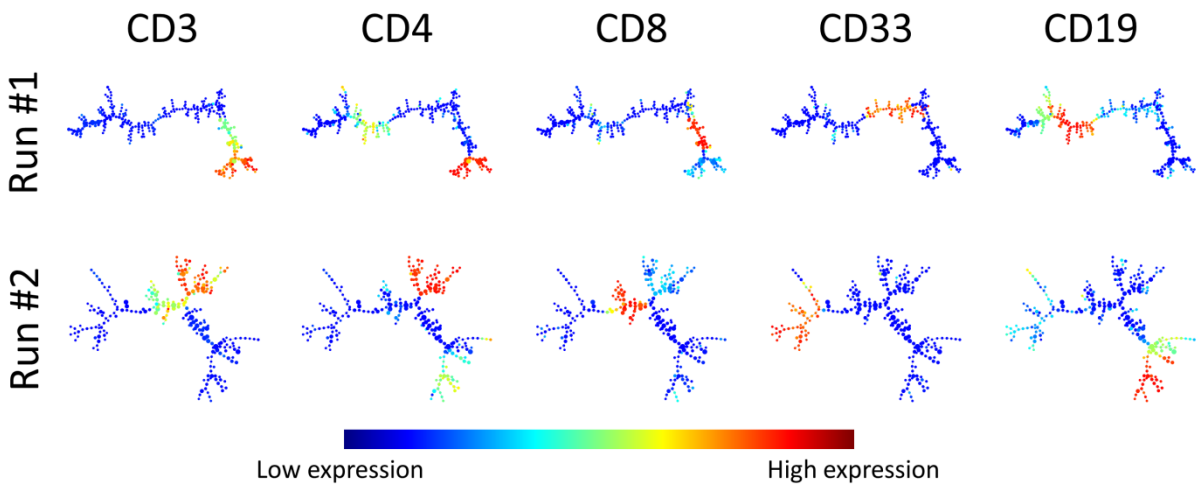


Figure 2-10. Two SPADE runs of Marrow1, colored by mean marker expression levels for each cluster. In this plot each point is a cluster of cells and edges represent the minimal spanning tree identified by SPADE. Each row represents one of the two SPADE runs and each column representes the marker whose mean value was used to color the clusters. Clusters group by immune subtype, but longer range distances are less conserved between runs, resulting in considerable differences between the SPADE tree from the two runs.

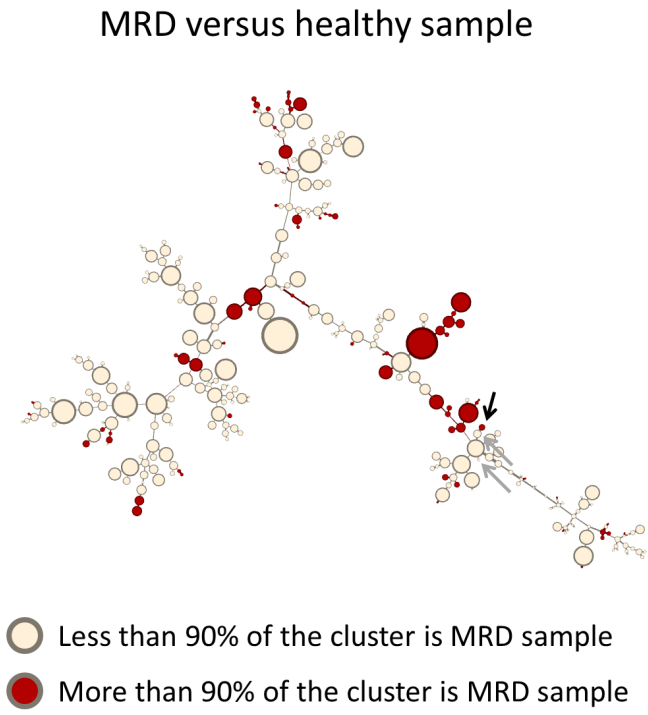


Figure 2-11. SPADE was applied to the same synthetic MRD sample used in Figure 2-8. The ALL-barcoded cells (the target cells) are spread over 3 different clusters (marked with arrows); the cluster with the highest percentage of ALL-barcoded cells is marked with a black arrow- approximately 90% of its cells are from the MRD sample (the remaining 10% are from the healthy control sample). However, there are 73 clusters that fit the 90% MRD criterion (red clusters), and the 3 ALL-barcoded-rich clusters do not stand out relative to the other 72 high-MRD clusters or the SPADE tree as a whole.

viSNE has a number of advantageous features for the analysis of single-cell data which lead it to succeed where the other methods failed. viSNE models the local neighborhood by examining all pairwise pro-Babilities, not just a subset that is defined by a user-provided threshold. This results in much better separation than other methods, especially in the case of similar cell types like $CD4^+$ and $CD8^+$ T cells. Additionally, viSNE has been designed with the curse of dimensionality in mind. As a result, the choice of a t-distribution makes the algorithm much more sensitive to small subsets, allowing us to subsample uniformly (thus preserving the original frequencies of cell populations), and leading to the detection of rare populations such as in the MRD settings. Finally, viSNE does not cluster the data, keeping the single-cell resolution and allowing us to visualize each individual cell. We can visualize marker levels, including signaling markers, on top of the viSNE map. viSNE's design philosophy is unique in the field of nonlinear dimensionality reduction, making it the best current choice for analysis of high-dimensional single-cell data.

2.4 Discussion

viSNE allows visualization of high-dimensional single-cell data on a two-dimensional map. This mapping takes advantage of the inherent structure of the data; for example different immune subsets reside in separate regions in high-dimensional space. Conventional analysis of cytometry data, which views only two dimensions at a time, ignores the higher-order structure and complex relationships between markers in the data. Whereas viSNE plots resemble conventional biaxial plots, their utility comes from combining and representing information from all dimensions simultaneously.

We found that the viSNE map is consistent across multiple healthy individuals, while cancer samples occupy regions distinct from healthy cells and from each other. We illustrated how viSNE can be utilized to characterize heterogeneity within cancer samples, mark disease progression from diagnosis to relapse, and identify rare cancer populations lurking among predominantly healthy cells.

Despite its extensive utility, as with all dimensionality reduction tools viSNE is inherently limited: low-dimensional mapping cannot represent all of the information in high-dimensional space. Therefore, viSNE only captures the most dominant structures. One way to gain more detail is to run viSNE on a well-defined subset of the data. For example, instead of analyzing several cancer samples together (Figure 2-3d), one can run viSNE on each sample separately (Figure 2-5a-b). Alternatively, it is possible to limit the mapping to only a subset of parameters of interest. Another consequence of dimensionality reduction is the "crowding pro-Blem", which typically limits the number of cells we can map to 30,000. This limits applications such as gating, since a subsampling of the cells is required, meaning only cells in the subsample can be classified. An effective solution for gating is combining viSNE with a clustering algorithm. We clustered the cells using FLOCK [57], a state-of-the-art clustering tool for cytometry data, and labeled the viSNE map according to this clustering (Figure 2-12). While FLOCK separated the immune subtypes, it splits each subtype into multiple clusters. The viSNE map helps interpret these clusters and their relationships.

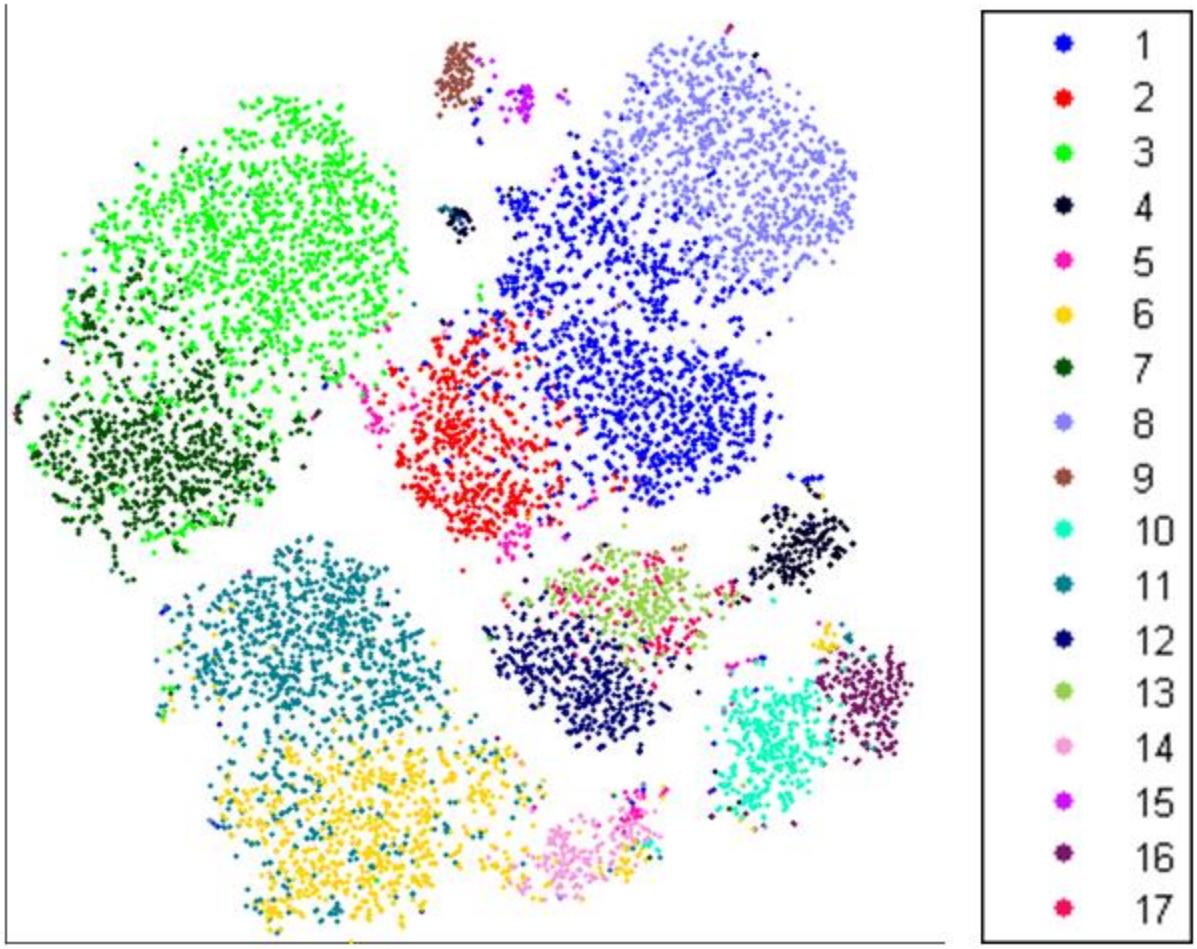


Figure 2-12. FLOCK clustering of mass cytometry data [57]**, as visualized by viSNE, each cell is colored by its cluster id. FLOCK separates the major subtypes.**
However, it suffers from over-clustering and breaks most cell subtypes into multiple clusters. The viSNE maps helps regroup these back together for interpretation.

viSNE is an unsupervised algorithm and does not require prior knowledge of the system. It is thus suitable for navigating less explored systems such as cancer. While structure in healthy samples is formed through an orderly program of development, cancer's derailed developmental program leads to loss of normal order and structure. viSNE helps characterize the plethora of abnormal phenotypes unique to each cancer by exploiting its ability to take all markers into account simultaneously, rather scanning through hundreds of biaxial plots, two markers at a time.

A characteristic feature that repeated across multiple cancer maps was the emergence of distinct gradients of marker expression levels that resemble developmental progression in healthy cells (Figure 2-1d). Comparing gradients in AML diagnosis and relapse samples (Figure 2-6b) supports the notion that the cells first gain CD34 and subsequently the cells acquire a large diversity of abnormal combinations of lineage-specific markers without attenuation of CD34. After identifying unexpected cancer populations using viSNE, one can design a sorting strategy for physical isolation and downstream characterization of these populations.

In the future, we expect viSNE to be instrumental in the analysis of mass cytometry data integrating the surface marker panel with a panel of functional markers that pro-Be signaling, cell cycle and metabolism, under many experimental perturbations (such as cytokines and drugs) [3]. In this scenario, viSNE's ability to distinguish rare subsets which comprise only a tiny fraction of the population (Figure 2-8) could be advantageous toward the identification and characterization of rare drug resistant subpopulations.

We demonstrated viSNE's capability to analyze mass cytometry and flow cytometry data. Biological research is trending towards dozens of dimensions in tens of thousands of cells. Making sense of these data is a daunting challenge that requires powerful computational approaches. The utility of viSNE will increase with the number of dimensions capable of being analyzed by mass cytometry and other technologies.

2.5 Materials and Methods

2.5.1 The t-SNE algorithm

The t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm maps points from high-dimensional space to low-dimensional space by minimizing the difference in all pairwise

similarities between points in high- and low-dimensional spaces [67]. The axes of the low-dimensional spaces are given in arbitrary units. The algorithm proceeds as follows.

t-SNE's input is a list of points in high-dimensional space, X. The algorithm begins by calculating the pairwise similarity matrix in high-dimensional space, P, and randomizing a starting position for each point in the low-dimensional space, Y_0. t-SNE proceeds to iteratively update the position of points in low-dimensional space: in iteration i, the similarity matrix in low-dimensional space, Q, is calculated according to the points' current positions (Y_{i-1}). Gradient descent is used to calculate the new position of each point, Y_i, in order to minimize the divergence between P and Q.

For each point x_i in high-dimensional space, t-SNE defines the similarity of x_i to x_j as defined below:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\Sigma_{k \neq i} \exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)} \text{ (Equation 1)}$$

σ_i is x_i's variance. For each x_i, t-SNE performs a binary search for the value of σ_i that produces a P_i with a fixed Perplexity (a parameter for the algorithm that is given by the user; an intuitive interpretation for the perplexity is a soft measure for the number of nearest neighbors to consider for each cell). The Perplexity is defined as:

$$Perp(P_i) = 2^{H(P_i)} \text{ (Equation 2)}$$

where $H(P_i)$ is Shannon's entropy:

$$H(P_i) = -\sum_j p_{j|i} log_2 p_{j|i} \text{ (Equation 3)}$$

The joint similarity of x_i and x_i is defined as:

$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$ (Equation 4)

For each pair of points in low-dimensional space, y_i and y_j, the similarity is defined as:

$q_{ij} = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\Sigma_{k \neq i}(1+\|y_i - y_j\|^2)^{-1}}$ (Equation 5)

While p_{ij} follows a Gaussian distribution, q_{ij} is calculated using a t-distribution.

t-SNE minimizes the Kullback-Leibler (KL) divergence between the joint pro-Bability distribution P (in the high-dimensional space) and the joint pro-Bability distribution Q (in the low-dimensional space). The KL divergence is defined as:

$KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$ (Equation 6)

The gradient of the KL divergence between P and Q is derived in [67]:

$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$ (Equation 7)

The optimization step may be interpreted as a set of springs. Each pair of points Y_i and Y_j is connected by a spring which repels or attracts the points from each other depending on whether the similarity between the points in the projection is lower or greater than the similarity in the high-dimensional space. The gradient reduces each point's springs into a single force. The heavy tailed t-distribution helps alleviate the "crowding pro-Blem" by exerting more force when pushing distant points further apart.

2.5.2 The viSNE implementation

viSNE is a distributed implementation of t-SNE that relies on the locality of t-SNE's calculations. Given N cores, viSNE splits the input into partitions where each partition contains

n/N points. Each core receives one partition. Instead of storing the entire matrix (n^2 values), each core only stores the submatrix consisting of its n/N points, times all other points (n^2/N values). For example, for n = 100,000 and N=64 cores, each core will need to store approximately 150 million values per matrix. For each iteration, each core locally updates Y_i for points in its partition and broadcasts these values to the other cores, guaranteeing that all cores have the updated similarity matrix.

The robustness and accuracy of t-SNE derives from the computation of all pairwise similarities. But, the similarity matrix comes at a heavy computational price, limiting the original implementation to 10,000 points. Our distributed implementation relies on the fact that each of t-SNE's computations are local and do not require the entire matrix. The technical computational limit of viSNE is 100,000 points. However, beyond 30,000 the limit is not computational, but rather the "crowding pro-Blem" [67]: the volume in high-dimensional space grows polynomial with the number of dimensions, and as a result a two-dimensional map cannot accommodate a large number of points while conserving the high-dimensional distances between them. Instead, distant points collapse onto nearby areas of the map, creating one large, dense region, with no separation between populations. To solve this, viSNE subsamples cells uniformly at random and maps the sampled population. The algorithm is robust to such subsampling and even after subsampling, we still detect rare subpopulations that constitute a mere 0.2% of the population.

2.5.3 The cyt visualization tool

cyt is an interactive visualization tool designed for the analysis of viSNE maps and the high-dimensional mass or flow cytometry data from which these maps were projected. It plots viSNE maps as scatter and density plots, and information can be overlaid onto this map by coloring cells according to various parameters, such as marker expression, source of sample or subtype. *cyt*

includes a gating feature that can be used with either biaxial plots (to generate a viSNE map on only a defined subset of the cells) or the viSNE map (to further study a population identified by viSNE). This is enabled by *cyt*'s modular design: once a gate is created it can be treated as an independent dataset and all of *cyt*'s features can be applied. The gates can be compared on a marker-by-marker basis using one-dimensional density plots, and *cyt* prioritizes the markers according to the L1 distance between marker distributions. This method quickly identifies key differences between populations. The combination of viSNE and *cyt* facilitates efficient examination of mass and flow cytometry data.

cyt contributes in correlating multiple viSNE maps of the same data. While viSNE consistently separates the various immune subtypes, their position on different maps could vary (most often due to rotations and reflections of the map). In healthy samples, we initially colored the immune subtypes using *cyt*, helping us label each sub-population and therefore compare between different viSNE maps of similar data. While the maps can vary in rotation and reflection, the actual population structure is preserved and *cyt* can be used for re-orientation. For cancer samples, which lack distinct subtypes, *cyt* lets us quickly identify which populations are similar to each other between multiple maps of the same sample based on their marker combination. *cyt* presents the data in an intuitive visual manner that allows the user to corroborate the viSNE maps.

Alternatively, to get multiple samples projected onto the exact same map, we can run a number of samples together in a single run of viSNE. This approach was used to generate Figures 2-3b and 2-3d. *cyt* can then be used to split these into multiple maps, one for each sample, that share coordinates (Supplementary Fig. 5). The advantage of running multiple samples together is one of the main reasons that the ability to run on more cells is an important feature of viSNE.

2.5.4 Mass cytometry data

Fresh, Ficoll-enriched human bone marrow was obtained from healthy donors from AllCells, Inc. (Emeryville, CA). Samples were obtained with informed consent in accordance with the Declaration of Helsinki and with accordance with Stanford University's review board. Leukemia bone marrow samples were obtained under IRB-approved protocols (protocol number 17552 under Stanford University's IRB) at St. Jude Children's Research Hospital, Memphis, TN (pediatric acute myeloid and lymphoblastic leukemia) or at Princess Margaret Hospital, Toronto, ON (adult acute myeloid leukemia). All samples were deidentified. The age and the sex of the donor, or any additional clinical information, were unknown at the time of the study.

Samples were processed as described in Bendall et al [12]. Briefly, cells were used fresh prior to mass cytometry experiments, or frozen in FCS with 10% DMSO, thawed and re-suspended in 90% RPMI with 10% FCS (supplemented with 20 U/mL sodium heparin (Sigma) and 0.025U/mL benzonase (Sigma) in the case of frozen samples), 1X L-glutamine and 1X penicillin/streptomycin (Invitrogen).

Cells were fixed with formaldehyde (PFA; Electron Microscopy Sciences, Hatfield, PA) added directly to growth media at a final concentration of 1.6% for 10 minutes at room temperature. Cells were then centrifuged at 500g for 5 minutes and washed once with staining media (PBS with 0.5% BSA, 0.02% sodium azide) to remove residual PFA, and blocked with Purified Human Fc Receptor Binding Inhibitor (eBioscience Inc., San Diego, CA) following manufacturer's instructions. Surface marker antibodies were added yielding 50 or 100 uL final reaction volumes and stained at room temperature for 30min (Table 2-2). Following staining, cells were washed 2 more times with cell staining media, permeabilized with 4°C methanol for at 10 min at 4°C, and then optionally stored at -80°C for later use. Cells were then washed twice in

cell staining media to remove remaining methanol, and stained with surface and phospho-specific antibodies in 50 or 100 μL for 30 min at room temperature. Cells were washed once in cell staining media, then stained with 1 mL of 1:5000 191/193Ir DNA intercalator(2) (www.dvssciences.com; DVS Sciences, Richmond Hill, Ontario, Canada) diluted in PBS with 1.6% PFA for 20 min at room temperature. Cells were then washed once with cell staining media and then finally with water alone before running on the CyTOF.

Target	Clone	Metal Isotope	Staining Concentration (ug/ml)
CD10	HI10a (BL)	**156 Gd**	1
CD114	LMM741 (BL)	**156 Gd**	2
CD117	104D2 (BL)	**171 Yb**	1
CD11b	ICRF44 (BL)	**144 Nd**	3
CD11c	3.9 (BL)	**154 Sm**	5
CD123	9F5 (BD)	**151 Eu**	3
CD13	L138 (BD)	**168 Er**	1
CD133	AC133 (MB)	**141 Pr**	3
CD14	M5E2 (BL)	**160 Gd**	2
CD15	W6D3 (BL)	**164 Dy**	2
CD16	3G8 (BL)	**165 Ho**	2
CD161	HP-3G10 (BL)	**150 Nd**	5
CD179a	HSL96 (BL)	**149 Sm**	2
CD19	HIB19 (BD)	**142 Nd**	1.5
CD2	TS1/8 (BL)	**152 Sm**	0.5
CD20	2H7 (BL)	**147 Sm**	3
CD20 cytosolic	H1 (BD)	147 Sm	3
CD22	HIB22 (BL)	**168 Er**	1
CD235a/b	HIR2 (BL)	**141 Pr**	2
CD3	S4.1 (QD)	**110 Cd**	1:200
CD3	UCHT1 (BL)	**170 Er**	0.5
CD33	P67.6 (BD)	**158 Gd; 173 Yb**	1.5
CD34	8G12 (BD)	**148 Nd**	3
CD38	HIT2 (BL)	**159 Tb; 168 Er**	1
CD4	RPA-T4 (BL)	**145 Nd**	3
CD41	HIP8 (BL)	**152 Sm**	1
CD44	G44-26 (BD)	**166 Er**	1
CD45	HI30 (BL)	**115 In; 154 Sm**	2
CD45RA	HI100 (BL)	**139 La**	3
CD47	B6H12 (BD)	**145 Nd; 172 Yb**	2
CD49d	9F10 (BL)	**144 Nd**	1
CD5	UCHT2 (BL)	**154 Sm**	1
CD56	B159 (BD)	**170 Er**	2
CD61	VI-PL2 (BD)	**169 Tm**	0.25
CD64	10.1 (BL)	**153 Eu**	1
CD7	M-T701 (BD)	**167 Er**	2
CD79b	CB3-1 (BL)	**146 Nd**	2
CD8	RPA-T8 (BL)	**146 Nd**	1.5

CD90	5E10 (BL)	**176 Yb**	5
CXCR4	12G5 (BL)	**175 Lu**	3
Flt3	BV10A4H2 (BL)	**150 Nd**	2
HLA-DR	L243 (BL)	**174 Yb**	2
IgD	IA6-2 (BL)	**145 Nd**	1
IgM	G20-127 (BD)	**153 Eu**	2
pre-BCR	HSL2 (BL)	**165 Ho**	3
TdT	E17-1519 (BD)	**151 Eu**	3
TIM-3	344823 (RD)	**169 Tm**	1

Table 2-2. Antibody sources, metal isotope and staining concentration for all of the antibodies used throughout the various experiments.
Vendors: Invitrogen Qdot655 (QD); Biolegend (BL); BD Biosciences (BD); Miltenyi Biotec (MB); R&D Systems (RD); DVS Sciences (DVS); VWR International (VWR); eBioscience (EB); Cell Signaling Technologies (CST).

2.5.5 Processing of mass cytometry data

Data was transformed using hyperbolic arcsin with a cofactor of five. Single cells were gated based on cell length and DNA content (to avoid debris and doublets) as described in Bendall et al. [12]. The expert manual classification of Marrow1 was taken from [12], where the complete biaxial plot gating strategy can be found.

