Learning Cell States from High-Dimensional Single-Cell Data

by

Jacob H. Levine

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Graduate School of Arts and Sciences

Columbia University

2016

© 2016

Jacob H. Levine

All rights reserved

Abstract

Learning Cell States from High-Dimensional Single-Cell Data

Jacob H. Levine

Recent developments in single-cell measurement technologies have yielded dramatic increases in throughput (measured cells per experiment) and dimensionality (measured features per cell). In particular, the introduction of mass cytometry has made possible the simultaneous quantification of dozens of protein species in millions of individual cells in a single experiment. The raw data produced by such high-dimensional single-cell measurements provide unprecedented potential to reveal the phenotypic heterogeneity of cellular systems. In order to realize this potential, novel computational techniques are required to extract knowledge from these complex data.

Analysis of single-cell data is a new challenge for computational biology, as early development in the field was tailored to technologies that sacrifice single-cell resolution, such as DNA microarrays. The challenges for single-cell data are quite distinct and require multidimensional modeling of complex population structure. Particular challenges include nonlinear relationships between measured features and non-convex subpopulations.

This thesis integrates methods from computational geometry and network analysis to develop a framework for identifying the population structure in high-dimensional single-cell data. At the center of this framework is PhenoGraph, and algorithmic approach to defining subpopulations, which when applied to healthy bone marrow data was shown to reconstruct known

immune cell types automatically without prior information. PhenoGraph demonstrated superior accuracy, robustness, and efficiency, compared to other methods.

The data-driven approach becomes truly powerful when applied to less characterized systems, such as malignancies, in which the tissue diverges from its healthy population composition. Applying PhenoGraph to bone marrow samples from a cohort of acute myeloid leukemia (AML) patients, the thesis presents several insights into the pathophysiology of AML, which were extracted by virtue of the computational isolation of leukemic subpopulations. For example, it is shown that leukemic subpopulations diverge from healthy bone marrow but not without bound: Leukemic cells are apparently free to explore only a restricted phenotypic space that mimics normal myeloid development. Further, the phenotypic composition of a sample is associated with its cytogenetics, demonstrating a genetic influence on the population structure of leukemic bone marrow.

The thesis goes on to show that functional heterogeneity of leukemic samples can be computationally inferred from molecular perturbation data. Using a variety of methods that build on PhenoGraph's foundations, the thesis presents a characterization of leukemic subpopulations based on an inferred stem-like signaling pattern. Through this analysis, it is shown that surface phenotypes often fail to reflect the true underlying functional state of the subpopulation, and that this functional stem-like state is in fact a powerful predictor of survival in large, independent cohorts.

Altogether, the thesis takes the existence and importance of cellular heterogeneity as its starting point and presents a mathematical framework and computational toolkit for analyzing samples from this perspective. It is shown that phenotypic and functional heterogeneity are robust characteristics of acute myeloid leukemia with clinically significant ramifications.

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

Rare is the single cell that exists in isolation; so too is the scientist who labors alone. The work presented in this dissertation is my contribution to a diverse, multidisciplinary endeavor spearheaded by Dana Pe'er and guided by her instinctive vision. As she has done for others before me, Dana brought me into her lab because she saw something others might not have seen. Insofar as this dissertation is successful, it is vindication of her ability to see potential in nascent scientists and to bring them together in a creative environment. Dana has repeatedly amazed me with insights on the widest spectrum of matters, from abstract theory to grounded pragmatism.

A corrollary of my indebtedness to Dana is my indebtedness to the extraordinary colleagues with whom she brought me into contact. I cannot catalogue what I have learned—both by instruction and by example—from these brilliant individuals, in particular: Bo-Juen Chen, Felix Sanchez-Garcia, Oren Litvin, Smita Krishnaswamy, Michelle Tadmor, and El-ad David Amir. Oren, Bo-Juen, and Felix were especially selfless in taking the time to transform not only what I know but also how I think.

The work presented in this dissertation is the outcome of enthusiastic collaboration with Garry Nolan and members of his lab at Stanford: Erin Simonds, Sean Bendall, and Kara Davis. All of the mass cytometry data that appears in this dissertation was collected in Garry's lab. Many of the ideas that became formalized in the methods presented here were generated by discussions with them. In particular, Erin and I worked closely to shape the studies of leukemia. Erin's contageous tenacity was vital to both the genesis and culmination of the project.

I am also grateful to the members of my thesis committee for providing a constructive variety of ideas, challenges, and encouragement. Carol Prives, Peter Sims, Saeed Tavazoie, and Miriam Merad—thank you.

Finally, I wish to express the deepest gratitude to my family, who supported me in every way while I completed this work.

For every complex problem there is an answer that is clear, simple, and wrong.

H. L. Mencken

Chapter 1

Introduction

1.1 Toward quantitative cell biology

> We do not merely suggest the application of these new technologies to classical cell biological questions, but rather that the fundamental approaches of systems biology, which are unbiased, large-scale, quantitative and multivariate, are integrated into the core of molecular cell biology in the future.
>
> — Liberali & Pelkmans [1]

1.1.1 There are no lonely cells in biology

Cellular differentiation is fundamental to multicellular life. Indeed, phenotypic divergence and functional specialization are so ubiquitous in multicellular populations that they are found in colonies of unicellular organisms [2]. There is no such thing as a multicellular organism without cellular differentiation.

The phenotypic and functional heterogeneity generated by cellular differentiation has been traditionally considered the purview of cell and developmental biology. Perhaps for this reason, progress in this area has lagged behind other branches of the biological sciences. For the decades of the twentieth century during which molecular biology flourished, cellular heterogeneity may have seemed not only tangential to central questions but perhaps counterpro-