2.5.6 viSNE analysis

Generating viSNE maps included the following steps (exact details can be found in Table 2-1). First, between 6,000 and 12,000 cells were uniformly subsampled from the data. After subsampling, viSNE was run for 500 iterations to project the data into 2D. Unlike t-SNE, PCA was not used as a preprocessing step. All runs used an identical random seed and the default t-SNE parameters (perplexity = 30, momentum = 0.5 for initial 250 iterations, momentum = 0.8 for remaining iterations, epsilon = 500, lie factor = 4 for initial 100 iterations, lie factor = 1 for remaining iterations). viSNE maps were visualized using *cyt*, which was also used to generate figures (color coding by immune cell subset (as in Figure 2-1b), by marker expression levels (as in Figure 2-1d) and in plotting expression level densities (as in Figure 2-3b).

2.5.7 Quantifying similarity between viSNE maps

We use the Jensen-Shannon (JS) divergence to quantify the similarity between viSNE maps. Each map is converted into a pro-Bability distribution. We define the similarity between two maps as the JS divergence between their respective distributions:

$JS(P||Q) = (KL(P||M) + KL(Q||M))/2$ (Equation 8)

Where M is:

$M = (P + Q)/2$ (Equation 9)

And KL is the Kullback-Leibler divergence:

$KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$ (Equation 10)

The JS divergence has a value of between zero and one. When JS(P||Q)=0, the pro-Bability distributions are identical. When JS(P||Q)=1, there is no overlap in the information encoded by P and Q.

2.5.8 A gating scheme for fluorescence-activated cell sorting

The viSNE map was utilized to devise a gating scheme for fluorescence-activated cell sorting (FACS) of the AML relapse sample (Figure 2-7). Due to the limits of flow cytometry, the gating scheme can only employ a limited number of channels and use hard thresholds. Through manual inspection of the viSNE map we identified four markers that lead to distinct subpopulations which could be of interest for downstream analysis: CD34, CD64, CD33 and CD7 (Figure 2-6). For each marker we defined a threshold for a binary negative/positive gate. The four binary gates were combined to create a total of sixteen composite gates covering all negative/positive

combinations. Only 6 composite gates had more than 20 cells. The cells residing in each of these 6 composite gates are color coded on the viSNE map of Figure 2-7.

2.5.9 Subsampling of synthetic MRD sample

We used two samples: the synthetic MRD sample (composed of 99.5% healthy bone marrow cells and 0.5% cells from a metal-barcoded ALL sample) and the control sample (100% healthy bone marrow cells taken from a different donor). To capture a higher proportion of ALL cells for the viSNE map, we devised the following subsampling procedure. The cells from the synthetic MRD sample and from the control sample were combined computationally and clustered using the Louvain algorithm [68]. Next, the clusters were weighted by the proportion of synthetic MRD sample cells in them; the higher the proportion of synthetic MRD sample cells, the higher the weight. Finally, 10,000 cells were chosen one at a time in a two-step process: one of the clusters was chosen randomly (biased by cluster weight) and a cell was uniformly chosen from that cluster. Note, the subsampling procedure is blind to the metal barcode; it can only access the mass cytometry measurement and the identity of the sample (synthetic MRD or control). Following the subsampling, viSNE was run as described above.

2.5.10 Additional algorithms

Isomap, LLE, KernelPCA and LLTSA were run using the Matlab Toolbox for Dimensionality Reduction [25]. FLOCK was compiled from the code available in the ImmPort FLOCK SourceForge page (http://sourceforge.net/projects/immportflock/). SPADE was run using the implementation available in Cytobank [69].

Chapter 3 The Wanderlust algorithm for trajectory detection

3.1 Introduction

3.1.1 The developmental trajectory

Given a healthy bone marrow sample composed of B-lineage cells, the *developmental trajectory* is the ordering of cells according to their developmental chronology. We can use the trajectory as scaffolding, upon which we can infer the order of key molecular and cellular events during development. For example, the transition between stages will be seen as a decrease in the levels of early-stage markers and an increase in later-stage markers. Proliferation, apoptosis, and the signaling involved in regulating development will all be reflected in changes in respective marker levels. The trajectory can also serve to characterize the timing of poorly-understood markers by examining changes in their expression relative to better understood markers.

There are several challenges underlying trajectory detection. These will be briefly described here and fully explored in the following paragraphs. First and foremost, the data is rife with statistical noise from multiple sources, both biological and technical; the most detrimental effect of this stochastic variation is the creation of *short circuits*- pairs of cells which are developmentally distant but close in the space of measured parameters. Second, the data involves several rare sub-populations that comprise a tiny portion of the data (often less than 1%); many approaches might incorrectly treat these as outliers, losing crucial regions of the trajectory. Third, the data is high-dimensional, thus immediately invalidating any manual-examination based approaches, since we

cannot physically visualize all of the required dimensions; additionally, computational methods should be able to scale to the large number of dimensions and cells in this data. Fourth, marker levels rise and fall in a coordinated fashion and their multivariate relations cannot be faithfully captured in a linear model, hence a successful computational approach must be non-linear in its nature. Fifth, cell size influences the measurement of all other measured parameters, leading to the appearance of spurious relationships in the data. We now elaborate on each of these points.

Biological data includes statistical noise from multiple sources, which can be broadly classified into two categories. The first, stochasticity, is the noise inherent to the biological system due to the small quantities of the molecules involved (sometimes as low as only a few copies of a molecular epitope per cell [70, 71]). The second, technological noise, is caused by imprecisions and limitations of the technology used. While some of it can be identified and compensated for via pre-processing, there is much variation that cannot be accounted for.

Noise raises a variety of pro-Blems for analysis and method development. For example, cells that are developmentally close might have different measurements. Vice versa, developmentally distant cells might appear similar to each other, a phenomenon we call a *short circuit* (figure 3-1a). A third implication, which is common to antibody-based technologies, is scaling: since antibodies have different binding affinities and molecular properties, antigen measurements can only be given in arbitrary units; it is impossible to compare exact amounts and the magnitude is different between antigens and between samples.

Another challenge is that some rare cell stages, such as stem cells, comprise less than a tenth of a percent of the population. Computational methods might classify such cells as outright outliers and remove them in their entirety from the analysis due to their low counts. Furthermore, a large

sample comprising hundreds of thousands of cells is required to guarantee that enough cells from each rare population are present, necessitating a computationally efficient algorithm to find the trajectory in a reasonable amount of time.

Another complication is the influence of cell size on marker measurements. Biological processes are concentration-based and therefore larger cells usually have more copies of required molecules. Without a reliable indication of cell size, it is impossible to differentiate between a large cell with relatively few copies of a molecule and a small cell with a high concentration of it, although the two might differ functionally. Two final, interwoven complications are the high-dimensionality of the data and the non-linear relationships between parameters (figure 3-1b). The large number of dimensions necessitates a computationally light method in order to receive results in a reasonable amount of time. However, the non-linear models required to correctly model the system are often complex and computationally intensive.

In summary, a successful method should be robust to the noise in the data, identify the rare cell subsets, consider relative rather than absolute marker amounts and accommodate, the nonlinear relationships between dimensions, all the while running in a reasonable computational time.

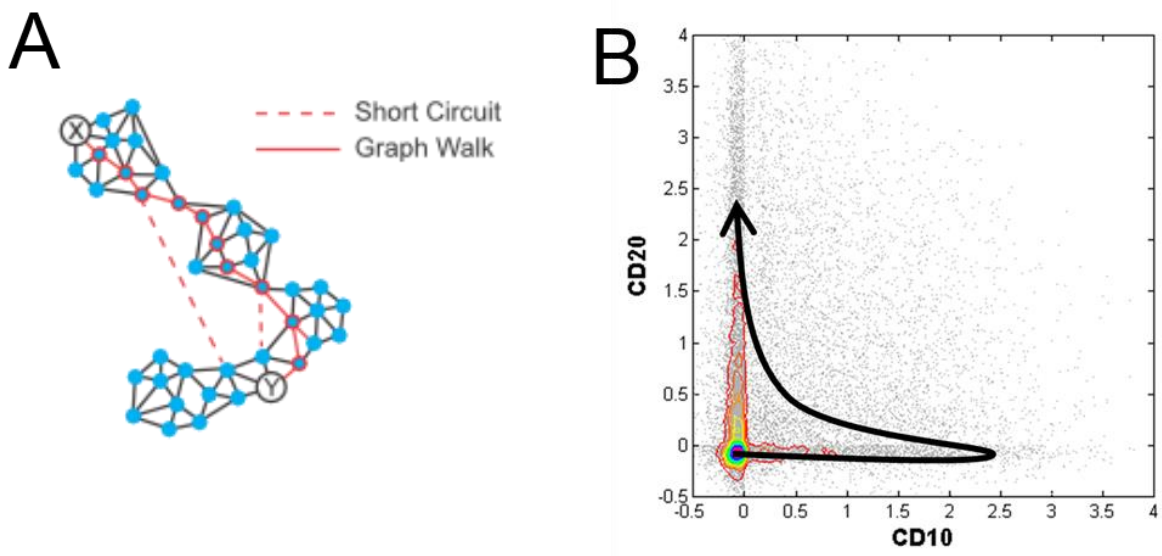


Figure 3-1. Non-linear relationship between markers.
(a) A toy example showing a short circuit. Black edges are biologically correct edges. The dashed red edge is a short circuit caused by a short distance in high-dimensional space due to noise. The correct graph walk (in red) will be superseded by the short circuit. (b) CD10 (x-axis) versus CD20 (y-axis) in B lineage cells. Grey dots represent cells. The contour plot is the kernel-smoothed estimation of the density. The black line shows the developmental process: $CD10^{-}CD20^{-}$ stem cells are followed $CD10^{+}$ progenitors; as cells mature, CD10 decreases while CD20 goes up. This process cannot be modeled by a linear relationship between CD10 and CD20.

3.1.2 Overview of existing methods

While a method that explicitly detects a trajectory does not exist, several computational methods rise as promising candidates for this task. PCA [29], Isomap [44], Kernel PCA [47] and other dimensionality-reduction methods have been covered in chapter 2; however, as described there, these either assume a linear relationship between parameters or do not scale to a large enough number of cells. Furthermore, none of them directly address the task of detecting a trajectory. Likewise, SPADE [56], which was also discussed in chapter 2, is non-robust and does not scale to the necessary number of cells; it also involves subsampling and up-sampling steps, both of which lose information and distort the data.

Ergodic rate analysis (ERA) [72] is a method designed for calculating the rate of molecular events based on single-cell measurements. ERA involves a density-based dimensionality-

reduction step that could conceptually be applied to trajectory detection. However, the integration that underlies this step assumes that the number of cells of a given stage is proportional to the amount of time spent in that stage. This ergodic assumption is broken in the case of a developmental system, where proliferation rates vary, rapid proliferation is followed by rapid cell death, and where cell subtype proportions are influenced by many factors (such as genetics and exposure to pathogens). Furthermore, ERA involves the calculation of the density function underlying the data, an operation whose computation speed scales exponentially with the number of dimensions. We note that ERA has only been successfully applied to a maximum of four dimensions.

Projection pursuit (PP) [73, 74] is an exploratory analysis method that aims to find interesting low-dimensional views (linear projections) of the data. The experimenter then visually examines the projections in an attempt to find those which include a relevant signal. PP requires a function *I*, called a projection index. *I* quantifies the how interesting a projection is. The algorithm then samples projections of the data and finds a set of projections with high *I*. Several papers explore possibilities for *I* and the sampling process [75-78], and the algorithm has been applied in contexts such as chemistry [79, 80], ecology [81], meteorology [82], and biology [83, 84]. However, PP suffers from three major shortcomings: first, it is based upon linear projections and as such cannot handle the non-linear relationship in developmental data; second, the algorithm is sensitive to the choice of projection index and sampling method [74]; and third, the algorithm is computationally intensive, as evidenced by the small number of data points (in the low hundreds) in all of the studies described.

3.2 Results

3.2.1 A graph-based approach to trajectory detection

To develop an appropriate trajectory detection algorithm, we make three assumptions about the data. First, a large healthy bone-marrow sample taken at a single time point will include cells throughout the entire continuous developmental process, including intermediate and rare cell populations. Second, B cell development in the marrow is non-branching and linear: cells can either proceed along development or undergo apoptosis. Third, B cell development is continuous; changes in protein expression are gradual, and therefore the transitions between stages are gradual. Given these assumptions, we can measure the levels of all of the markers needed to identify the different stages in millions of cells and should then be able to trace the trajectory.

Based on these assumptions, we propose a graph representation of the data as a conceptual framework for a trajectory detection algorithm. We convert the data into a k-nearest neighbor graph (k-NNG). Each cell is represented as a node and is connected to its k neighbors, the cells most similar to it, via an edge whose weight is set by the similarity. We define the shortest-path distance between a pair of cells as the length of the path between the nodes that minimizes the sum of weights of its constituent edges. The shortest-path distance is composed of transitions through neighbors, where each transition is a small gradual step. A conceptual design for a trajectory detection algorithm would be to start from the earliest cell in the data and order the rest of the cells according to their shortest-path distance from that earliest cell. Based on our third assumption this approach provides an approximation to the developmental order between cells.

The graph representation addresses several of the data's pro-Blems. The conversion into a graph is cheap, and shortest-path calculations are fast due to the graph's sparsity; this enables the analysis of large datasets in reasonable time [85]. Our model is based on similarity between cells rather than on relationships between parameters and can therefore handle non-linearity. The magnitude of the noise is proportional to the distance and since the local neighborhood is based on short distances this approach is less susceptible to noise (since short distances are reliable, and our representation of long distances as graph traversal mitigates noise from long distances). Finally, even if two developmentally close cells are not direct neighbors they still reside in the same region in the high-dimensional space. Therefore, they will be separated by a small number of close neighbors and the shortest-path distance between them will be low, circumventing much of the unexplained variability in the data.

However, the k-NNG is still susceptible to short circuits as such cells might be connected via an edge. The shortest-path between developmentally distant cells will go through the short circuits, leading to incorrect distances. A possible solution for this pro-Blem is to use a random-walk based distance measure. However, random walks are computationally intensive as they go both towards and away from the target and are not practical in such a large graph. Instead, we fight fire with fire and utilize randomness to address the noise in the data. Short circuits are rare; therefore, a random subset of the graph is likely to include only a few short circuits. We extend the graph representation into an ensemble of l-out-of-k-nearest neighbor graphs (l-k-NNG)[1]. An l-k-NNG is generated by starting with the k-NNG and iterating over each node in the graph, randomly keeping only l of its k-nearest neighbors. On average, a given short circuit will only exist in 2l/k of the graphs (the pro-Bability that that specific edge will be one of the l edges

[1] Please note that l stands for lowercase l, not for the digit 1.

chosen out of the k edges, for each of the two nodes that the edge connects); in these graphs the shortest-path distances will be distorted by that short circuit. By picking l lower than k, each short circuit only appears in a few l-k-NNGs and influences a different set of cells. We can average out its effect by taking the mean over all graphs.

Shortest-path distances raise two complications. First, distances do not have a direction while the trajectory does. In order to address this issue, we require a user-defined early cell. The early cell is assumed to reside toward the beginning of the trajectory and is used for orientation. Second, the shortest-path distance variability increases with the distance. As nodes get farther from each other, the accuracy of their short-path distance decreases, since mistakes accumulate. Therefore, the distance between two distant cells is less reliable than between two close cells. We utilize landmarks to support our trajectory by breaking it into shorter distances. We randomly flag a small subset of as landmarks, following a uniform distribution. In the naïve approach only the early cell was used to decide on each other cell's position. The landmark cells serve as reinforcements to the early cell: the position of each cell will now be calculated as the average of its distance from all of the landmarks. Additionally, we weigh the contribution of each landmark according to its distance from the cell, further reinforcing the influence of short distances and reducing the influence of long distances. Since at least some of the landmarks will be close to the cell, this allows us to get a better approximation of its position.

3.2.2 An outline of Wanderlust

Wanderlust begins with a two-step initialization step (Figure 3-2, top left). First, a set of cells is randomly chosen as landmarks. Then, the data is transformed into an ensemble of l-k-NNGs. The algorithm proceeds by iteratively calculating the trajectory in each of the graphs separately: for each cell (referred to as a target), the target's position along the trajectory is first set to the

shortest-path distance from a user-defined early cell *s*. The target's position is refined according to the shortest-path distance from each landmark. The distances are weighed so that landmarks closer to the target contribute more to the calculation (as they are less susceptible to the noise inherent in the shortest-path distance). However, the landmarks are themselves cells. Therefore, their position will change following the refinement based on the same calculation that was applied to the rest of the cells. Since cell positions depend on landmark positions, the shift in landmarks might obsolete the new calculated positions. Therefore, the refinement step is repeated with the new landmark positions until the positions of all cells converge. Once the trajectory calculation step completes in all of the graphs, the output trajectory is set to the average over all graph trajectories.

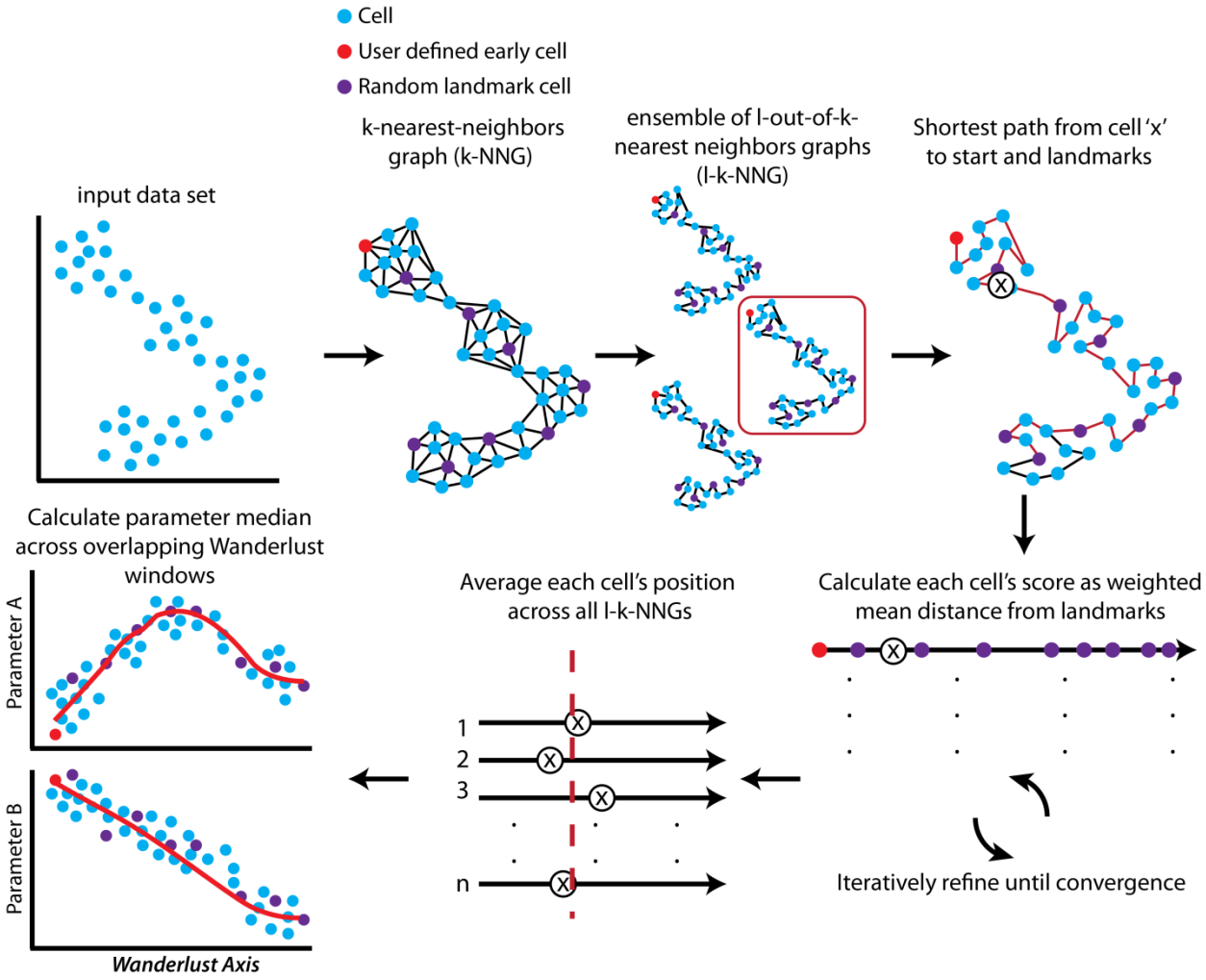


Figure 3-2. Description of the Wanderlust algorithm.
The input data is presented as the toy scatter plot to the top left. The order of the arrows follows the different stages of the algorithm. First, Wanderlust transforms the data into an ensemble of graphs, and flags a random subset of cells as landmarks (marked in purple). The list of landmarks is constant for the rest of the run of the algorithm. Calculation proceeds separately for each graph (portrayed for the highlighted graph here). In each graph, a user defined early cell (marked in red) is used to calculate an orientation trajectory. The orientation trajectory is iteratively refined using the landmarks. The final trajectory score is an average over the trajectory of all graphs. In order to examine trends across the trajectory, we define the trace of each marker as the median marker intensity in overlapping windows across the trajectory.

3.2.3 Formal description of the Wanderlust algorithm

Wanderlust receives as input a list of N points in D dimensions. Each point is a cell represented by a vector of length D, where each element is a measurement of the intensity of one marker. The algorithm assumes that the cells lie upon a one-dimensional developmental trajectory. In addition, Wanderlust receives an early cell s that serves as a starting point for the trajectory

detection. As its name implies, *s* is expected to originate from the beginning of the trajectory. For each cell, Wanderlust outputs a continuous trajectory score which provides the cell's temporal position across development: *s* has a score of zero and the most mature cell has a score of one, with the rest of the cells in between. Section 3.2.4 includes a pseudo-code of the algorithm including a summary of user-supplied parameters.

Wanderlust is composed of two steps: an initialization step and an iterative trajectory detection step. In the initialization step, Wanderlust flags a set of cells to serve as landmarks. The landmark selection is done uniformly at random and therefore uses no prior information about the data its underlying developmental process. The landmarks buffer against noise: each cell is going to have landmarks nearby, reducing the variability in calculating its position across the trajectory.

Next, the data is converted to a k-nearest-neighbor graph (k-NNG): each cell is represented by a node and is connected via an edge to the k cells most similar to it. The edge weights are equal to the distance between the two nodes. The graph is represented as an adjacency matrix, where each row and each column are a cell, and the value at position *(k, l)* corresponds to the weight of the edge between nodes *k* and *l* [86]. The k-NNG is used as a template for the generation of an ensemble of l-out-of-k-nearest-neighbor graphs (l-k-NNG). A single l-k-NNG is generated by randomly and uniformly picking l neighbors out of the k-nearest-neighbors for each cell. As with the landmarks, the construction of a random ensemble mitigates noise, since a spurious edge in the k-NNG will be absent from most l-k-NNGs randomly derived from it.

After landmarks have been chosen and the l-k-NNG ensemble constructed, the trajectory calculation step begins. This is an iterative process that is done separately for each l-k-NNG. First, we define the shortest-path distance between each pair of nodes *(s, t)* as:

$$shortest - path - distance(s,t) = min_P \sum_{e \epsilon P} w(e)$$

where *P* is a path between *s* and *t*, *e* is an edge and *w(e)* is the weight of *e*. We calculate the shortest-path distances using Dijkstra's algorithm [85] which has a running time of O(|E|+|V|log|V|), when |E| is the number of edges and |V| is the number of nodes. Briefly, Dijkstra's algorithm initializes the distance from *s* to all other nodes in the graph to infinity. The algorithm recursively scans the graph, at each step updating the distances to any nodes that can be reached. The algorithm stops when *t* is reached.

For each cell, the trajectory score is initialized to the shortest-path distance to that cell from the early cell that was supplied as a parameter to the algorithm. We define this score as the initial orientation trajectory.

Next, for each cell (referred to as target), the shortest-path distance is calculated between each landmark and the target. However, distance does not have a direction: we cannot separate between the cases where the target precedes a landmark and where the target follows a landmark (figure 3-3, top and center). Therefore, an orientation step follows, where we utilize the initial orientation trajectory to decide on cell ordering relative to each landmark (figure 3-3, bottom):

given early cell s then for each target cell t and landmark l,

if d(s,t) < d(s,l) : t precedes l
otherwise : t follows l

Additionally, graph traversal is in itself a source of noise and is proportional to the shortest-path distance: as the distance between two nodes increases many more possible paths exist between them, leading to higher variance in the traversed distance. Therefore, the distance from the target to a nearby landmark has lower variability than the distance to a distant one. This can be leveraged in minimizing the noise by defining a weight for each landmark:

$$w_{l,t} = \frac{d(l,t)^2}{\sum_m d(l,m)^2}$$

The summation at the denominator is over all target cells *m*; the weight of each landmark is exponentially proportional to its distance from the target. The trajectory score for *t* is the weighted average over all landmark distances:

$$traj_t = \sum_l \frac{d(l,t)}{nl} w_{l,t}$$

where the summation is over all landmarks *l* and *nl* is the total number of landmarks. In this weighing scheme closer landmarks, whose distance to the target is less noisy, have a higher weight in its trajectory score..

During this step the starting cell and each landmark also become target cells and their trajectory score changes. We use this trajectory score as a new orientation trajectory and repeat the orientation step. Since the graph itself does not change the shortest-path distances do not change and we can use the existing distances in the repeated trajectory score calculation. Landmark positions continue to change with each orientation step, which is repeated until landmark positions converge:

$$converge?\,(L_{t-1}, L_t) = \rho_{L_{t-1},L_t} == 1$$

where L_i is a vector of landmark positions at orientation step i and ρ is Pearson's correlation.

The l-k-NNG's trajectory is equal to the trajectory of the last orientation iteration.

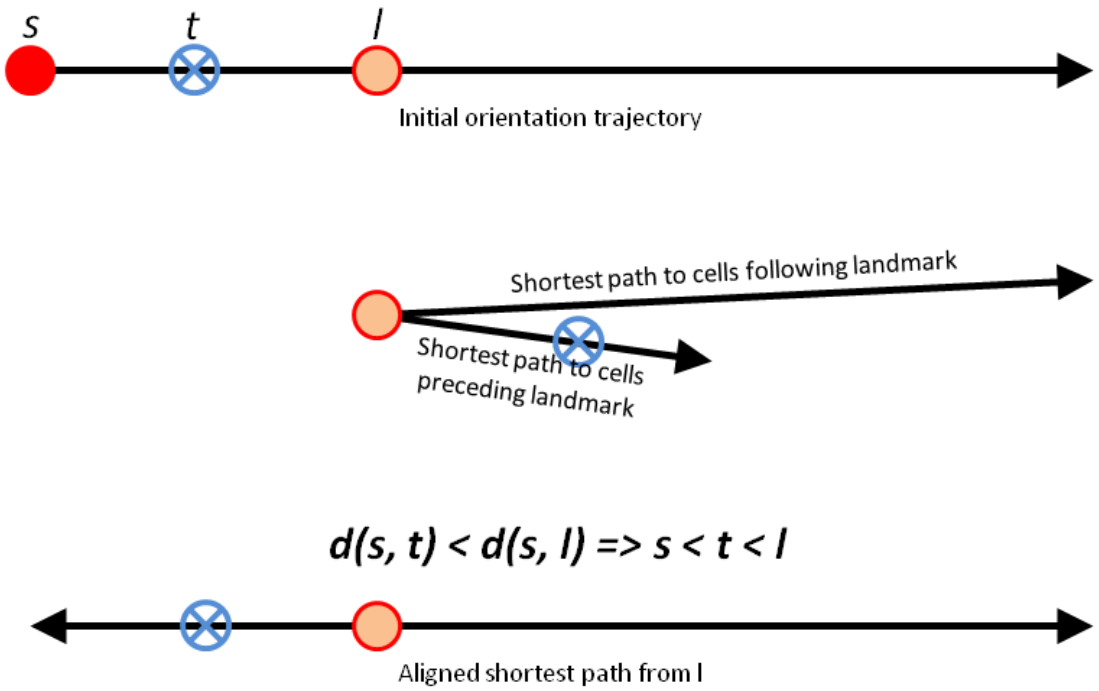


Figure 3-3. The orientation step of the Wanderlust algorithm.
Top: the initial orientation trajectory from the early cell (*s*) to a target cell (*t*) and a landmark (*l*). center: the shortest-path distance from the landmark to the target cell. Since distance has no direction we cannot identify whether the target cell precedes or follows the landmark. Bottom: According to the initial orientation trajectory, the distance from the early cell to the target is lower than its distance to the landmark. Therefore, the target must be between the early cell and the landmark and we can orient the shortest-path distance from l accordingly.

Finally, after a trajectory is iteratively calculated for each l-k-NGG graph, the output trajectory is set to the average over the trajectory scores of all l-k-NNG graphs.

3.2.4 Pseudo-code of Wanderlust

Input: data set of cells $X = \{x_1, \ldots, x_n\}$, starting cell s, number of landmarks nl, distance function $dist$, ensemble parameters: size of ensemble ng, number of nearest neighbors k, subset size l
Output: trajectory score $S = \{s_1, \ldots, s_n\}$ for each cell x_i

Initialization:

pick from X nl cells uniformly at random to serve as landmarks; set s as first landmark → $\{l_1 = s, \ldots, l_{nl}\}$

calculate k-nearest-neighbor graph of $X \rightarrow G$

randomly generate ng l-out-of-k-nearest-neighbor graphs → $G_{1, \ldots,} G_{ng}$

Trajectory calculation:

for each l-out-of-k-nearest-neighbor graph

calculate shortest-path distance from each landmark l_j to each point $x_i \rightarrow D = d_{ij}$

set w_{ij} to $|d_{ij}|^2 / \Sigma_k |d_{ik}|^2$

realign: calculate realigned distance T = t_{ij}. for each landmark l_j

$t_{ij} = d_{ij}$ if $d_{1i} > d_{1j}$, $-d_{ij}$ otherwise

$t_{ij} = t_{ij} + d_{1j}$

set $traj_{1,i}$ to $\Sigma_j t_{ij}/nl * w_{ij}$

repeat until convergence

realign t_{ij} using $traj_{(iter-1),i}$

set $traj_{iter,i}$ to $\Sigma_j t_{ij}/n_l * w_{ij}$

set the graph's trajectory to $traj$ of the last iteration

return average over all graph trajectories in ensemble → S

3.2.5 Wanderlust accurately recapitulates the trajectory in synthetic data

To evaluate Wanderlust's performance, we first applied it to synthetic data composed of a series of simulated datasets. A curved, one dimensional simulated trajectory, embedded in 3-dimensions, was generated by starting at position (1, 1, 1) and randomly traversing the space for 10,000 steps. After each step the current position was added to the trajectory as a point. All datasets had the same solution trajectory (figure 3-4, top). Each dataset additionally included seven dimensions of normally-distributed noise. The mean of each noise dimensions was zero. The magnitude of each noise dimension was defined as the standard deviation divided by the range of the solution trajectory. Each dataset used the same magnitude for all seven noise channels. A total of eight datasets were generated with increasing magnitude, from zero (no noise) to one (noise magnitude equals the range of the solution trajectory). In total, this synthetic data included eight datasets, each of which had 10,000 points and ten dimensions; regardless of the noise magnitude, each dataset included seven noise dimensions and only three trajectory

dimensions. The algorithm parameters were the same as those used in later analysis of biological data.

We computed a Wanderlust trajectory for each synthetic dataset and compared the resulting trajectory to the solution trajectory (figure 3-4, bottom). When there was no noise, the algorithm's output was almost identical to the solution (Pearson's ρ=1). Wanderlust continued to faithfully recapture the trajectory as noise levels increased to magnitude as high as 0.2 (Pearson's ρ=0.97). When the magnitude was increased to 0.5, output quality decreased as the entire first half of the trajectory was given a similar score of 0.2 by the algorithm. The second half, however, was well-modeled, giving a reasonable view of the system (Pearson's ρ=0.86). As expected, when the magnitude reached 1, the algorithm was no longer able to detect the solution trajectory. Over the eight datasets, Wanderlust perfectly detected the solution trajectory in six datasets, reached high correlation with one dataset, and failed to detect the trajectory in the last dataset (where the magnitude of the noise was equal to seven times the solution trajectory).

We next wanted to test whether Wanderlust continued to correctly detect the trajectory even when the data included short circuits. We chose the synthetic dataset with the lowest noise magnitude, 0.01 (Pearson's ρ=0.99), as a template for the generation of sixteen short-circuit datasets, since we wanted to isolate the effect of the short circuits (figure 3-5a). Dataset generation included two parameters. The first parameter, N, was the number of short circuits (N=50, 100, 500, 1,000). The length of each short-circuit was randomly sampled from an exponential distribution whose mean, μ, was the second parameter (μ=0.01, 0.05, 0.09, 0.13). When μ=0.01, very few of the short circuits were long-range. However, as μ increased, short-circuit length increases, and, more specifically, more long-range short circuits appeared. We

expected the detection quality to be inversely correlated with the number of short circuits and the proportion of long-range short circuits.

The Wanderlust trajectories were well-correlated with the solution trajectory (Pearson's $\rho>0.95$) in ten of the sixteen datasets (Figure 3-5b). As long as most of the short circuits were short-range ($\mu=0.01, 0.05$) the number of short circuits (N) had only a slight effect on the algorithm's output. When $\mu=0.09$, Wanderlust detected the trajectory well until N increased to 500 (Figure 3-5b, third row). Even when N was equal to 500 or 1,000, the algorithm gave a reasonable solution (Pearson's $\rho=0.92$).

We observed an interesting inversion observed when $\mu=0.13$ and most of the short circuits were long-range (Figure 3-5b, fourth row): with a small number of short circuits (N=50), Wanderlust modeled the first half of the trajectory well, but then backtracked and included the second half of the solution trajectory as a plateau, leading to a ridge in the scatter plot (Pearson's $\rho=0.59$); however, as N increased, so did the correlation between the algorithm's trajectory and the solution (Pearson's $\rho=0.74$, 0.83 and 0.85 when N=100, 500 and 1,000, respectively). When examining the scatter plots we saw that as more short circuits were added, less of the trajectory was modeled well, with the rest becoming a plateau (similar to the magnitude=1 dataset in Figure 3-4). The improved correlation is an artifact caused as the ridge in N=50 changes into noise.

Overall, Wanderlust detected the solution trajectory in the synthetic data in almost all of the datasets. Despite increasing levels of noise and the incorporation of varying quantities of short circuits of varying lengths, the algorithm reached a high correlation with the embedded solution trajectory. Wanderlust's high degree of robustness allowed it to overcome many of the challenges we expect trajectory detection to face in an experimental system.

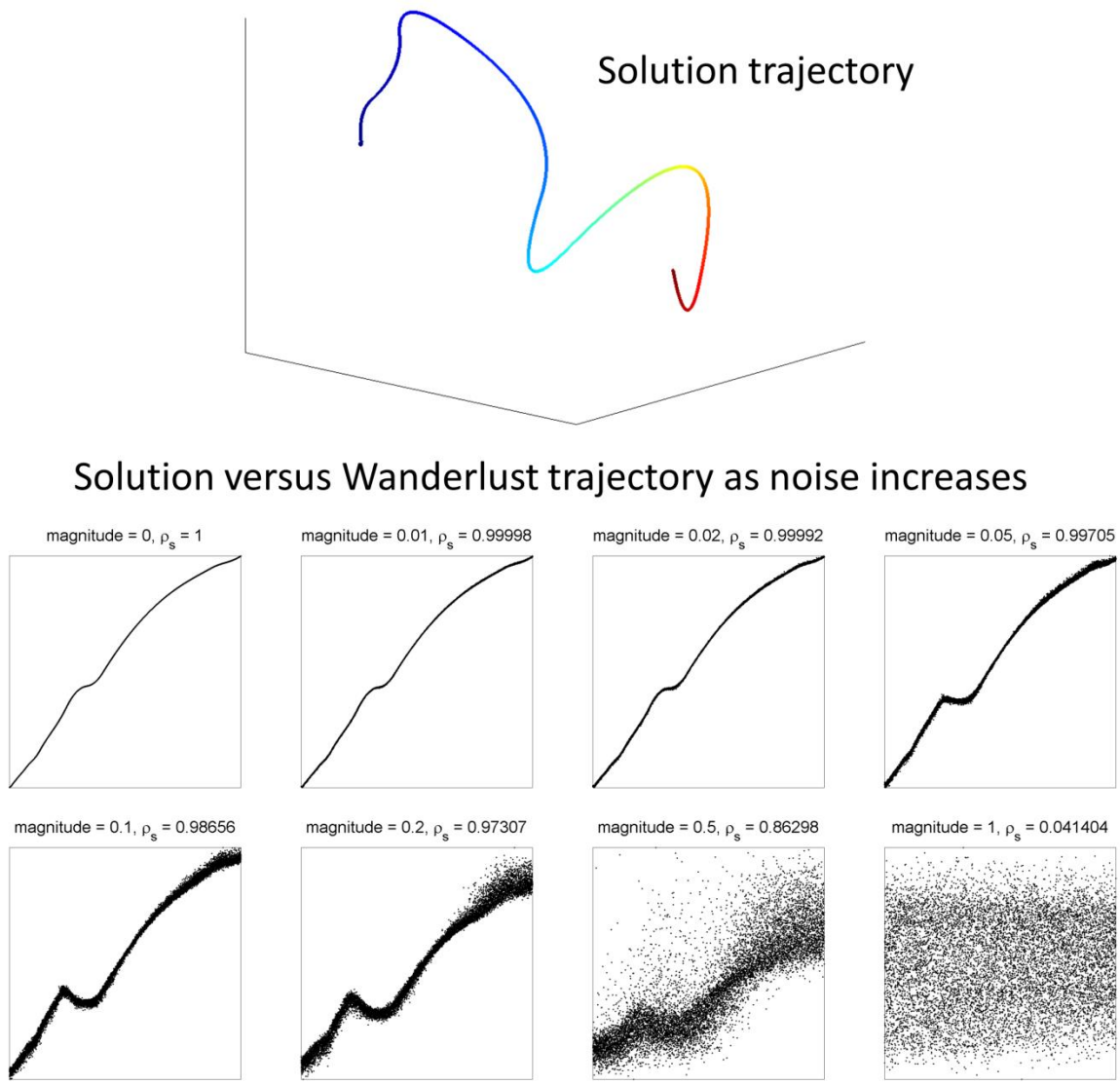


Figure 3-4. Trajectory detection in synthetic data with increasing amounts of noise.
Top: The synthetic datasets are composed of a 1-dimensional curve (trajectory) embedded three-dimensions (beginning colored in blue, end colored in red) and seven dimensions of normally distributed noise with $\mu=0$ and increasing magnitude (σ/range of solution trajectory; magnitude = 0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1). Bottom: Each plot corresponds to a Wanderlust run on one of the synthetic datasets. The magnitude of the noise and the Pearson correlation between the solution trajectory and the Wanderlust trajectory are indicated in each plot's title. The X-axes are the solution trajectory (cells are ordered by the solution) and the Y-axes are the Wanderlust trajectories (ordered by Wanderlust). Each dot is a cell. The Wanderlust trajectories are highly correlated with the solution trajectory even when the noise's magnitude equals half of the maximum.

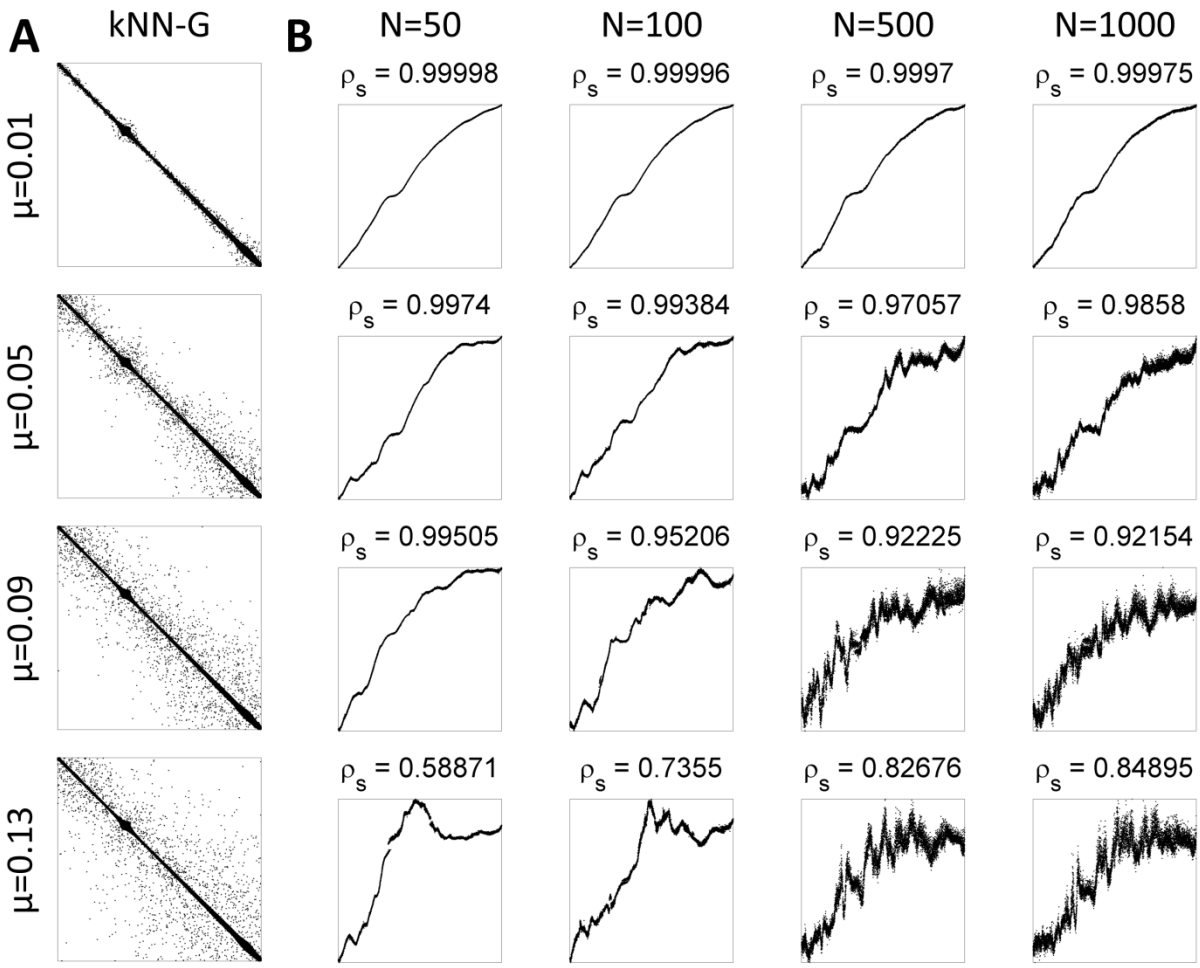


Figure 3-5. Wanderlust is resilient to short circuits.
The dataset with noise magnitude = 0.01 served as the basis for 16 synthetic datasets with varying amounts of short circuits (top, N=50, 100, 500, 1,000). The short circuit lengths were exponentially distributed with increasing μ (left, μ=0.01, 0.05, 0.09, 0.13). (a) The adjacency matrix of the 30-nearest-neighbors graph of each N=1,000 dataset. Each point is an edge. The X- and Y-axes are node numbers, ordered by the solution trajectory. The diagonal is composed of the real neighbors while the surrounding cloud is the short circuits. As μ increases the short circuits connect points which are more distant across the solution trajectory. (b) Wanderlust runs on these synthetic datasets. The X-axes are the solution trajectory and the Y-axes are the Wanderlust trajectories. Pearson's correlation values are given at each plot's title. As the number of short circuits increases, Wanderlust remains well correlated with the solution trajectory unless too many short circuits are long-range. Even when the data includes many long-range short circuits, the algorithm provides a reasonable trajectory.

Chapter 4 Trajectory detection orders hallmarks of early human B cell development

4.1 Introduction

B lymphocytes, whose central role is producing immunoglobulin, are a crucial component in the body's adaptive immune system. [1, 87]. An immunoglobulin, or antibody, is a Y-shaped protein that can bind to foreign objects (such as pathogens) and either neutralize them outright or assist other parts of the immune system in identifying them. The antibody is composed of two regions: the variable (V) region and the constant (C) region. The V region, which includes the antigen-binding site, defines the target of the antibody. Its sequence is generated during B cell development through a random process of genetic recombination. The C region, also referred to as the antibody isotype or class, defines the effector function of the antibody, including its distribution in the body and its functional activity.

The early development of B cells in humans, mice, and most other mammals occurs in the bone marrow. The immune cells pass through a series of stages: a hematopoietic stem cell, followed by a progenitor cell, a pro-B cell, a pre-B cell, and finally an immature B cell, which migrates out of the marrow [88-90]. Immature B cells then circulate between the various peripheral lymphoid tissues (such as the white pulp of the spleen or the lymph nodes) until they settle into either B-cell regions called follicles (in which case they are called follicular B cells) or the marginal zone of the spleen (in which case they are called marginal zone B cells) [91]. Follicular B cells are activated by a combination of antigens and signals from T helper cells and then congregate in germinal centers where they proliferate; their antibody's V region is further

tailored to the target antigen (a process called somatic hypermutation), and their antibody's C region might change to a different isotype (class switching). The cells can later differentiate into plasma cells, terminally differentiated cells whose purpose is the production of large amounts of antibody, or into memory B cells, slowly-dividing cells that can be used against later encounters with the same foreign object. Marginal zone B cells are resting, mature cells which are recruited to the early adaptive immune response in a T cell independent manner in response to specific types of extracellular bacterial pathogens.

The formation of a productive immunoglobulin is achieved through the genetic recombination of different V, D and J gene segments [92, 93]. First the heavy chain locus of the immunoglobulin (IgH) is rearranged during the pro-B cell stage: one D and one J segment are recombined by deleting the DNA between them followed by the joining of one V segment (again by deleting the DNA between the D and the V segments) [94, 95]. Likewise the V and J segments of the light chain loci are combined during the pre-B cell stage [93]. The recombination process is well-regulated through multiple mechanisms, including transcription, signaling, and cell proliferation and apoptosis [96-102]. Because of their fairly linear differentiation path and localization to the bone marrow as the primary site of development, B cells provide an interesting model for studying development of cells.

Different B cell stages can be identified through the expression of certain cell surface proteins, such as CD34 in stem cells and CD20 in more mature B cells [59, 89]. Much of the current understanding of human B cell development is based on studies using murine models, which allow large numbers of cells to be analyzed but are limited due to their intrinsic differences from human B cell development [101, 103]. Current state-of-the-art cytometry technology utilizes fluorescence to measure a small number of proteins simultaneously [8, 10, 11], limiting our

ability to examine multiple cell stages within a single sample. To bridge these gaps, we utilized mass cytometry in order to measure the expression level of more than 30 proteins in hundreds of thousands of B lineage cells in primary human bone marrow from healthy donors [12, 14]. However, this approach led to two computational challenges. First, the data were composed of a high number of nonlinearly correlated dimensions. Second, both the biological system and the technology itself were sources of noise that obscured the true developmental signal. A computational method for the exploration of B cell development using mass cytometry was needed to consider these statistical properties of the data.

As discussed above, we developed Wanderlust, a graph-based trajectory detection algorithm. Wanderlust's input is measurements of all cells in a system during a single time-point. The algorithm then identifies the trajectory underlying the system: that is, the developmental ordering of the cells. Wanderlust does not make any assumptions regarding the distribution of the data and is resistant to noise. It is composed of two steps: an initialization step, where the data is transformed into an ensemble of graphs, and a trajectory calculation step, where shortest-path distances are used to iteratively refine the trajectory for each graph. The average over the trajectories of all graphs is returned as the final trajectory score- the position of each cell across the developmental trajectory.

We showed that the Wanderlust trajectory is an accurate depiction of B-cell development. Furthermore, the use of the Wanderlust trajectory aided in the identification of novel population of early B cells. Coupling the Wanderlust algorithm with high dimensional mass cytometry data allowed the observation of the timing of developmental checkpoints and the coordination of protein expression across development in an unprecedented systematic fashion. Wanderlust is a

powerful algorithm for the identification of an underlying time element in a static snapshot of a dynamic system.

4.2 Results

4.2.1 Wanderlust captures the features of B-cell lymphopoiesis

Given Wanderlust's performance on synthetic data, we investigated its effectiveness in the analysis of a complex tissue by applying it to a cohort of primary human marrow aspirates and focusing on the development of the B-cell lineage. Bone marrow B lymphopoiesis, being a non-branching process with all involved cells present in a single tissue, represented an ideal real-world test case for the algorithm.

Each acquired healthy bone marrow sample was enriched for B-lineage cells via magnetic bead depletion, where several non-B-lineage cell subtypes were removed: red blood cells, myeloid cells and T cells. The mass cytometry panel included cell surface markers (CD45, CD19, CD22, IgD, CD79b, CD20, CD34, CD179a, CD72, IgM-i, Kappa, CD10, Lambda, CD179b, CD49d, CD24, CD127, CD38, CD40, CD117, HLADR and IgM-s), signaling markers (cPARP, pPLCg, pSrc, pSTAT5, pAKT, pSyk, pErk12, pP38, pCreb), cell cycle markers (Ki67, pCreb), and TdT and Rag1, two enzymes which are responsible for the genomic rearrangement in B cells. An additional post-acquisition computational enrichment stage which was based on cell length and DNA content removed debris, doublet events, and dividing cells. Any remaining non-B-lineage cells were removed using a dump channel that included canonical markers for several non-B-lineage cell subtypes (CD3, CD235, CD61, CD66b, CD33, CD11c, CD16). After these cleaning steps, the data included between 20,000 and 300,000 cells.

To evaluate the Wanderlust trajectory, we applied the algorithm to the 21 phenotypic markers in each sample. The list of input markers and Wanderlust parameter values, including the distance metric used, are given in the Materials and Methods section. We visualized marker expression trends across the trajectory by calculating the trace of each marker over the trajectory using a sliding window. The trajectory was first divided into one hundred overlapping, equidistant windows. The width of each window was 8% of the total trajectory width (window widths between 6% and 12% were tested and provide comparable results). For each marker, the median marker intensity was calculated in each window (Figure 4-1a). In addition, the trajectory was divided into ten non-overlapping windows. In each of these windows, we examined the biaxial plot of three surface marker pairs: CD34 versus CD38, CD19 versus CD10 and CD20 versus surface IgM (figure 4-1b). This second visualization using biaxial plots followed the established view of B-cell development and serves as support for the Wanderlust trace.

To anchor the trajectory to known early B-cell development we examined markers whose levels are well-characterized (Figure 4-1). Wanderlust faithfully captured known trends across the progression [31, 47-49]: CD34 is the earliest marker expressed and is rapidly followed by CD38 expression, as demonstrated by both the Wanderlust trace and in the biaxial plots. CD10 expression is next. Shortly after CD19 levels begin to increase. As CD34 levels decline, IgH rises on the surface of the more mature cells that have productively rearranged the heavy chain locus of the immunoglobulin gene. CD20 expression rises in concert with a decrease in CD10 and the expression of the kappa light chain. Kappa expression rises and peaks and is followed by the expression of lambda, the alternative light chain component. In addition to recapitulating known trends, the trajectory serves as a quantitative scaffold that clarifies the relative timing and co-expression of phenotypic markers across B cell development; for example, we see that CD10

peaks while CD34 decreases, and that CD19 plateaus when IgM reaches half of its maximum intensity. The trajectory is a continuous estimate of marker expression levels in relation to each other.

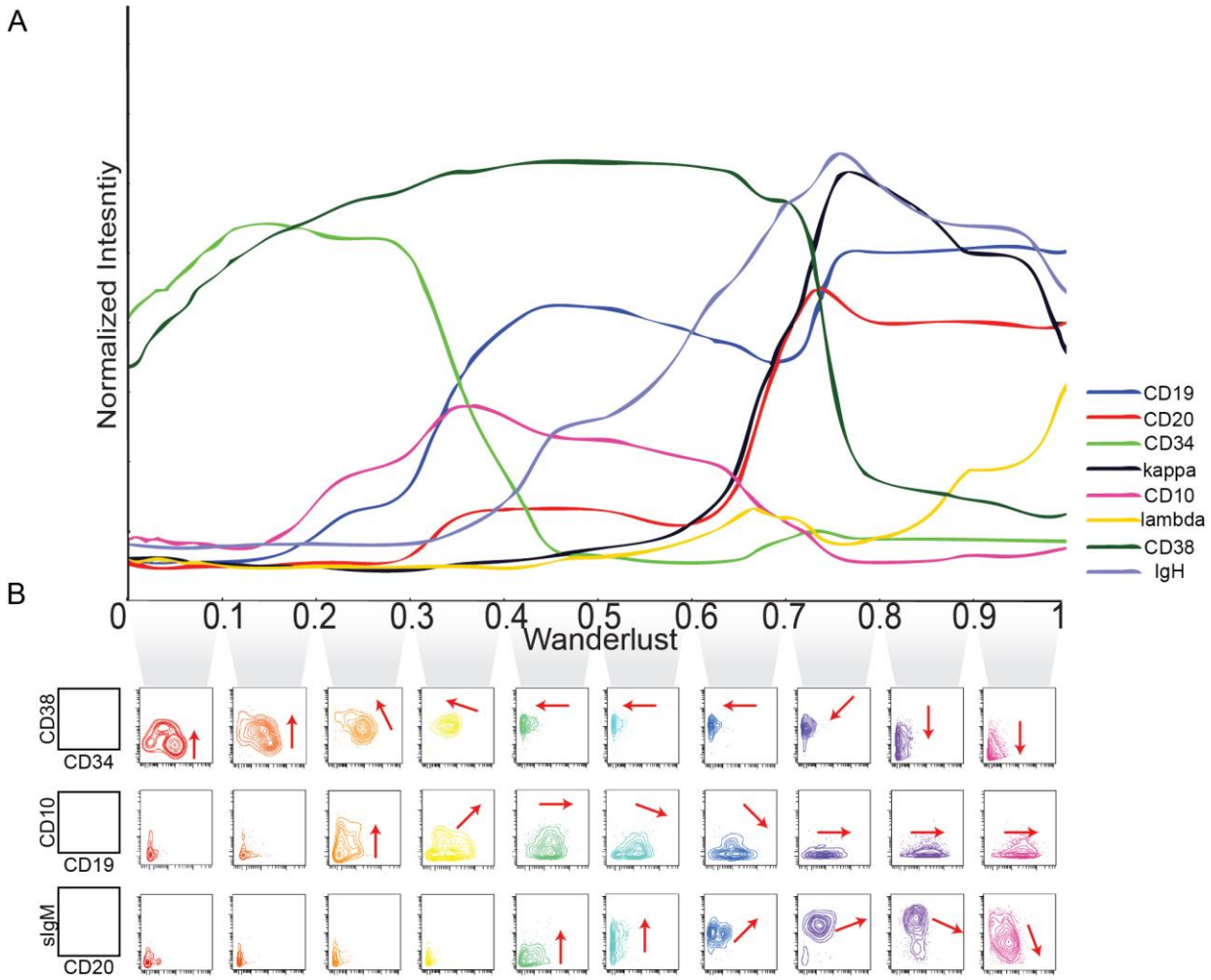


Figure 4-1. Wanderlust detects the trajectory of B-cell development.
The Wanderlust algorithm applied to B-lineage cell data from a healthy human bone marrow sample. (a) The traces of well-characterized developmental markers. The X-axis is the trajectory, divided into 100 overlapping, equidistant windows. Y-axis is the median normalized marker intensity in each window. Marker expression levels follow known trends. (b) Biaxial plots of CD34 versus CD38, CD19 versus CD10 and CD20 versus surface IgM in ten discrete windows across the trajectory. The red arrows show the expression changes from the previous window. Expression levels shift according to known trends as development progresses.

The variation in the expression of each phenotypic marker as a function of the Wanderlust trajectory was remarkably low (Figure 4-2). At any given point, the distribution of B-cell centric

epitopes was relatively low, indicating their combined concordance in the model and the ability of the algorithm to leverage multi-dimensional information to create a highly organized trajectory of cellular development. This tightness was especially apparent with TdT, which was not used as input to the algorithm. Additionally, examining variability across the trace reveals whether changes in marker expression are gradual (for example, CD38 and CD24) are more threshold-like in nature (such as the decrease in CD34 or the increase in IgM). The low variability across the traces further reinforces the accuracy of the Wanderlust trajectory.

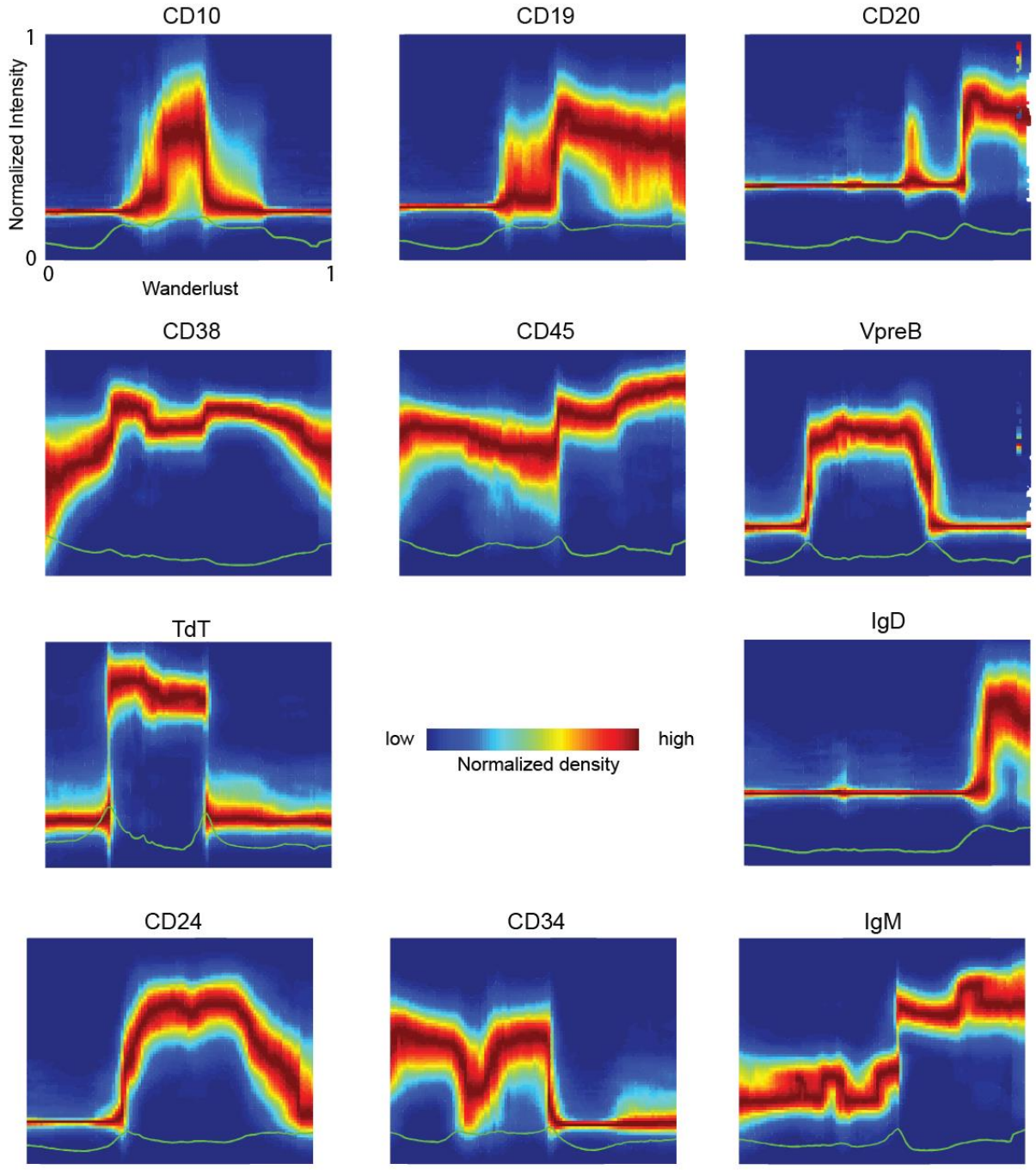


Figure 4-2. The marker expression distribution across the trace over the trajectory is tight.
Each panel corresponds to one marker (given at the title). The trajectory has been divided into 100 overlapping windows. In each window, the distribution of each marker has been calculated using a kernel-density estimator. X-axis is the trajectory, Y-axis is the density estimation at each window. Red denotes the highest density, blue the lowest (zero). The green line indicates the standard deviation of expression across the trajectory.

4.2.2 Wanderlust is robust over multiple runs and different samples

The Wanderlust algorithm initialization includes two stages that require random choices: the l-k-NNG ensemble generation (choosing of l out of the k-nearest neighbors for each node) and the landmark selection. These random processes could influence the output trajectory and lead to different results. To evaluate the robustness of the algorithm and reject this possibility, we re-ran the Wanderlust algorithm five times using the same healthy bone marrow sample data. The five runs were executed independently and each started from a different seed for the random number generator. The cell orderings over the five runs were almost identical (figure 4-3) with Pearson correlation greater than 0.99. While the second half of the trajectory seems to have more variation between pairs of runs, this was a visual artifact caused by the higher number of cells in that region of the trajectory (which was composed of more mature cells). In summary, Wanderlust is a robust algorithm that provides a consistent trajectory over multiple runs on the data.

Next, we wanted to verify that the Wanderlust trajectory is consistent over multiple individuals, thus representing human B-cell development, rather than that of a specific individual. We applied the algorithm to data from four healthy bone marrow samples that were acquired from different healthy individuals. However, a direct comparison between the trajectories is misleading. If a sample includes many cells of a given subtype, these cells will occupy a larger region of the Wanderlust output than other, less-represented subtypes. When examining multiple bone marrow samples, the proportions of different cell subtypes vary according to many factors (such as genetics, exposure to pathogens, and others). The altered proportions will lead to scaling discrepancies between the output trajectories, where less-populated areas shrink and higher-

populated areas expand. Therefore, the ordering provided by the trajectory is relative within the sample and cannot be applied to other samples.

We addressed this issue using the mean of all marker cross-correlations. Given a marker, for each sample we calculated the cross-correlation between that marker's trace in the sample and its trace in an arbitrarily chosen base sample. The trajectory was shifted such that the mean of all cross-correlations was maximized (figure 4-4). We next examined the cross-correlation corrected traces of several markers across the four trajectories. Even after shifting by the aggregate cross-correlation, the mean Pearson's correlation between the marker traces over the four samples remained high ($\rho > 0.9$ for all markers). The traces for TdT, which was not used in the Wanderlust analysis, were especially correlated ($\rho = 0.94$). Apart from that, no one marker seemed to be better coordinated across the samples. Likewise, no members of sample pairs seemed to be better correlated with each other: for example, while the IgM and TdT traces for samples b and c neatly overlap, their CD24 and IgD traces are dissimilar. Combining figures 4-3 and 4-4, we see that the Wanderlust trajectory remains consistent within and between healthy bone-marrow samples.

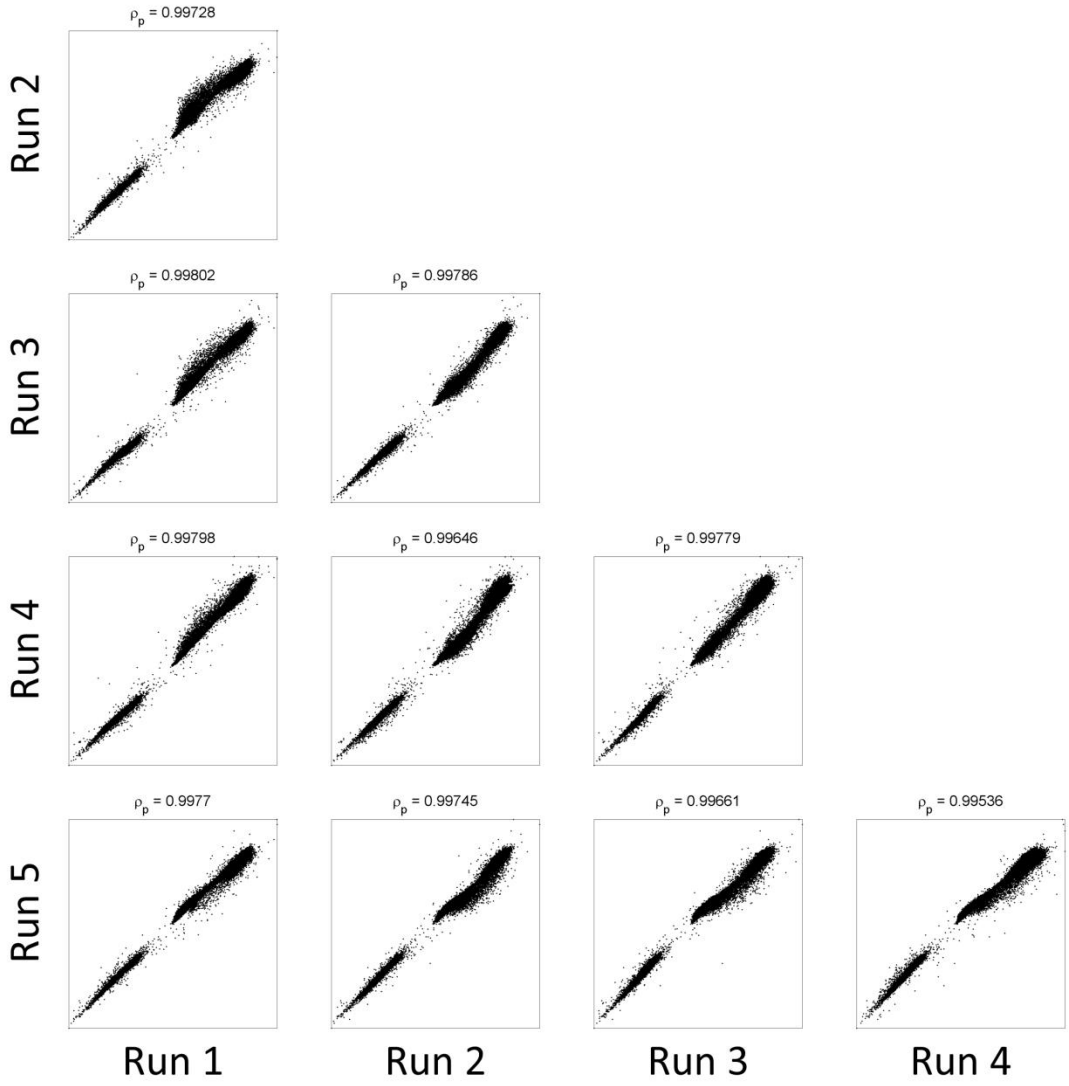


Figure 4-3. Wanderlust outputs a consistent trajectory over multiple runs and different samples. Repeated Wanderlust runs for the same sample, using different random seeds. The X-axes and the Y-axes are the algorithm's output for the respective runs. Pearson's correlation is given in each plot's title. The correlation never drops below 0.99.

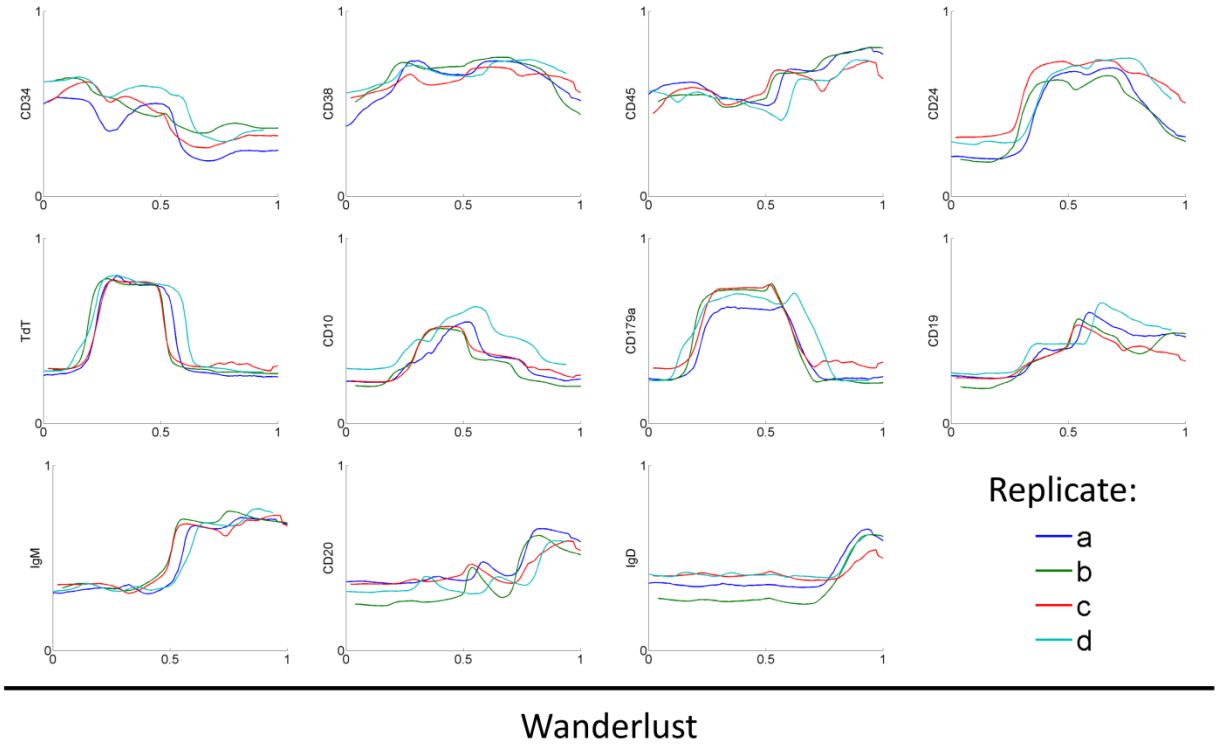


Figure 4-4. Wanderlust outputs a consistent trajectory over different samples.
Marker traces across the Wanderlust trajectory for 4 different healthy human bone marrow samples (denoted a to d). The trajectories were aligned using cross-correlation. The traces are almost identical between the samples (Pearson's ρ never drops below 0.9).).

4.2.3 Wanderlust is robust over a wide range of parameter choices

Wanderlust requires the user to supply several parameters along with the input data. Key among these is the early cell that serves as the start of the initial orientation trajectory. However, locating an absolute starting point for B-cell development or other biological processes is challenging. First, the definition of the starting cell might be inaccurate or unavailable; for example, only a subset of the stem-cell markers might be known. Noise in the data might result in a later cell having higher values in the relevant markers, obscuring the true starting cell. Alternatively, the measured panel might not even include the markers needed due to technical reasons. When applying Wanderlust to the healthy B-cell data, we supplied the algorithm with a

$CD34^{+}Lin^{-}$ stem cell. This is an approximate starting cell and the data might have other cells preceding it in the developmental chronology.

We used the output from the Wanderlust run described in figure 4-1 as a baseline trajectory for testing the influence of the early-cell parameter on the algorithm's output. We re-ran Wanderlust ten times. Each time, the early cell (s) was shifted by 0.1 across the baseline trajectory and the output was compared to our initial run (figure 4-5). As long as s remained within the first third of the baseline trajectory the two trajectories overlapped (Pearson's ρ=0.99, 0.99 and 0.98 for s=0.1, 0.2 and 0.3, respectively). When s is set to the midsection of the baseline trajectory (s=0.4, 0.5, 0.6 or 0.7), we see that Wanderlust breaks the output trajectory in half: the algorithm models the second half of the trajectory well, then backtracks and models the second half in reverse. The latest cells are connected to the earliest cells, with the middle becoming a starting point. Finally, when s is a later cell (s=0.8, 0.9, 1.0), Wanderlust detects the reverse trajectory, starting from the latest cells and going back to the earliest cells (Pearon's ρ=-0.98, -0.96 and -0.95, respectively). The algorithm has a meaningful output for a wide range of early cell choices, and only an approximate early cell is needed to detect the trajectory. Moreover, Wanderlust can recover an accurate reverse trajectory when starting from a late cell and going towards an early cell.

The generation of the l-k-NNG ensemble involves two parameters: k, the number of nearest neighbors in the initial k-NNG, and l, the neighbor subset size for each node in each graph. These parameters have several ramifications on the algorithm's performance. If k is too low, the k-NNG might be disconnected. Likewise, if l is too low, too many edges will be removed and the l-k-NNG will not be connected. In both cases some of the cells will be unreachable by the graph walk and will not be ordered at all by the algorithm. On the other hand, a high k will increase the number of short-circuits in the k-NNG. Similarly, if l is too close to k, more short circuits will

overlap across the l-k-NNG ensemble. From a complexity perspective, increasing l reduces the sparsity factor of the graph, leading to slower run times. In light of the above, k/l choice is crucial for accurate results.

All prior Wanderlust runs used the values k=30 and l=5 (each cell had 30 neighbors in the initial k-NNG, 5 of which were chosen in each l-k-NNG in the ensemble). We tested all combinations of k and l values over a set of values (k=20, 30, 50, 100, l=5, 10, 20, 30). The Wanderlust trajectory generated in figure 4-1 was again used as the comparison baseline. The correlation between the figure 4-1 trajectory and each k/l combination trajectory was high (figure 4-6, Pearson's ρ greater than 0.99), showing that the algorithm is generally consistent over choices of these two parameters. However, when examining the scatter plot for k=20, l=20 (figure 4-6, bottom left), we see the effect of a short circuit: a cloud of cells diverges from the baseline toward the end of the trajectory, showing that that region is not modeled well. Since l was identical to k, there is no l-k-NNG ensemble and the algorithm is susceptible to this short circuit. As long as l is lower than k, the l-k-NNG ensemble is used, leading to accurate trajectories.

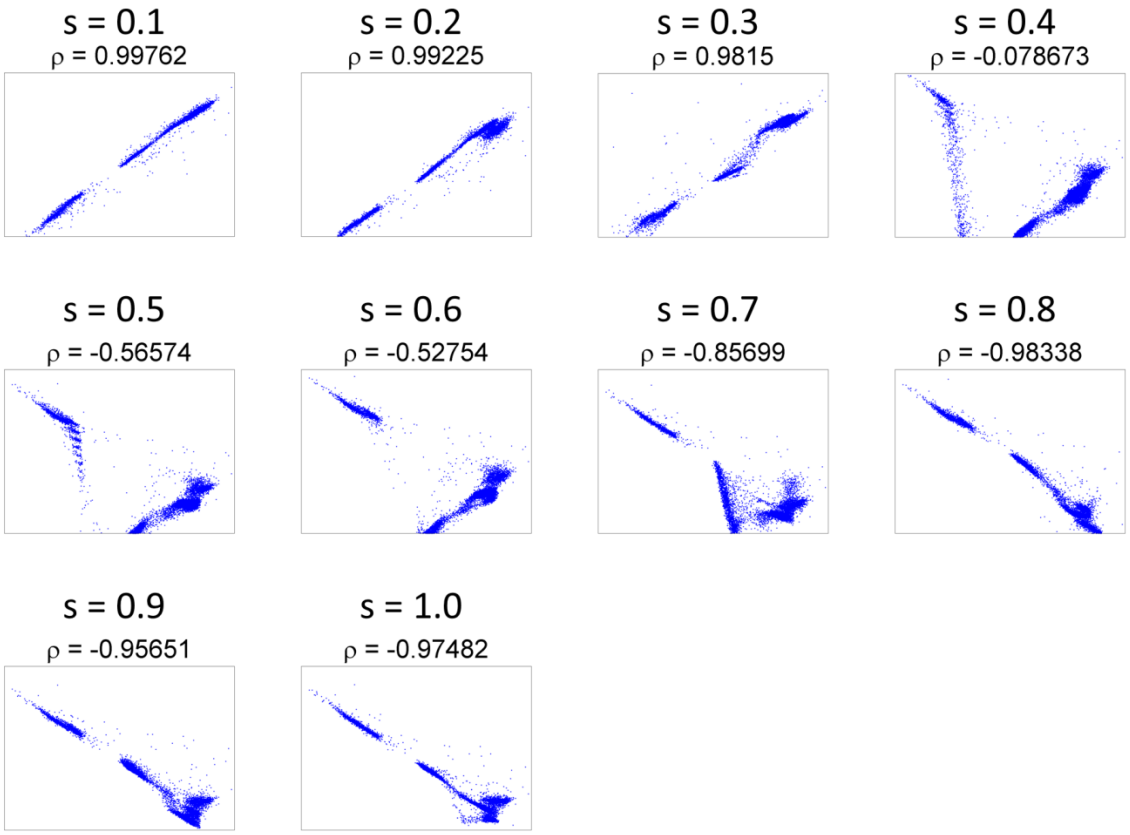


Figure 4-5. Wanderlust is robust to early-cell parameter choice.
The Wanderlust algorithm has been rerun ten times with the early-cell parameter advancing across the baseline trajectory from figure 4-1. X-axes are baseline trajectory, Y-axes are Wanderlust's output for given s. Each dot is a cell along the trajectory. Pearson's correlation and "starting point" location is given in title. The Wanderlust trajectory is well correlated with the baseline for s=0.1, 0.2 and 0.3, and is inversely correlated (the algorithm detects the reverse trajectory) for s=0.8, 0.9 and 1.0. For other values of s, Wanderlust broke the trajectory in half and modeled each half correctly.

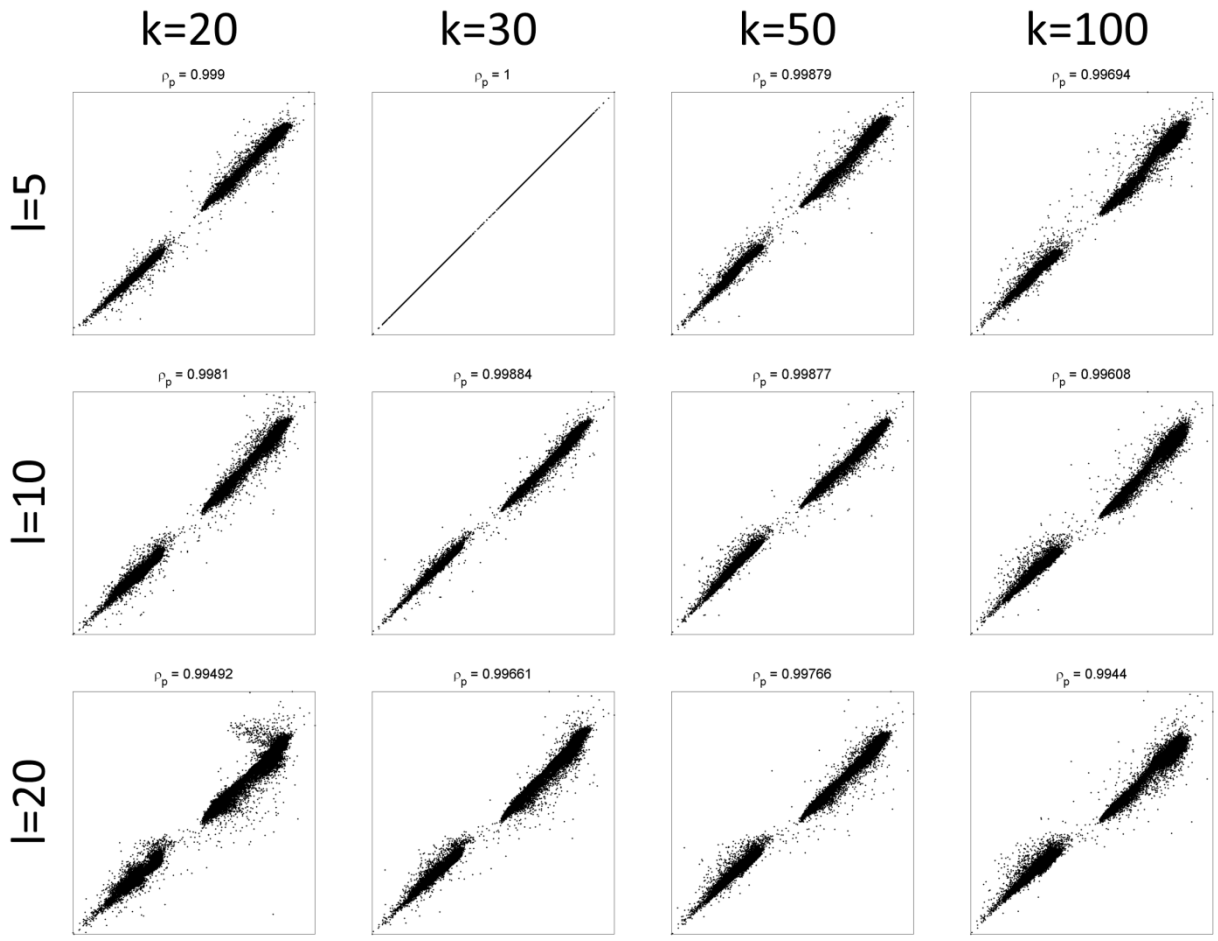


Figure 4-6. Wanderlust is robust to k/l parameter choice.
The Wanderlust algorithm has been rerun twelve times with different k/l parameter values. The figure 4-1 trajectory used k=30 and l=5 as parameters (top row, second from left). X-axes are the baseline trajectory, Y-axes are the output trajectory for the k/l parameter choices denoted above and to the left. Pearson's correlation values are given in the titles. The correlation is greater than 0.99 for all parameter combinations. A short-circuit can be seen when k=20, l=20 (bottom left plot).

The last set of parameters is Ng, the number of graphs in the l-k-NNG ensemble, and Nl, the number of landmarks. The number of graphs is again tied to Wanderlust's resistance to short circuits: a certain minimum number of graphs are needed or a short circuit might randomly appear in enough of them to skew the output trajectory. The number of landmarks is related to the algorithm's ability to reduce the variability caused by using the shortest-path distance. Since short distances are more reliable than long distances, enough landmarks are needed to guarantee that each cell has a landmark nearby. For purposes of optimizing accuracy, there is no downside

to increasing Ng and Nl, as there is no threshold above which these parameters will diminish the quality of the output trajectory. For purposes of optimizing runtime, however, there can be a downside to increasing Ng and Nl, because both are factors in Wanderlust's complexity.

To evaluate robustness, we tested multiple values for Ng and Nl (figure 4-7). The rest of the parameters were identical to the run in figure 4-1, which was used as the baseline, and in which Ng=20 and Nl=20. When Ng=1, Wanderlust performs poorly irrespectively of the number of landmarks utilized (figure 4-7, left). The algorithm follows the set of short circuits randomly chosen in that l-k-NNG, leading to a distorted view of the second half of the trajectory. The layered structure of the scatter plot allows us to follow the number of short circuits in each graph (for example, for Ng=1, Nl=5, there are seven short circuits in the graph). The trajectory improves when Ng=10, although the variability of its second half is very high when Nl=5. This trend of high noise between later cells continues as long as Nl=5, irrespective of Ng. Finally, when using at least ten graphs and at least twenty landmarks, Wanderlust outputs a consistent, high-quality trajectory, marking these values as the threshold for robust trajectory detection in the healthy B-cell dataset.

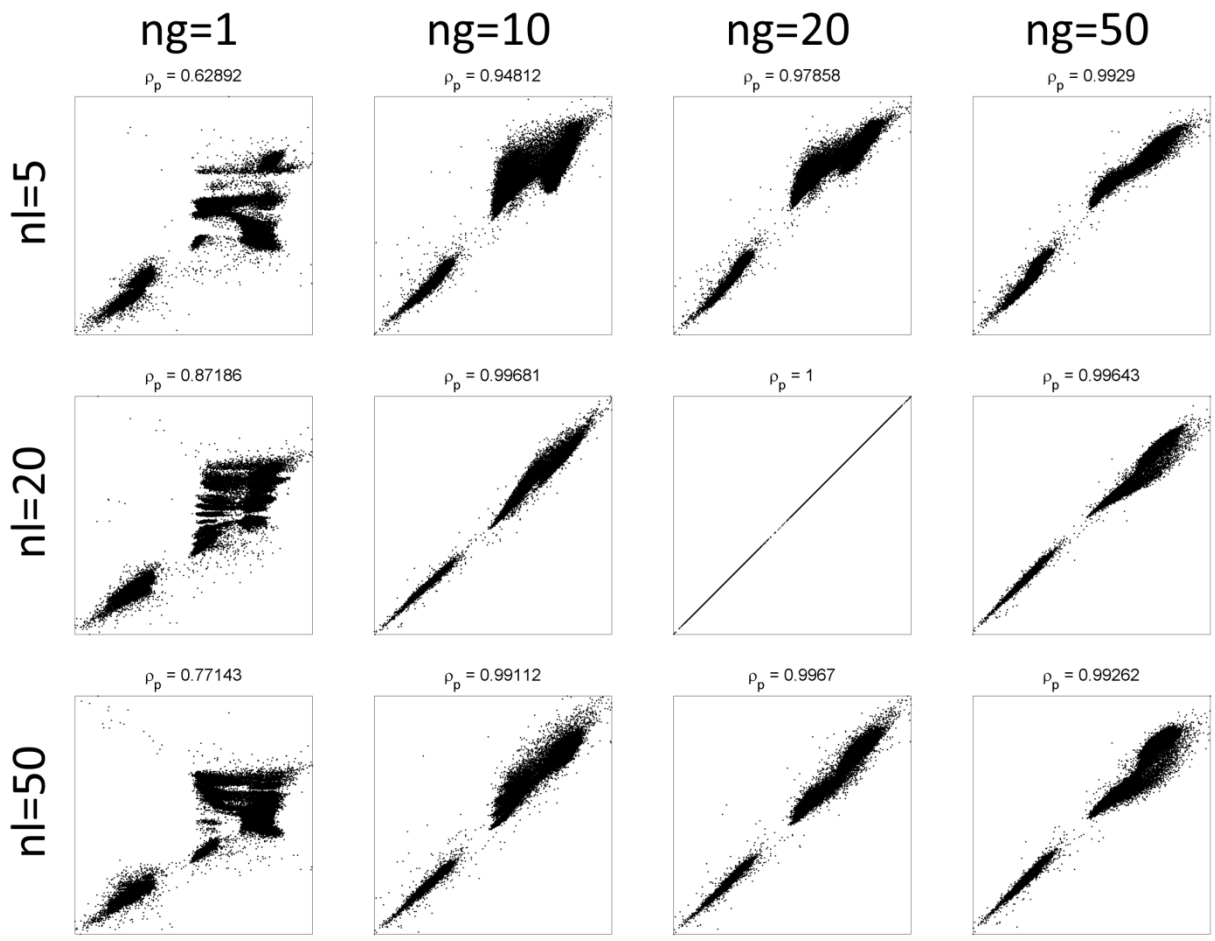
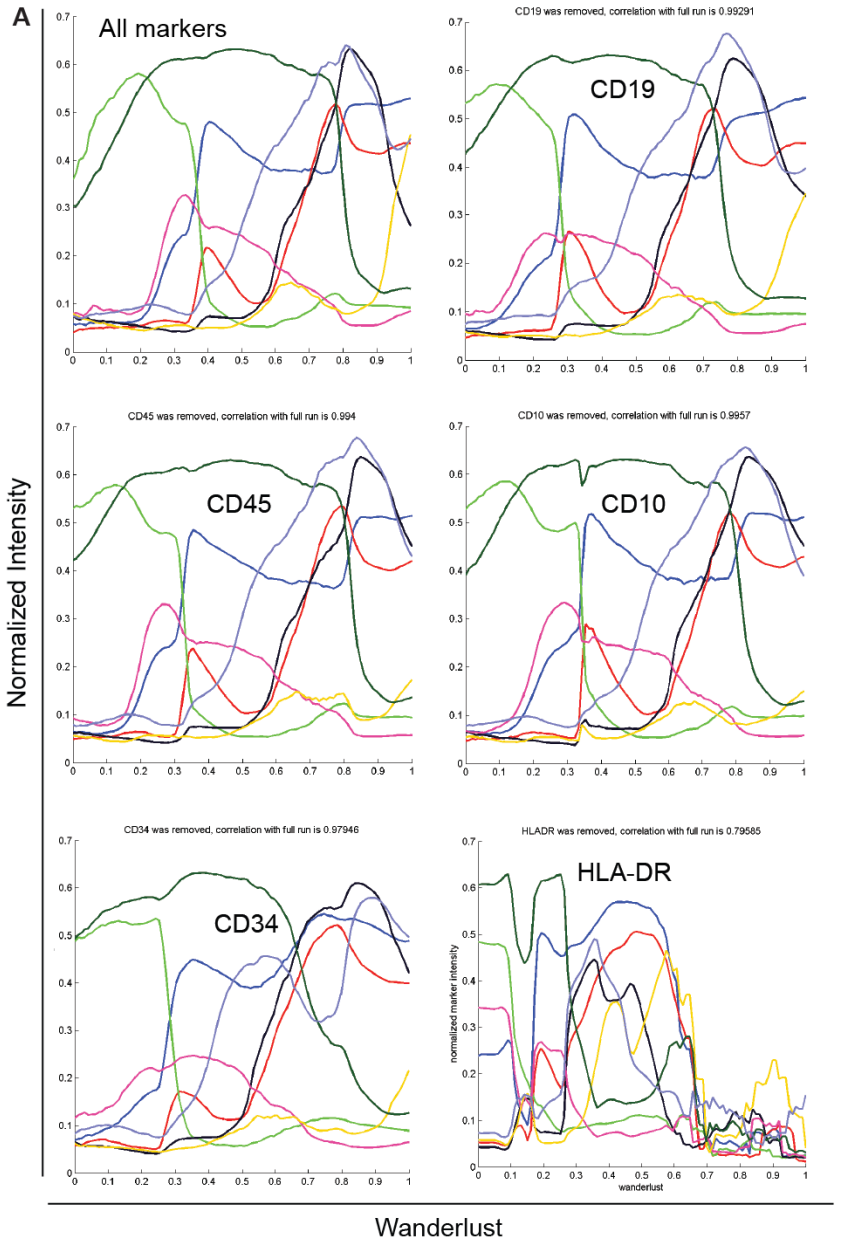


Figure 4-7. Wanderlust is robust to Ng/Nl parameter values past a certain threshold.
Wanderlust has been rerun a dozen times with different Ng/Nl parameter values. The comparison baseline is the trajectory from figure 4-1 (where Ng=20, Nl=20). X-axes are the baseline trajectory and Y-axes are the Wanderlust trajectory for that Ng/Nl parameter combination. Pearson's correlation values are given in the title. The algorithm fails the model the trajectory when Ng=1. The second half of the trajectory has high variation when Nl=5. In all other cases Wanderlust outputs a faithful representation of the trajectory.

4.2.4 Wanderlust is robust to marker selection

Marker selection is a central part of experiment design. Canonical markers are considered crucial in the identification of certain cell stages. Additionally, while mass cytometry offers a substantial increase in panel size, we cannot comprehensively include all of the surface markers that are expressed by the variety of immune system cell types, or even by just B-lineage cells. Since Wanderlust is only given a subset of possible markers, this raises the possibility that a missing linchpin marker will skew the algorithm's result.

To test the robustness of the trajectory to our selection of phenotypic features we independently ran Wanderlust multiple times. Each time, we removed one marker and calculated both marker traces and the correlation between the Wanderlust output and the original trajectory (Figure 4-8). Exclusion of any one individual maker had little effect on the results of the overall trajectory as evidenced by the strong correlation with the original model (Pearson's $\rho > 0.97$). Qualitatively, marker traces are identical between the different runs and faithfully follow the continuum of B-cell development. The only exception is HLA-DR, a component of all antigen-presenting cells, which had the greatest influence on the algorithm output with its exclusion dropping the correlation to 0.796. Notably, HLA-DR did not appear to affect the relative ordering of B-cell stages, only their position across the Wanderlust trajectory, suggesting it has a role in partitioning unrelated cell-types. Overall, for the representative Wanderlust trajectory discussed here, no single cellular marker served as a linchpin in the analysis.



B

Phenotypic Marker	Correlation to Original
CD10	0.996
CD19	0.993
CD20	0.994
CD22	0.994
CD24	0.986
CD34	0.979
CD38	0.936
CD40	0.986
CD45	0.994
CD49d	0.994
CD72	0.995
CD79b	0.994
CD117	0.991
CD127	0.995
CD179a	0.988
CD179b	0.993
HLA-DR	0.796
IgD	0.987
IgMi	0.990
IgMs	0.992
kappa	0.987
lambda	0.978

Figure 4-8. Wanderlust does not rely on any individual marker.
(a) Resulting Wanderlust trajectories from leave one out testing in which CD19, CD45, CD10, CD34 and HLA-DR were omitted from the trajectory algorithm. (b) Correlation to original trajectory (Figure 4-1) when indicated marker is omitted from the construction of the trajectory. Omission of HLA-DR (red box) in the construction of the trajectory has the largest effect on the overall trajectory (correlation 0.796).

4.2.5 Wanderlust uncovers and orders emerging B cell precursors

Following the robustness analysis, we used the Wanderlust trajectory to guide the identification of distinct, early populations and determine their relative ordering in developmental time. Due to TdT's role in the earliest activities known to define mammalian B cell emergence, we hypothesized that its combination with CD24 and other progenitor markers could serve as a novel set of identifiers to dissect early fractions of human B cells emerging in the marrow. We used Wanderlust to guide a series of biaxial gates based on CD34, CD38, CD24 and TdT, so that the resulting fractions best match Wanderlust's order of these cells (Figure 4-9a, left). Drilling down on the $CD34^{+}CD38^{+}$ fraction using the combined expression C24 and TdT revealed four distinct populations of cells. According to Wanderlust, these were early cellular fractions that sequentially occupy populations II-V, beginning with TdT expression, followed by cell surface expression of CD24 and finally loss of TdT as they proceed through development.

Enabled by the dimensionality of our single cell data, the expression of additional phenotypic markers could be used to support the progression of these populations and their identity as definitive early B cells. Intracellular λ5 (CD179b), a component of the surrogate light chain (sLC) of the pre B cell receptor (pre-BCR), is expressed first within population II (Figure 4-9b). As cells progressed to population III, the expression of CD10 emerges. By the time cells reach population IV, almost all co-express sLC proteins which coincides with maximal expression of CD10, aligning these cells with the pro B cell stage of development. Finally, as cells reach population V, most cells have lost Vpre-B, λ5, CD10 and all express the IgH protein intra-cellularly as they enter the later stages of Pre-B cell development (Figure 4-9b). Consequently, the Wanderlust trajectory of these populations (II-V) is confirmed by the co-expression patterns of the proteins typical of B-lineage development.

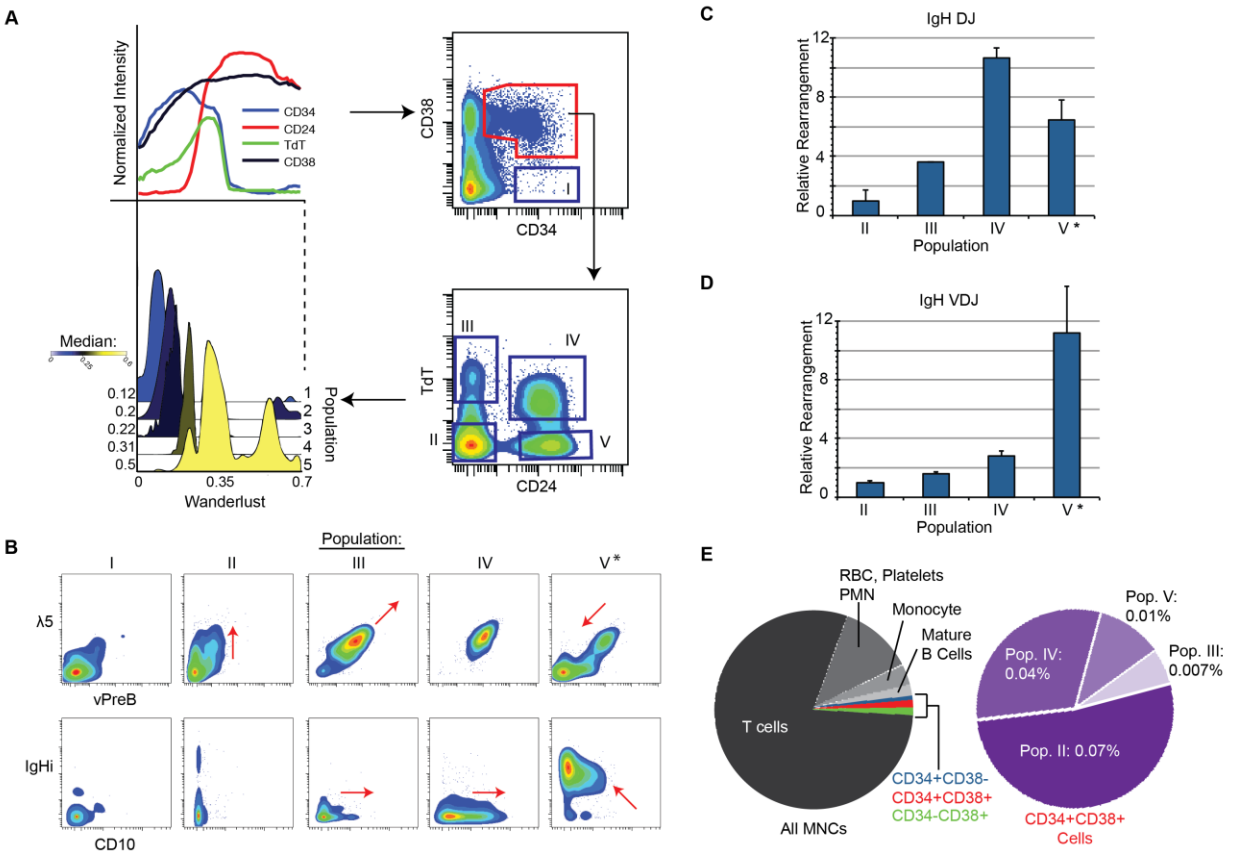


Figure 4-9. Wanderlust uncovers rare B cell progenitors prior to the expression of CD10 or CD19. (a) Wanderlust trace demonstrating CD24 expression rising sharply following the expression of CD34, CD38, and TdT (upper left panel). Examining the expression of TdT and CD24 on the CD34+CD38+ fraction on a contour plot demonstrates four easily gatable populations. An overlay of the median Wanderlust values for the cells contained within these gates (dark blue gates, labeled as population I-V), reveals the progression in maturity from cells in Fraction I (median WL 0.12) to Fraction II (median WL 0.2) to Fraction III (median WL 0.22) to Fraction IV (median WL 0.31) to and Fraction V (median WL value 0.5). (b) Expression of additional early B cell markers supports the B cell identity of these fractions. Once in population II, expression of λ5 rises on a minority of cells and by population IV, 100% of cells co-express both components of the surrogate light chain of the pre-B cell receptor, λ5 and Vpre-B. By population V, about half of cells have downregulated expression of these proteins. CD10 expression rises as cells transit population III-IV and by the time cells reach population V, the majority express intracellular immunoglobulin heavy chain (IgHi). (c) Results of relative amount of IgH DJ rearrangement as determined by qPCR analysis of genomic DNA from prospectively isolated cells from populations II-IV. Results are representative of two biological replicates and normalized to population II. (d) Results of relative amount of IgH DJ rearrangement as determined by qPCR analysis of genomic DNA from prospectively isolated cells from populations II-IV. Results are representative of two biological replicates and normalized to population II. (e) Grey pie chart displays the relative fractions of cell types contained in ficolled mononuclear cells from human bone marrow. Red slice demarcates CD34+CD38+ fraction. Purple pie chart displays the percentage of cells in populations II-IV that are contained within the CD34+CD38+ fraction.

4.2.6 VDJ Recombination confirms Wanderlust's ordering of novel early human B cell populations

We sought to further confirm the Wanderlust trajectory by examining the rearrangement of the germline IgH locus, which is the target of TdT and a measure of B cell identity. We developed a quantitative polymerase chain reaction (qPCR) assay that could linearly quantify the relative proportions of DJ- and VDJ-arranged cells in a mixed population. We used fluorescence activated cell sorting to isolate populations II-V from healthy bone marrow preparations from two individuals, extracted genomic DNA from each fraction and determined IgH rearrangement status using the qPCR assay.

We found progressive rearrangement of the IgH locus from population II to population V. The majority of cells had detectable DJ rearrangement by the time they reached population IV and completed VDJ rearrangement upon reaching population V (Figure 4-9c, d). This was consistent with the observation that virtually all cells in fraction V displayed intracellular expression of IgH protein (Figure 4-9b, asterisk). Thus, by establishing the progressive rearrangement of IgH in these populations, we confirmed that the Wanderlust trajectory not only facilitated the characterization of the earliest human B lymphocytes, but also accurately ordered their correct developmental timing, all from the analysis of a single human marrow, without synchronization or manipulation.

The ability to both identify and order emerging human B-lymphocytes across numerous individuals was particularly notable given their sparsity. Figure 3E highlights the rarity of these early B-cell fractions relative to the whole bone marrow. In particular, fraction III comprised only 0.007% of the mononuclear cell fraction of whole bone marrow. The fact that fraction III occurs prior to CD19 expression (Figure 4-9a, b) in combination with inconsistent expression of

CD10 (Figure 4-9b) offers an explanation as to why these fractions have not been described previously, phenotypically or functionally.

4.2.7 Wanderlust reveals pSTAT5 response to IL-7 is confined to rare B cell precursors

We sought to further characterize these early B-cell precursors, whose role in human B lymphopoiesis had been elusive to date. Mass cytometry allowed us to simultaneously measure not only surface markers, but also internal functional proteins and their modifications in the same cells. To functionally characterize early B cells and how these respond to stimuli, we collected data following perturbation with the cytokine IL-7. We focused on the activation of STAT5 by IL-7 via its phosphorylation site due to its critical regulatory role in mouse lymphopoiesis [104-107]. In mice, disruption of this pathway results in arrest of B cell maturation at the pro-B cell stage [98]. However, in the human, the precise developmental timing of this pathway and its regulatory role remain unclear.

Investigation of signaling response to IL-7 across the four early B cell populations II-V revealed that cells within population III displayed an almost exclusive induction of STAT5 phosphorylation (Figure 4-10a) – a striking observation considering population III represents less than one in ten thousand cells in the marrow. Moreover, this pinpointed response was consistent across seven individual marrows. We note that pSTAT5 and other functional markers were not used to construct the Wanderlust trajectory and therefore this pattern of pSTAT5 induction was not enforced by the algorithm, but rather was revealed due to the precise phenotypic ordering of cells.

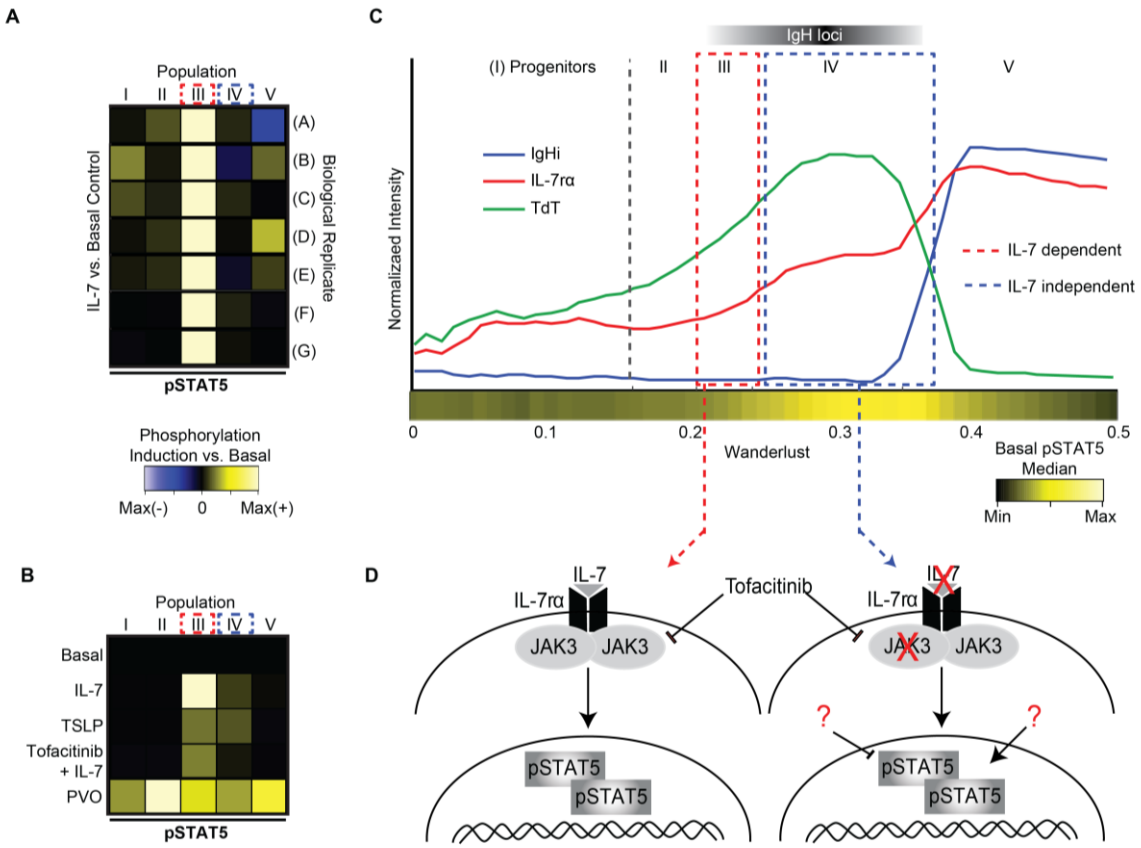


Figure 4-10. Regulatory signaling re-wires across development.
(a) Compared to basal control, IL-7 induces maximal induction of pSTAT5 in population III (CD34+CD38+TdT+) consistently across seven individual biologic replicate human bone marrow samples. (b) Population III and population IV can also respond to thymic stromal lymphoprotein (TSLP) by induction of pSTAT5. Jak 1/3 inhibitor, tofacitinib, abrogates some IL-7-induced pSTAT5 signal in population III but has no effect on other developmental fractions. (c) Across the early half of the Wanderlust trajectory (0-0.5) expression of the IL-7rα rises early along with TdT but prior to expression of IgHi. Heatbar indicates basal pSTAT5 expression levels across early trajectory. Developmental timing of populations I-V is indicated at the top of the trace along with timing of IgH loci rearrangement. (d) Overlying the timing of populations III and IV demonstrates a switch of the regulatory network such that cells occupying population III demonstrate a Jak-dependent induction of pSTAT5 in response to IL-7 exposure whereas when cells transition to occupy population IV they are no longer responsive to IL-7 but maintain high basal pSTAT5 levels suggesting inhibition of a negative regulator or an alternative route of stimulation.

4.2.8 Wanderlust captures STAT5 network rewiring

Since the IL-7/STAT5 response was limited to a specific fraction, we sought to further characterize STAT5 regulation, relative to adjacent fractions. Activation of pSTAT5 by phosphorylation in mature lymphocytes is mediated by Janus kinase [108]. To confirm a Janus kinase (JAK) mediated mechanism of STAT5 control, we used the JAK inhibitor Tofacitinib,

combined with IL-7 stimulation. As expected, within population III, STAT5 activation could be attenuated by treatment with Tofacitinib, but not the SRC inhibitor Dasatinib, indicating a JAK mediated mechanism in these cells (Figure 4-10b).

Populations III's induction of pSTAT5 coincides with the cells gaining expression of the IL-7 receptor (CD127), where all CD127 positive cells strongly induce pSTAT5 in response to IL-7, explaining the lack of induction in earlier populations. IL-7 receptor levels do not peak at population III, but rather continue to rise in populations IV and V (Figure 4-10c, red line), yet *ex vivo* IL-7 stimulation no longer induces pSTAT5 in these later populations. Following the induction of pSTAT5 by IL-7 in population III, subsequent cells occupying fraction IV display a higher basal level of pSTAT5 (Figure 4-10c, bottom yellow bar). To test if pSTAT5 levels are saturated in later fractions, we used the pan tyrosine phosphatase inhibitor pervanadate (PVO). In the presence of PVO, the levels of pSTAT5 rose in all CD34+ progenitor B cell fractions, across biological replicates (Figure 4-10b). Additionally, cells in fractions III and IV yielded a similar STAT5 phosphorylation pattern in response to a thymic stromal lymphoprotein (TSLP) (Figure 4-10b), a ligand that shares the IL-7rα chain and is known to activate STAT5 [105].

Therefore, though STAT5 activation by IL-7 was restricted to the cells of population III, it was neither due to the lack of IL-7 receptor on later cell fractions, nor to their lack of STAT5 expression or ability to phosphorylate it. Together these observations illustrate a STAT5 network rewiring over the development of B cell precursors (Figure 4-10d). First, STAT5 phosphorylation is initially dependent upon IL-7 in a JAK mediated mechanism (population III). Then, despite continued expression of the IL-7 receptor, STAT5 phosphorylation becomes independent of IL-7 (population IV), yet remains basally high relative to developmentally adjacent cells.

4.2.9 STAT5 network rewiring occurs during immunoglobulin rearrangement

Previous studies in mouse have implicated the IL-7-dependent STAT5 induction in both the initiation of IgH rearrangement and the suppression of the kappa (light chain) locus [98, 109]. Using the trajectory as a scaffold, we could overlay the different measured elements and examine their relative timing. The peak expression of TdT (Figure 4-10c) indicates that cells in population IV are actively rearranging the IgH locus of the immunoglobulin gene, which is further supported by our qPCR assessment of IgH loci rearrangement (Figure 4-9c-d).

Thus, the switch in regulation of STAT5 activation overlaps with the cell undergoing genomic rearrangement. An IL-7-dependent stage of STAT5 activation at population III as IgH loci rearrangement is initiated, followed by an IL-7-independent stage during germline gene rearrangement (a perilous cell state that requires careful regulation). Following completion of IgH rearrangement, expression of the IgH protein rapidly rises in cells transitioning into population V (Figure 4-10c, blue line). Thus, when cells are analyzed by Wanderlust, we can observe coordinated rewiring of the regulatory signaling network across these rare, early B cell populations (Figure 4-10d).

4.2.10 Derivative analysis of Wanderlust reveals coordination points in development

The coordinated expression of phenotypic markers coupled with re-wiring of regulatory signaling implied that these events coalesced around developmental checkpoints controlling the progression of B-cell lymphopoiesis. This highly multiplexed dataset combined with the developmental ordering revealed by Wanderlust allowed us to examine the concurrent timing of protein expression across B-cell development. By approximating each parameter's derivative along the Wanderlust trajectory, we were able to quantify the rise and fall in their expression (Figure 4-11a). Then, by clustering each parameter based on the absolute value of this derivative

across the trajectory, we uncovered '*coordination points'*, when the change in expression of multiple proteins coalesce across B cell development (Figure 4-11b).

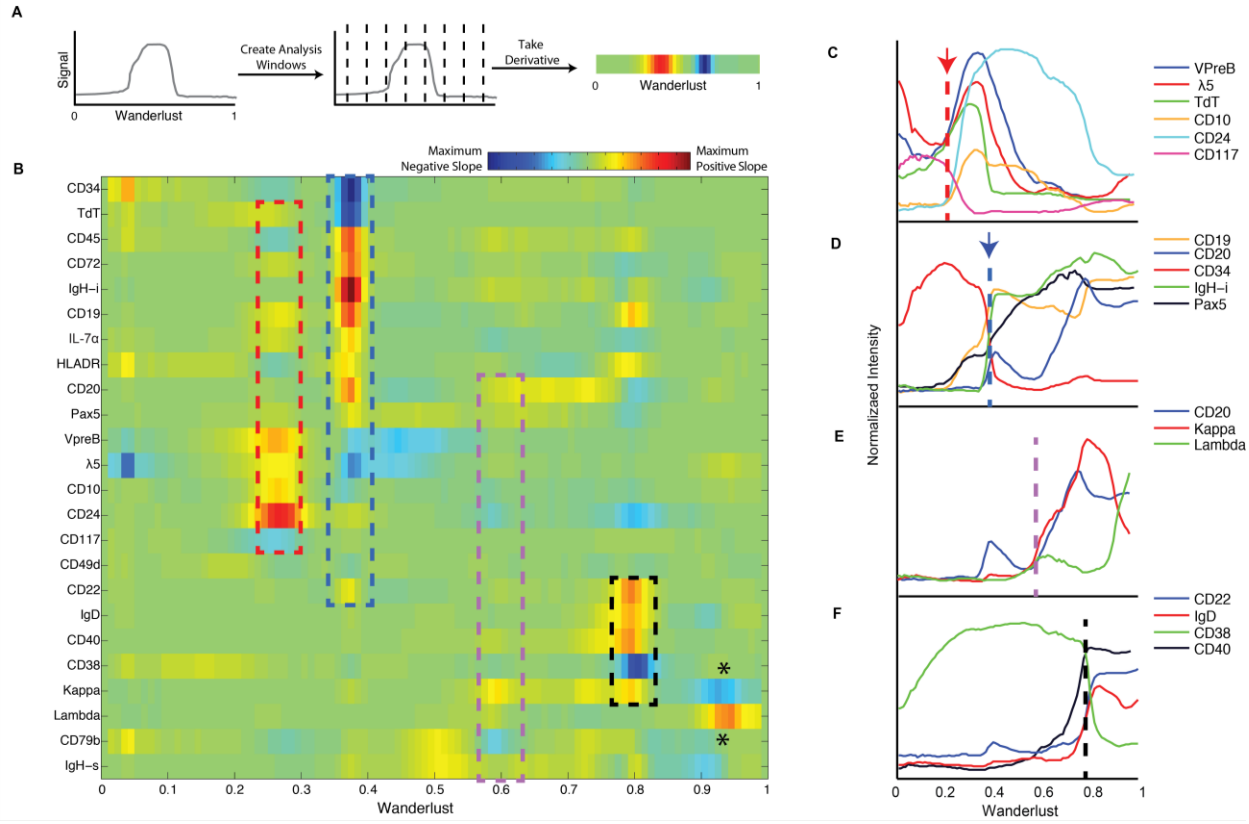


Figure 4-11. Coordination of protein expression across B cell development.
(a) The first derivative was calculated for the windows across each marker's trace. These values are expressed as a heatmap with red indicating a positive slope (increasing expression) and blue indicating a negative slope (decreasing expression). (b) Results of the first derivative analysis. Markers were hierarchically clustered. (c) Coordination point (red dashed line and box) at approximately 0.25 Wanderlust with rise in Vpre-B, λ5, TdT, CD10, and CD24 and fall in CD117. (d) Coordination point between 0.3 and 0.4 (blue dashed line and box) with drop in CD34 expression and rise in CD19, CD20, IgHi, and Pax5. (e) Coordination point at 0.6 (purple dashed line and box) showing the rise in CD20 with rise in kappa light chain protein. Lambda light chain expression rises later. Asterisks call out decline in kappa light chain expression co-incident with rise in lambda light chain. (f) Coordination point (black dashed line and box) at 0.8 with drop in CD38 expression, rise in CD40, IgD and CD22 expression.

At least four major inflection points of coordinated epitope expression can be identified across the trajectory (Figure 4-11b, dashed boxes). The first occurs between 0.2 and 0.3 (Figure 4-11b, red box) with the induction in expression of TdT, CD10, Vpre-B (CD179a), λ5 (CD179b) and the down regulation of CD117 expression, the receptor for stem cell factor (Figure 4-11c, red

line). This coincides with fraction III, the IL-7-dependent pSTAT5 cells (Figure 4-10), the overall combination of markers represents cells at the early pro-B cell stage of development [110] and occurs just prior to the timimg of IgH locus rearrangement. In the next coordination point, CD45, CD72, CD19, and Pax5 rise in concert with intracellular expression of IgH (essentially intracellular heavy chain IgM) while CD34, TdT, Vpre-B, and λ5 expression fall (Figure 4-11b and d, blue box and line). This combination of marker flux is consistent with cells that are passing through the pre-B cell stage and are preparing to rearrange the light chain locus of the immunoglobulin.

Light chain rearrangement is at the center of the latter two coordination points. The first of these (Figure 4-11b, purple box) coincides with kappa light chain protein expression, which mirrors the trajectory of CD20, signifying that the expression of CD20 occurs in concert with BCR light chain rearrangement and expression (Figure 4-11e, purple line). Cells that do not successfully express kappa switch to lambda light chain, which is both consistent with the known biology and correctly ordered by Wanderlust in this example (Figure 4-11b, asterisks and Figure 4-11e). Lastly, the synchronized induction of CD40, CD22, IgD expression coupled with the loss of CD38 expression (Figure 4-11b and f, black box & line) cements the emerging cells as naïve, immature B cells preparing to enter peripheral lymphoid organs [110].

In summary, Wanderlust successfully analyzed a dynamically asynchronous cellular system providing a holistic picture of the coordination of a complex system, even for the most transient and rare cell types. Such fine timing of both regulatory and cellular events suggest these coordination points might act as checkpoints between cell states.

4.2.11 Coordination points predict a checkpoint for B-cell developmental progression

Coordinated aspects of cell cycle and programmed cell death come into play throughout tissue homeostasis, to guide the transition through developmental checkpoints, as well as to control the size of each cellular compartment in a given tissue. Using Wanderlust to overlay simultaneously measured indicators of cell proliferation (Ki67) and apoptosis (cleaved poly ADP ribose polymerase– cPARP) revealed yet another level of functional coordination across these nascent human B cell fractions (Figure 4-12a). Just prior to the first inflection point (Figure 4-12a, red arrow), the cells are characterized by IL-7 dependent activation of STAT5 phosphorylation (Figure 4-10). Remarkably, this coordination point marks a transition from a state of high to low proliferation, as assessed by decreasing Ki67 expression (Figure 4-12a, background shade). This drop in proliferation leads directly into fraction IV where Vpre-B and λ5 rise (Figure 4-12a), signifying the transition into pro-B cells, where IgH gene locus rearrangement activity has been established (Figure 4-9c-d). While this pro-B cell checkpoint has never been clearly demonstrated in the human, we see it here for the first time as the cells transition between fractions III and IV (gain of CD24).

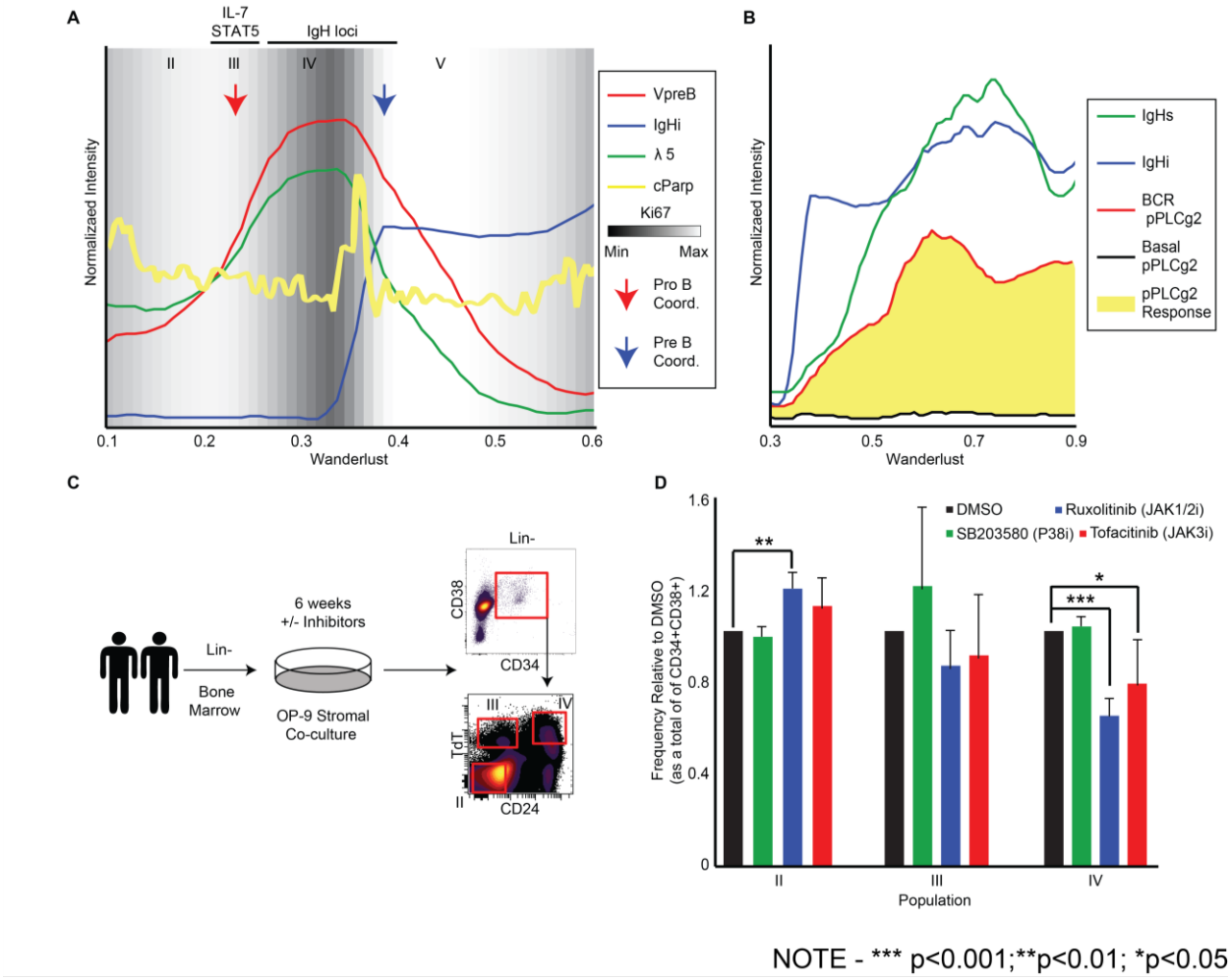


Figure 4-12. Regulatory signaling influences cell fate decisions in developing B cells.
Wanderlust traces for Vpre-B, λ5 and IgHi across the early trajectory. Background of plot shaded by heatbar indicating Ki67 (proliferative antigen) levels. Alignment of populations II-V across early trajectory indicates high Ki67 levels in population III just prior to Vpre-B and λ5 expression followed by quiescence during IgH rearrangement in population IV. Immediately prior to IgHi expression a spike in cleaved PARP (yellow line) occurs. This is followed by another burst of Ki67 expression. Red arrow indicates putative timing of pro-B cell coordination point. Blue arrow indicates putative timing of pre-B cell coordination point. (b) Wanderlust trace of the late trajectory showing IgH expression intracellularly (blue line) followed by IgH expression on the cell surface (green line). Basal levels of pPLCg2, a downstream marker of B cell receptor signaling is indicated in black. In response to cross linking of the B cell receptor, pPLCg2 is induced to the level of the red line and the fold change induction is shown in yellow. (c) Primary B cell co-culture was performed using two individual human bone marrows following lineage depletion cultured for 6 weeks on an OP-9 stromal layer in the presence or absence of inhibitors. Cells were analyzed by flow cytometry after the completion of the six-week culture period. (d) Frequency of cells occupying populations II-IV after six weeks of culture in the presence of DMSO control or targeted kinase inhibitors (SB203580-pP38i, Ruxolitinib-JAK1/2i, Tofacitinib-JAK1/3i). Treatment with Ruxolitinib results in increased cells occupying population II ($p<0.01$) and less cells occupying population IV ($p<0.001$). Treatment with tofacitinib also results in fewer cells in population IV compared to DMSO control ($p<0.05$).

The second coordination point occurs after the IgH locus has been completely rearranged, as indicated by the intracellular expression of IgH protein. Based on Ki67 the cells re-enter a state of proliferation as they pass through this checkpoint, expanding the pool of pre-B cells, which have productively formed an IgH (Figure 4-12a, blue arrow) [90]. Interestingly, just preceding this pre-B cell expansion, there is a discrete spike in cell death as indicated by a surge in single cells with higher cPARP (Figure 4-12a, yellow line), consistent with cells that could not form a productive IgH rearrangement and thus were unable to pass through this developmental checkpoint.

In concert with expression of Vpre-B and λ5, the newly expressed IgH now composes a complete pre-B cell receptor (pre-BCR). Mapping cells following B cell receptor cross-linking onto Wanderlust demonstrates that precisely paralleling the surface expression of the IgH (IgHs), cells are able to induce massive phospholipase C (PLC) gamma 2 phosphorylation (Figure 4-12b, red) as compared to the basal state (Figure 4-12b, black). Thus, with the presence of pre-BCR on the surface of the cells, they have yet again re-wired their regulatory signaling and have now become responsive to receptor cross-linking (Figure 4-12b, ~0.4 on Wanderlust).

4.2.12 *ex vivo differentiation assay confirms pro-B cell checkpoint*

We sought to further confirm the requirement of these checkpoints by interrogating the earliest pro-B cell checkpoint, as identified here, using an *ex vivo* differentiation assay (Figure 4-12c). Base on the re-wired regulatory signaling across fractions II to V (Figure 4-10), we hypothesized that a blockade of IL-7 dependent activation of STAT5 would inhibit the progression of cells through the pro-B checkpoint, in turn reducing the number of cells transiting from fraction II through the IL-7-dependent fraction III to fraction IV and beyond. Lin- human bone marrow samples from two donors with an additional depletion of BCR expressing lineages were cultured

with kinase inhibitors on OP-9 stromal cell feeders for six weeks and then the relative proportions of fractions II through IV was assessed by flow cytometry.

Both Ruxolitinib (JAK1/2 inhibitor) and Tofacitinib (JAK1/3 inhibitor) restricted progression of cells from population II through to Population IV (Figure 4-12d). Both JAK inhibitors significantly decreased the frequency of cells in fraction IV, relative to a DMSO control. At the same time, there was a significant accumulation of the cells in population II when incubated with JAK inhibitors. The P38 inhibitor, did not have a significant influence on the allocation of cells across the 3 fractions though it did promote significant, albeit compartment independent, cellular expansion.

Altogether, these *ex vivo* culture assays demonstrated that while the cells are capable of differentiation, as predicted in our pro-B cell checkpoint model, to successfully complete their developmental progression they depend on JAK mediated signals. Thus, free from the constraints of conventional genetics (i.e., gene- or cell-specific deletion), timing of expression and regulation (as simultaneously measured by single cell mass cytometry and ordered by Wanderlust) can identify physiologically relevant checkpoints and provide mechanistic information as to their regulation.

4.3 Materials and Methods

4.3.1 Primary Human Marrow

For mass cytometry analysis, fluorescence activated cell sorting (for qPCR), and ex vivo cell cultures, fresh whole human bone marrow (BM) was obtained from healthy donors from AllCells, Inc. (Emeryville, CA). Samples were Ficolled and processed as described in Bendall et al [12] and used fresh or cryopreserved. Cryopreserved cells were thawed in 90% RPMI with 10% Fetal calf serum (FCS) (supplemented with 20 U/mL sodium heparin (Sigma) and 0.025U/mL benzonase (Sigma),1X L-glutamine and 1X penicillin/streptomycin (Invitrogen). Fresh samples were re-suspended in 90% RPMI with 10% FCS.

4.3.2 Lineage Depletion

Where indicated, BM mononuclear cell (MNC) preparations were lineage depleted prior to cell culture or cytometric analysis. In all cases except for the ex vivo cell culture, the preparations were fixed with 1.6% formaldehyde for 10min (PFA; Electron Microscopy Sciences, Hatfield, PA) prior to depletion. Cells were then washed with staining media (CSM: PBS with 0.5% BSA, 0.02% sodium azide). For depletion, cells were stained with biotin-conjugated antibodies for 30 minutes at a concentration of 5 million cells per 100ul. Cells were washed with CSM twice then incubated with BD Streptavidin Particles Plus (BD Biosciences, San Jose, CA) at the manufacturer's recommended concentration for 30 minutes at room temperature. Particle-labeled cells were resuspended in CSM to approximately 2-8x107 cells/ml and placed in a magnetic holder for seven minutes. The supernatant was transferred to a new tube and the beads/cells were washed and resuspended and placed back in the magnetic holder for an additional round of

depletion and supernatant recovery. This washing procedure was repeated 2 times. Cells from the supernatant were then concentrated by centrifugation at 250g for 5 minutes.

4.3.3 Mass cytometry analysis

BM-MNCs were first stained for viability using cisplatin. Cells were then rested for 30min at 37°C and perturbed with various stimuli and inhibitors prior to analysis. Following perturbation, cells were immediately fixed with 1.6% paraformaldehyde (PFA; Electron Microscopy Sciences, Hatfield, PA) for 10 minutes. Cells were then washed with CSM and Fc receptor block was performed using Human TruStain FcX (Biolegend) following manufacturer's instructions. Cells were then stained for surface proteins at room temperature for 30min. Following staining, cells were washed twice with CSM and permeabilized with 4°C methanol for 10 minutes at 4°C. Cells were then washed twice with CSM and stained for intracellular proteins for 30 min at room temperature. Cells were washed with CSM and stained with 1 mL of 1:5000 of 2000x Ir DNA intercalator (DVS Sciences) diluted in PBS with 1.6% PFA for 20 minutes at room temperature up to overnight at 4°C.

4.3.4 Mass cytometry panel

The mass cytometry panel included the following markers:

Marker name	**Brief description**	**Ref**
Surface markers		
CD45	Pan-leukocyte marker	[59]
CD19	Pan-B-cell marker	[59]
CD22	Pan-B-cell marker	[59]

IgD	Immunoglobulin; marker of immature cells migrating out of the marrow	[89]
CD79b	Part of the BCR	[89]
CD20	Mature B-cells marker	[59]
CD34	Stem cells marker and marker of precursor stages of various cell lineages	[59]
CD179a	Vpre-B, part of the surrogate light chain	[111]
CD179b	Lambda5, part of the surrogate light chain	[111]
CD72	Pan-B-cell marker	[89]
IgM	BCR heavy chain	[89]
Kappa	Light chain component	[94]
Lambda	Light chain component	[94]
CD10	Precursor and early B-cells marker	[59]
CD49d	Integrin involved in homing to the bone-marrow niche	[112]
CD24	Pan-B-cell marker	[89]
CD127	IL-7-R, involved in pro-B cell signaling	[113]
CD38	B-cell progenitor marker	[114]
CD40	Antigen-presenting cells marker	[89]
CD117	Precursor cells marker	[59]
HLA-DR	Antigen-presenting cell surface receptor	[114]
Signaling markers		
pCreb	Transcription factor involved in pre-BCR-mediated cell expansion	[89]
pPLCg2	Lipid metabolizing effector enzyme. Generates second messenger	

	(IP_3) that activate Erk	
pSrc	Protein tyrosine kinase. Transmits the pre-BCR signal for cell expansion	
pAkt	Phosphorylates Foxo1, a transcription factor for Rag1/2, leading to its degradation after pro-B and pre-B cell rearrangements	
pSyk	Protein kinase. Activates pPLCg2	
pErk1/2	Extracellular signal-related kinase. Phosphorylates several transcriptional regulatory proteins involved in B-cell development	
pP38	Mitogen-activated protein kinase. Phosphorylates various transcription factors involved in B-cell development	
pStat5	Part of the IL-7 signaling pathway involved in pro-B cell development	[98]
Cell cycle markers		
cParp	An indicator of programmed cell death	[115]
Ki67	Proliferation marker	[116]
Genomic rearrangement enzymes		
Rag1	Activates V(D)J rearrangement along with Rag2	[94]
TdT	DNA polymerase that adds nucleotides to the 3' terminus of DNA	[59]

Table 4-1. List of markers used in mass cytometry

The table lists all of the markers used in the mass cytometry panel, organized by their function (cell surface, signaling, cell cycle and genomic rearrangement).

4.3.5 Mass cytometry analysis data pre-processing

Prior to CyTOF analysis cells were washed once with CSM and then twice with ddH2O. To make all samples maximally comparable, all data was acquired using internal metal isotope bead

standards. Cell events were acquired at approximately 500 events per second on a CyTOF I (DVS Sciences). Each patient sample was individually normalized to the internal bead standards prior to analysis. Mass cytometry data cleaning was performed at www.cytobank.org. To remove dead cells, debris, and non-B cell types data was gated based on cell length and DNA content as described in Bendall et al. [12] and for cisplatin negativity. Remaining cells were filtered for high expression levels of CD3 (T cells), CD33 (myeloid), CD11c (dendritic cells), CD16 (NK cells), CD235 (erythrocytes) and CD61 (platelets) prior to Wanderlust analysis or manual population gating.

4.3.6 Wanderlust parameters

The following parameters were used in all Wanderlust runs unless otherwise stated. As demonstrated in the results section, each of these values, including the early cell selection, can be set to a wide range of values while still resulting in accurate trajectory detection.

Parameter name	**Description**	**Value**
s	The early cell that is used in calculating the initial orientation trajectory Synthetic data: (1, 1, 1), the point from which the simulated trajectory originates. Mass cytometry healthy bone marrow data: the cell with *max(CD34-IgM)* *	
k	Number of neighbors of each node in k-nearest neighbors graph.	30
l	Number of neighbors selected for each node in each l-k-nearest neighbors graph in the ensemble	5
ng	Number of graphs in the l-k-nearest neighbors graph ensemble	20

nl	Number of landmarks	20
* This criterion was set in order to choose a stem cell (CD34+) while avoiding a possible stem cell-mature cell doublet (which would be CD34+IgM+, a non-physiological condition).		

Table 4-2. Default Wanderlust parameters
A list of the Wanderlust parameters, explanation of each parameter and the default parameter used in all experiments (unless otherwise stated).

All ten dimensions were used as input for the synthetic data. For the B-lineage data, the following surface markers were used: CD45, CD19, CD22, IgD, CD79b, CD20, CD34, CD179a, CD72, intracellular and surface IgM, Kappa, CD10, Lambda, CD179b, CD49d, CD24, CD127, CD38, CD40, CD117, HLADR.

The output trajectory was normalized using the following equation:

$$w_{norm} = \frac{w - p_5}{p_{95} - p_5}$$

where w is the output trajectory and p_i is the *ith* percentile of the output trajectory. By using the 5th/95th percentiles we avoid cases where a distant outlier cells skews the rest of the trajectory.

4.3.7 Cosine distance

Euclidean distance was used in Wanderlust runs on the synthetic data. However, marker intensities in data acquired via mass cytometry are given in arbitrary units, that is, the scaling of each channel is independent of the other channels. Therefore, the proportions between markers are arbitrary as well (figure 4-13a). As a result, measuring cell-to-cell distances using a norm (such as the L^1 norm, cityblock distance, or the L^2 norm, Euclidean distance) could lead to different distance distributions depending on the specific scale (figure 4-13b, top). Additionally, due to the varying scales, markers become weighed by the dynamic range of their acquisition channel, which is unrelated to physical dynamic range or to the marker's importance.

To address this pro-Blem, we provide Wanderlust with the cosine distance, which is defined for a pair of cells (*s, t)* as:

$$d_{(s,t)} = 1 - \frac{x_s x_t{}'}{\sqrt{(x_s x_s')(x_t x_t')}}$$

where x_i is the vector representing cell *i*. $d_{(s,t)}$ is equal to *1-cos(θ)*, where θ is the angle between x_s and x_t in the high-dimensional space, hence the name cosine distance. Cosine distances are scale-independent and their distance distribution remains relatively constant after scaling changes (figure 4-13b, bottom).

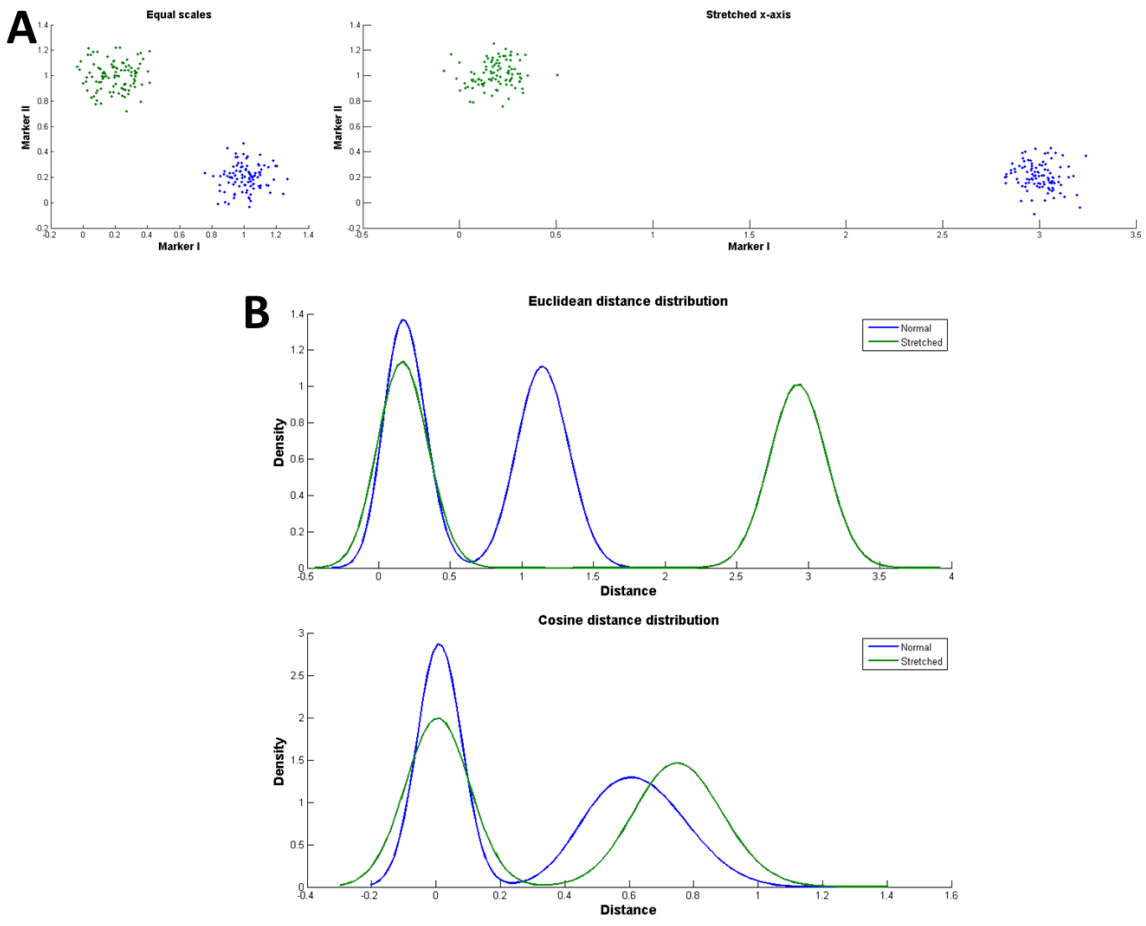


Figure 4-13. Cosine distances are scale-independent.

(a) Left: In this toy example two separate clusters have been randomly sampled from a two-dimensional Gaussian with center (0.2, 1) (green) or (1, 0.2) (blue). The two dimensions have equal scales. Right: The X-axis has been stretched by a factor of three, shifting the blue cluster to the right. (b) Top: The L2 norm (Euclidean distance) distribution in the equal scale data (blue) and the stretch X-axis data (green). We see that intra-cluster distances remain the same, while inter-cluster distances are shifted by the same factor as the stretching of the X-axis. Bottom: The cosine distance distribution in each dataset. Intra-cluster distances still remain the same and the inter-cluster distances are much closer to each other.

4.3.8 Calculation of marker trace across the trajectory

We divided the Wanderlust trajectory into 100 equidistant windows. The width of each window was equal to 8% of the total trajectory width (each window included all cells whose trajectory score was +/-0.04 of the window's center). For each marker, the marker's trace is defined as the marker median in each window, normalized to the 0-1 range by subtracting the minimum and dividing by the maximum marker value.

4.3.9 Cross-correlation of trajectories across individuals

Cell subtype proportions across the developmental trajectory might change between individuals due to many factors, such as genetics or recent exposure to pathogens. As a result regions of the Wanderlust trajectory might become shorter or longer, depending on proportions between cell subtypes (figure 4-14a). In order to synchronize the trajectory across individuals, we calculated the cross-correlation between each sample and an arbitrarily chosen base sample. The cross-correlation was calculated as the mean of all marker cross-correlations. Then, the trajectory score of each cell in each sample was shifted by the value that maximized the cross-correlation (figure 4-14b).

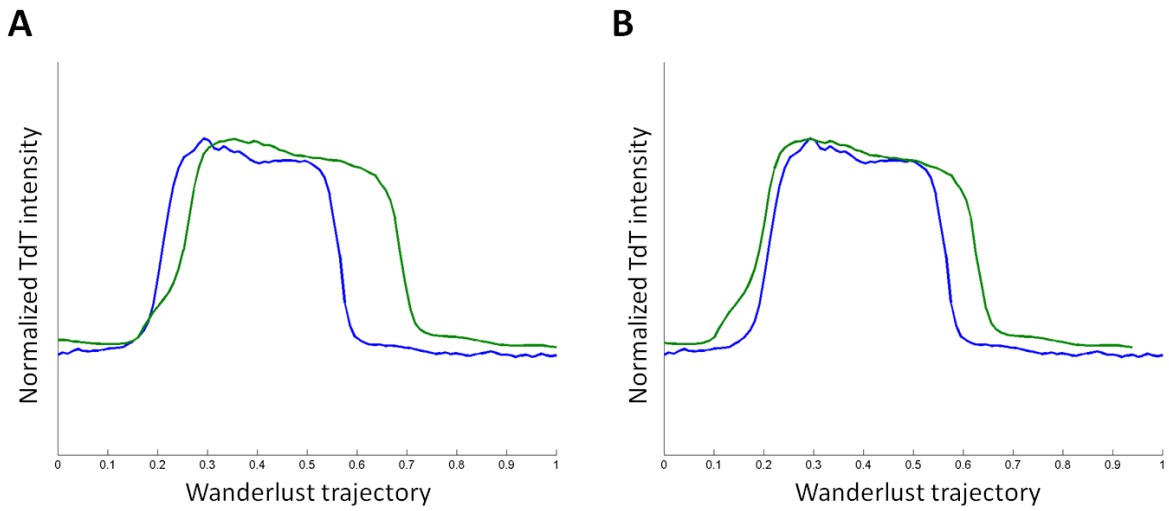


Figure 4-14. Cross-correlation allows comparison of trajectories between samples.
Wanderlust has been applied to two healthy bone marrow samples, in blue and green, respectively. (a) The TdT trace over the trajectory in each sample, before cross-correlation shift. The green sample has more TdT^+ cells, leading to a wider TdT^+ region in the trajectory. Pearson's ρ=0.8. (b) After shifting by the maximal cross-correlation between the two samples, the TdT^+ section has a higher overlap. Pearson's ρ=0.89.

4.4 Discussion

Wanderlust detects the trajectory, the underlying temporal element, in a system. The algorithm is resilient to noise, consistent between samples and scalable to up to tens of millions of points. It extracts the trajectory from a snapshot of the system rather than from time-series data and only requires an approximate starting point as prior information. The Wanderlust trajectory is continuous; in addition to mapping stable cell states, it also provides information about the transitions between them. The combination of these characteristics makes Wanderlust ideal for the exploration of any system that undergoes a developmental process.

We have shown that Wanderlust finds the developmental trajectory of human B-cell development in the bone marrow. The trajectory is consistent with our current understanding of this process. Furthermore, Wanderlust provides a quantitative, high-resolution ordering of surface marker expression, signaling and recombination events, including markers whose timing

and relevance was previously unknown. In addition to validating many observations from mice in human, Wanderlust unveils a hitherto inaccessible systems-level landscape of early B-cell development: the coordination between the many mechanisms that regulate this process, including surface marker expression, signaling, proliferation and apoptosis. The Wanderlust trajectory represents the most comprehensive analysis of the human system to date, unifying all relevant cellular features and regulatory behaviors of early B-cell development in the human, and lays a roadmap for its further exploration.

The continuous approach diverges from existing models of development. The trajectory captures expression as trends: markers rise and fall in patterns that correspond to the cell's behavior. Instead of examining cell stages, we see transitions that cannot be classified under a discrete partitioning. This perspective can also be extended to the relationships between markers. We can utilize the trajectory in order to examine regulation in a holistic manner, as the coordination of and interaction between the different regulatory mechanisms. By providing a unified framework, the trajectory highlights transition periods when the cell is undergoing a regulatory or genetic transformation.

4.4.1 Detection of more complicated trajectory structures

Branching is an exciting future direction for Wanderlust. The current version of the algorithm assumes that the developmental process is composed of a series of consecutive stages that lead to a single fate. A more sophisticated model could aid in the exploration of systems that involve branching at different stages of the development process and lead to multiple fates. One approach for the design of such a model will be to follow chains of cells that break the linearity assumption. The model will start with multiple "naïve" Wanderlust runs from different points across the system and look for triplets of distances where A equals B, B equals C, but A does not

equal C. This final inequality signifies a potential branching point between the triplet's members. Multiple such relationships will pinpoint the branching point's position.

In the context of early B-cell development, such an extension might be able to identify the development of B-1 cells, which are B lymphocytes that are involved in the humoral immune response [113, 117]. Most of the body's population of B-1 cells originates from fetal precursors that undergo constant proliferation; however, pre-B cells could develop into B-1 cells with very low efficiency [118, 119]. The developmental process of B-1 cells in adults and the purpose of this process are still unclear. A Wanderlust algorithm that incorporates branching should be able to detect such a tiny population. Additional applications for a branching Wanderlust include the development of activated B cells in the germinal center and their branching into different antibody isotypes, the complete immune system, or the progression of other types of stem cells.

4.4.2 Application of the developmental trajectory to disease

The intimate connection between the healthy and the abnormal is exemplified by cancer, a condition which originates from a deviant developmental process. As such, mapping the normal immune system lays the foundation for understanding the disease. The trajectory lead to identification of coordination points where the cell transitions from one developmental stage to another, coupled to the regulatory signaling involved in this process. This suggests specific periods of risk for transforming to malignancy, should regulatory protections fail. The next step would be to harness this information, for example, in the experimental dissection of cancer by interfering with the key players identified through Wanderlust in critical moments through development. The newly-gained understanding of normal B-cell development will be instrumental in exploring disease originating from the B cell lineage.

A branching version of Wanderlust could be applied directly to single-cell measurements of malignant samples. The algorithm will detect the disease's developmental process, which will in turn be overlaid on top of the healthy trajectory using methods similar to those we utilized in comparing normal samples. By applying the algorithm to cancer in its earliest stages, for example a sample taken from a mouse, we should be able to identify whether the cancer originated from a specific early cell stage, which regulatory mechanisms have been abolished by the disease and which are still operational. Alternatively, the Wanderlust trajectory of a more advanced form of the disease will shed new light on its corrupted developmental process, how its cells proliferate and whether and how they die. Answering these questions would assist in directing future research toward the most promising treatments.

4.4.3 A universe of development

Wanderlust requires very few pieces of information in order to detect the trajectory: single-cell measurements and a starting point. While we examined the highly-studied system of B-cell development, the algorithm could be applied to less-explored systems such as the development of induced-pluripotent stem cells (iPSCs). The analysis in that context will proceed along similar steps to those taken with B-cell development: a population of iPSCs will be grown and assayed via a mass cytometer, possibly at different time points. This system has the distinct advantage that the selection of the starting cell is trivial. The Wanderlust trajectory will lead to new insights on how iPSCs differentiate into the various lineages and how their development differs from regular stem cells. A more exciting possibility is detecting the trajectory a somatic cell undergoes following transfection with stem cell genes, revealing the details of the process in which iPSCs are created.

Two features of the algorithm are advantageous in the context of unfamiliar developmental systems. First, we supplied Wanderlust with an early starting cell (a CD34+ CD38- stem cell). However, as seen in figure 4-5, the algorithm could instead begin from a late cell and map the trajectory from a known finale back to its beginning. This variation is relevant in the context of non-hematopoietic development, where the stem cells are not known but where the mature cells are both plentiful and easily identified, such as in mesenchymal development. Two, overlaying known information over the combination of multiple marker traces pinpoints the position of *de novo* transition states. As in the present work, validation of these states could be done experimentally. The combination of trajectory detection from a mature cell and demarcation of transition states will be pivotal in exploring a new system.

Finally, Wanderlust could be utilized outside the realms of biology, in any situation in which we are exploring a dynamic system but are unable to acquire time-series data. The algorithm is a hypothesis-generation method with broad applications.

One of the biggest challenges in research is deciding where to look. In order to ask the right questions, we need some initial understanding of the structure of the pro-Blem being examined. The existence of a developmental process provides us with a powerful scaffold to that end. Based on this insight, the trajectory captures the central dynamic in the data: Wanderlust provides the crucial starting point needed to eventually untangle the entire system.

Chapter 5 Conclusions

The advent of high-dimensional, single-cell methods marks a new era in the life sciences. Flow cytometry and related microscopy techniques have taught us to appreciate the roles of heterogeneity and stochasticity, two concepts that were the realm of theory until a decade ago. Recent advances push the number of dimensions ever higher by abandoning the limited fluorescence-based protocols in favor of novel methods. Mass cytometry, single-cell RNA qualification, and other technologies can assay up to a hundred parameters in each cell, opening a window into mechanisms that were previously inaccessible. However, new technologies lead to new challenges as current statistical tools cannot handle the tsunami of new data being generated. In order to reach the full potential of the latest breakthrough technologies, we need to adopt a fresh outlook on how to process, analyze and integrate "big" biological data.

Here we presented two new computational methods for the exploration of high-dimensional data. The first, viSNE, is a dimensionality reduction algorithm that maps the data into two dimensions. By projecting complex relationships into a familiar scatter plot, viSNE provides a visualization of information that would not be immediately accessible otherwise. The viSNE map highlights the separation between different cell subtypes (both healthy and malignant), draws the progression of cancer from diagnosis to relapse and accentuates a tiny cancer population in a minimal residual disease setting. The second algorithm, Wanderlust, detects the developmental ordering (the trajectory) of cells without requiring any time-series experiments. We used Wanderlust to identify the progression of healthy B cell development. It reveals the coordination between multiple regulatory mechanisms in key checkpoints across development (such as the

coordination between signaling, proliferation and apoptosis in the pro B-cell checkpoint). These two algorithms translate nearly impervious data into intuitive constructs.

Taken together, viSNE and Wanderlust offer a completely new outlook on biological data. Classic experimentation techniques follow the production, regulation, function and degradation of a handful of molecules at a time. Genomic-based techniques, such as microarrays and sequencing, scale this philosophy into a large set of assayed molecules. However, they still examine individual molecules, or, at most, clusters or modules of molecules. The work presented here is a leap toward a true systematic view that takes the entirety of information and synthesizes it into a single coherent picture. These two algorithms incorporate all of the parameters examined into a single profile (with viSNE) or trajectory (in the case of Wanderlust). The discussion no longer revolves around a single protein or even a group of genes, but rather around their combination, and how this combination changes after perturbation by a treatment or by disease. Furthermore, since viSNE and Wanderlust preserve the single-cell resolution, we can use their low-dimensional representation in the design of validation experiments using well-established, classic approaches. Biology did not evolve one molecule at a time. To reveal its full potential and get a comprehensive understanding of life, we must follow this holistic philosophy.

Technology is advancing in a staggering pace: we are constantly breaking the boundaries of what is possible and measure more parameters in higher resolutions than ever. This trend will only accelerate as time goes by. We have shown that viSNE and Wanderlust are robust in the algorithmic sense; however, maybe more importantly, they are also robust to the advent of our technology. Both algorithms are theoretically capable of processing tens of thousands of dimensions over millions of cells in reasonable computational times. As biology enters the realms of “big data”, this scalability-oriented mindset will prove critical. Any analytical tool that

ignores it will become obsolete, along with whatever instrumentation existed at the time of its conception. The design of viSNE and Wanderlust guarantees that they will stay in the forefront for years to come.

The life sciences are undergoing a paradigm shift. When our starting point was confined by the available technology we were prone to lock into accepted dogma. However, we are entering a new era, where entire biological systems can be profiled and visualized using high-dimensional, single-cell technologies. The current work is a beginning, a proof of concept for exploring a small subset of what is possible. It leads to many further questions: How is the complete hematopoietic developmental process regulated? What is the role of coordination points in the emergence of other cell types? How are these processes perturbed by cancer and by immune deficiencies? Furthermore, we should not restrict ourselves to the immune system. Every organ in our body is a microcosm that can be charted using viSNE and whose trajectory detected with Wanderlust. In an ideal scenario, we will follow the greatest developmental process of all, the growth of a fetus and the differentiation of all of the many subsystems in our bodies. viSNE and Wanderlust are harbingers of this exciting new fountain of biological knowledge.

References

1. Murphy, K., P. Travers, and M. Walport, *Janeway's Immunobiology, chapter 1*. 7th Edition ed. 2008, New York: Garland Science.
2. Kondo, M., et al., *Biology of hematopoietic stem cells and progenitors: implications for clinical application.* Annu Rev Immunol, 2003. **21**: p. 759-806.
3. Irish, J.M., et al., *Single cell profiling of potentiated phospho-protein networks in cancer cells.* Cell, 2004. **118**(2): p. 217-28.
4. Sachs, K., et al., *Causal protein-signaling networks derived from multiparameter single-cell data.* Science, 2005. **308**(5721): p. 523-9.
5. Majeti, R., C.Y. Park, and I.L. Weissman, *Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood.* Cell Stem Cell, 2007. **1**(6): p. 635-45.
6. Tarnok, A., H. Ulrich, and J. Bocsi, *Phenotypes of stem cells from diverse origin.* Cytometry A, 2010. **77**(1): p. 6-10.
7. O'Brien, C.A., A. Kreso, and J.E. Dick, *Cancer stem cells in solid tumors: an overview.* Semin Radiat Oncol, 2009. **19**(2): p. 71-7.
8. Bendall, S.C., et al., *A deep profiler's guide to cytometry.* Trends Immunol, 2012. **33**(7): p. 323-32.
9. Gattinoni, L., et al., *A human memory T cell subset with stem cell-like properties.* Nat Med, 2011. **17**(10): p. 1290-7.
10. Mahnke, Y., P. Chattopadhyay, and M. Roederer, *Publication of optimized multicolor immunofluorescence panels.* Cytometry A, 2010. **77**(9): p. 814-8.
11. Roederer, M., *Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats.* Cytometry, 2001. **45**(3): p. 194-205.
12. Bendall, S.C., et al., *Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum.* Science, 2011. **332**(6030): p. 687-96.
13. Benoist, C. and N. Hacohen, *Immunology. Flow cytometry, amped up.* Science, 2011. **332**(6030): p. 677-8.
14. Bandura, D.R., et al., *Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry.* Anal Chem, 2009. **81**(16): p. 6813-22.
15. Lou, X., et al., *Polymer-based elemental tags for sensitive bioassays.* Angew Chem Int Ed Engl, 2007. **46**(32): p. 6111-4.
16. Cornett, D.S., et al., *MALDI imaging mass spectrometry: molecular snapshots of biochemical systems.* Nat Methods, 2007. **4**(10): p. 828-33.
17. Mercer, J., et al., *RNAi Screening Reveals Proteasome- and Cullin3-Dependent Stages in Vaccinia Virus Infection.* Cell Rep, 2012. **2**(4): p. 1036-47.
18. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nat Rev Genet, 2011. **12**(2): p. 87-98.
19. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
20. Dalerba, P., et al., *Single-cell dissection of transcriptional heterogeneity in human colon tumors.* Nat Biotechnol, 2011. **29**(12): p. 1120-7.

21. Lubeck, E. and L. Cai, *Single-cell systems biology by super-resolution imaging and combinatorial labeling.* Nat Methods, 2012. **9**(7): p. 743-8.
22. Bendall, S.C. and G.P. Nolan, *From single cells to deep phenotypes in cancer.* Nat Biotechnol, 2012. **30**(7): p. 639-47.
23. McClean, M.N., et al., *Cross-talk and decision making in MAP kinase pathways.* Nat Genet, 2007. **39**(3): p. 409-14.
24. Amit, I., et al., *A module of negative feedback regulators defines growth factor signaling.* Nat Genet, 2007. **39**(4): p. 503-12.
25. Van der Maaten, L., E. Postma, and H. Van Den Herik, *Dimensionality reduction: A comparative review.* Journal of Machine Learning Research, 2009. **10**: p. 1-41.
26. Pan, S.J., J.T. Kwok, and Q. Yang. *Transfer Learning via Dimensionality Reduction*. in *AAAI*. 2008.
27. Bengio, Y., *Learning Deep Architectures for AI.* Found. Trends Mach. Learn., 2009. **2**(1): p. 1-127.
28. Erhan, D., et al., *Why Does Unsupervised Pre-training Help Deep Learning?* J. Mach. Learn. Res., 2010. **11**: p. 625-660.
29. Jolliffe, I., *Principal component analysis*. 2005: Wiley Online Library.
30. Grassberger, P. and I. Procaccia, *Measuring the strangeness of strange attractors.* Physica D: Nonlinear Phenomena, 1983. **9**(1): p. 189-208.
31. Fukunaga, K. and D.R. Olsen, *An algorithm for finding intrinsic dimensionality of data.* Computers, IEEE Transactions on, 1971. **100**(2): p. 176-183.
32. Pettis, K.W., et al., *An intrinsic dimensionality estimator from near-neighbor information.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1979(1): p. 25-37.
33. Harman, H.H., *Modern factor analysis.* 1960.
34. Lau, K.S., et al., *In vivo systems analysis identifies spatial and temporal aspects of the modulation of TNF-alpha-induced apoptosis and proliferation by MAPKs.* Sci Signal, 2011. **4**(165): p. ra16.
35. Bar-Even, A., et al., *Noise in protein expression scales with natural protein abundance.* Nat Genet, 2006. **38**(6): p. 636-43.
36. Andersen, T., et al., *Ecological thresholds and regime shifts: approaches to identification.* Trends in Ecology & Evolution, 2009. **24**(1): p. 49-57.
37. Raytchev, B., I. Yoda, and K. Sakaue. *Head pose estimation by nonlinear manifold learning*. in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. 2004. IEEE.
38. Jeong, M., J.H. Choi, and B.H. Koh, *Isomap-based damage classification of cantilevered beam using modal frequency changes.* Structural Control and Health Monitoring, 2013.
39. Mosconi, F., et al., *Some nonlinear challenges in biology.* Nonlinearity, 2008. **21**(8): p. T131.
40. Levin, S.A., et al., *Mathematical and computational challenges in population biology and ecosystems science.* Science, 1997. **275**(5298): p. 334-43.
41. Burges, C.J., *Dimension reduction: A guided tour.* Machine Learning, 2009. **2**(4): p. 275-365.
42. Lee, J.M., *Introduction to smooth manifolds*. Vol. 218. 2012: Springer.
43. Saul, L.K., et al., *Spectral methods for dimensionality reduction.* Semisupervised learning, 2006: p. 293-308.

44. Tenenbaum, J.B., V. de Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction.* Science, 2000. **290**(5500): p. 2319-23.
45. Kruskal, J.B., *Nonmetric multidimensional scaling: a numerical method.* Psychometrika, 1964. **29**(2): p. 115-129.
46. Weinberger, K.Q. and L.K. Saul, *Unsupervised learning of image manifolds by semidefinite programming.* International Journal of Computer Vision, 2006. **70**(1): p. 77-90.
47. Shawe-Taylor, J. and N. Cristianini, *Kernel methods for pattern analysis*. 2004: Cambridge university press.
48. Paprotny, A. and J. Garcke. *On a connection between maximum variance unfolding, shortest path pro-Blems and isomap*. in *International Conference on Artificial Intelligence and Statistics*. 2012.
49. Steinwart, I. and A. Christmann, *Support vector machines*. 2008: Springer.
50. Roweis, S.T. and L.K. Saul, *Nonlinear dimensionality reduction by locally linear embedding.* Science, 2000. **290**(5500): p. 2323-6.
51. Donoho, D.L. and C. Grimes, *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data.* Proceedings of the National Academy of Sciences, 2003. **100**(10): p. 5591-5596.
52. Hinton, G.E. and S.T. Roweis. *Stochastic neighbor embedding*. in *Advances in neural information processing systems*. 2002.
53. Amir el, A.D., et al., *viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia.* Nat Biotechnol, 2013. **31**(6): p. 545-52.
54. Petilla Interneuron Nomenclature, G., et al., *Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex.* Nat Rev Neurosci, 2008. **9**(7): p. 557-68.
55. Herzenberg, L.A., et al., *Interpreting flow cytometry data: a guide for the perplexed.* Nat Immunol, 2006. **7**(7): p. 681-5.
56. Qiu, P., et al., *Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE.* Nat Biotechnol, 2011. **29**(10): p. 886-91.
57. Qian, Y., et al., *Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data.* Cytometry B Clin Cytom, 2010. **78 Suppl 1**: p. S69-82.
58. Pyne, S., et al., *Automated high-dimensional flow cytometric data analysis.* Proc Natl Acad Sci U S A, 2009. **106**(21): p. 8519-24.
59. van Lochem, E.G., et al., *Immunophenotypic differentiation patterns of normal hematopoiesis in human bone marrow: reference patterns for age-related changes and disease-induced shifts.* Cytometry B Clin Cytom, 2004. **60**(1): p. 1-13.
60. Wakita, S., et al., *Mutations of the epigenetics modifying gene (DNMT3a, TET2, IDH1/2) at diagnosis may induce FLT3-ITD at relapse in de novo acute myeloid leukemia.* Leukemia, 2012.
61. Campana, D., *Status of minimal residual disease testing in childhood haematological malignancies.* Br J Haematol, 2008. **143**(4): p. 481-9.

62. Borowitz, M.J., et al., *Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: a Children's Oncology Group study.* Blood, 2008. **111**(12): p. 5477-85.
63. Ossenkoppele, G.J., A.A. van de Loosdrecht, and G.J. Schuurhuis, *Review of the relevance of aberrant antigen expression by flow cytometry in myeloid neoplasms.* Br J Haematol, 2011. **153**(4): p. 421-36.
64. Loken, M.R., et al., *Residual disease detected by multidimensional flow cytometry signifies high relapse risk in patients with de novo acute myeloid leukemia: a report from Children's Oncology Group.* Blood, 2012. **120**(8): p. 1581-8.
65. Bodenmiller, B., et al., *Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators.* Nat Biotechnol, 2012. **30**(9): p. 858-67.
66. Van der Maaten, L., *An introduction to dimensionality reduction using matlab.* Report, 2007. **1201**: p. 07-07.
67. Van der Maaten, L. and G. Hinton, *Visualizing data using t-SNE.* Journal of Machine Learning Research, 2008. **9**(2579-2605): p. 85.
68. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of Statistical Mechanics: Theory and Experiment, 2008. **2008**(10): p. P10008.
69. Kotecha, N., P.O. Krutzik, and J.M. Irish, *Web-based analysis and publication of flow cytometry experiments.* Curr Protoc Cytom, 2010. **Chapter 10**: p. Unit10 17.
70. Elowitz, M.B., et al., *Stochastic gene expression in a single cell.* Science, 2002. **297**(5584): p. 1183-6.
71. Raser, J.M. and E.K. O'Shea, *Control of stochasticity in eukaryotic gene expression.* Science, 2004. **304**(5678): p. 1811-4.
72. Kafri, R., et al., *Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle.* Nature, 2013. **494**(7438): p. 480-3.
73. Friedman, J.H. and W. Stuetzle, *Projection pursuit regression.* Journal of the American statistical Association, 1981. **76**(376): p. 817-823.
74. Huber, P.J., *Projection pursuit.* The annals of Statistics, 1985: p. 435-475.
75. Lee, E.-K., et al., *Projection pursuit for exploratory supervised classification.* Journal of Computational and Graphical Statistics, 2005. **14**(4).
76. Nason, G., *Three-dimensional projection pursuit.* Applied Statistics, 1995: p. 411-430.
77. Friedman, J.H., *Exploratory projection pursuit.* Journal of the American statistical Association, 1987. **82**(397): p. 249-266.
78. Jimenez, L.O. and D. Landgrebe, *High dimensional feature reduction via projection pursuit.* ECE Technical Reports, 1996: p. 103.
79. Yuan, Y., et al., *Prediction of CCR5 receptor binding affinity of substituted 1-(3, 3-diphenylpropyl)-piperidinyl amides and ureas based on the heuristic method, support vector machine and projection pursuit regression.* European journal of medicinal chemistry, 2009. **44**(1): p. 25-34.
80. Ren, Y., et al., *QSPR study on the melting points of a diverse set of potential ionic liquids by projection pursuit regression.* QSAR & Combinatorial Science, 2009. **28**(11-12): p. 1237-1244.
81. Clements, A.-M. and M. Jones, *An ecological example of the application of projection pursuit to compositional data.* Vegetatio, 1991. **95**(2): p. 101-107.

82. Christiansen, B., *Is the atmosphere interesting? A projection pursuit study of the circulation in the northern hemisphere winter.* Journal of Climate, 2009. **22**(5): p. 1239-1254.
83. Leban, G., et al., *Vizrank: Data visualization guided by machine learning.* Data Mining and Knowledge Discovery, 2006. **13**(2): p. 119-136.
84. Faith, J., R. Mintram, and M. Angelova, *Targeted projection pursuit for visualizing gene expression data classifications.* Bioinformatics, 2006. **22**(21): p. 2667-73.
85. Dijkstra, E.W., *A note on two pro-Blems in connexion with graphs.* Numerische mathematik, 1959. **1**(1): p. 269-271.
86. Leiserson, C.E., et al., *Introduction to algorithms, section 22.1*. 2001: The MIT press.
87. Vaughan, A.T., A. Roghanian, and M.S. Cragg, *B cells--masters of the immunoverse.* Int J Biochem Cell Biol, 2011. **43**(3): p. 280-5.
88. Hardy, R.R. and K. Hayakawa, *B cell development pathways.* Annu Rev Immunol, 2001. **19**: p. 595-621.
89. LeBien, T.W. and T.F. Tedder, *B lymphocytes: how they develop and function.* Blood, 2008. **112**(5): p. 1570-80.
90. LeBien, T.W., *Fates of human B-cell precursors.* Blood, 2000. **96**(1): p. 9-23.
91. Murphy, K., P. Travers, and M. Walport, *Janeway's Immunobiology, chapter 9.* 7th Edition ed. 2008, New York: Garland Science.
92. Alt, F.W., et al., *VDJ recombination.* Immunol Today, 1992. **13**(8): p. 306-14.
93. Willerford, D.M., W. Swat, and F.W. Alt, *Developmental regulation of V(D)J recombination and lymphocyte differentiation.* Curr Opin Genet Dev, 1996. **6**(5): p. 603-9.
94. Alt, F.W., et al., *Ordered rearrangement of immunoglobulin heavy chain variable region segments.* EMBO J, 1984. **3**(6): p. 1209-19.
95. Tonegawa, S., *Somatic generation of antibody diversity.* Nature, 1983. **302**(5909): p. 575-81.
96. Schatz, D.G., M.A. Oettinger, and D. Baltimore, *The V(D)J recombination activating gene, RAG-1.* Cell, 1989. **59**(6): p. 1035-48.
97. Nutt, S.L. and B.L. Kee, *The transcriptional regulation of B cell lineage commitment.* Immunity, 2007. **26**(6): p. 715-25.
98. Malin, S., S. McManus, and M. Busslinger, *STAT5 in B cell development and leukemia.* Curr Opin Immunol, 2010. **22**(2): p. 168-76.
99. Oettinger, M.A., et al., *RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination.* Science, 1990. **248**(4962): p. 1517-23.
100. Monroe, J.G., *ITAM-mediated tonic signalling through pre-BCR and BCR complexes.* Nat Rev Immunol, 2006. **6**(4): p. 283-94.
101. Milne, C.D. and C.J. Paige, *IL-7: a key regulator of B lymphopoiesis.* Semin Immunol, 2006. **18**(1): p. 20-30.
102. Novershtern, N., et al., *Densely interconnected transcriptional circuits control cell states in human hematopoiesis.* Cell, 2011. **144**(2): p. 296-309.
103. Okuno, Y., et al., *Differential regulation of the human and murine CD34 genes in hematopoietic stem cells.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 6246-51.
104. Fry, T.J. and C.L. Mackall, *Interleukin-7: from bench to clinic.* Blood, 2002. **99**(11): p. 3892-904.

105. Kang, J. and S.D. Der, *Cytokine functions in the formative stages of a lymphocyte's life.* Curr Opin Immunol, 2004. **16**(2): p. 180-90.
106. Parrish, Y.K., et al., *IL-7 Dependence in human B lymphopoiesis increases during progression of ontogeny from cord blood to bone marrow.* J Immunol, 2009. **182**(7): p. 4255-66.
107. Corfe, S.A. and C.J. Paige, *The many roles of IL-7 in B cell development; mediator of survival, proliferation and differentiation.* Semin Immunol, 2012. **24**(3): p. 198-208.
108. Johnson, S.E., et al., *Murine and human IL-7 activate STAT5 and induce proliferation of normal human pro-B cells.* J Immunol, 2005. **175**(11): p. 7325-31.
109. Bertolino, E., et al., *Regulation of interleukin 7-dependent immunoglobulin heavy-chain variable gene rearrangements by transcription factor STAT5.* Nat Immunol, 2005. **6**(8): p. 836-43.
110. Cobaleda, C. and I. Sanchez-Garcia, *B-cell acute lymphoblastic leukaemia: towards understanding its cellular origin.* Bioessays, 2009. **31**(6): p. 600-9.
111. Karasuyama, H., A. Rolink, and F. Melchers, *Surrogate light chain in B cell development.* Adv Immunol, 1996. **63**: p. 1-41.
112. Bonig, H., et al., *Increased numbers of circulating hematopoietic stem/progenitor cells are chronically maintained in patients treated with the CD49d blocking antibody natalizumab.* Blood, 2008. **111**(7): p. 3439-3441.
113. Hardy, R.R., P.W. Kincade, and K. Dorshkind, *The protean nature of cells in the B lymphocyte lineage.* Immunity, 2007. **26**(6): p. 703-14.
114. Doulatov, S., et al., *Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development.* Nat Immunol, 2010. **11**(7): p. 585-593.
115. Yu, S.W., et al., *Mediation of poly(ADP-ribose) polymerase-1-dependent cell death by apoptosis-inducing factor.* Science, 2002. **297**(5579): p. 259-63.
116. Scholzen, T. and J. Gerdes, *The Ki-67 protein: from the known and the unknown.* J Cell Physiol, 2000. **182**(3): p. 311-22.
117. Martin, F. and J.F. Kearney, *B1 cells: similarities and differences with other B cell subsets.* Curr Opin Immunol, 2001. **13**(2): p. 195-201.
118. Tung, J.W., et al., *Phenotypically distinct B cell development pathways map to the three B cell lineages in the mouse.* Proc Natl Acad Sci U S A, 2006. **103**(16): p. 6293-8.
119. Hardy, R.R., *B-1 B cell development.* J Immunol, 2006. **177**(5): p. 2749-54.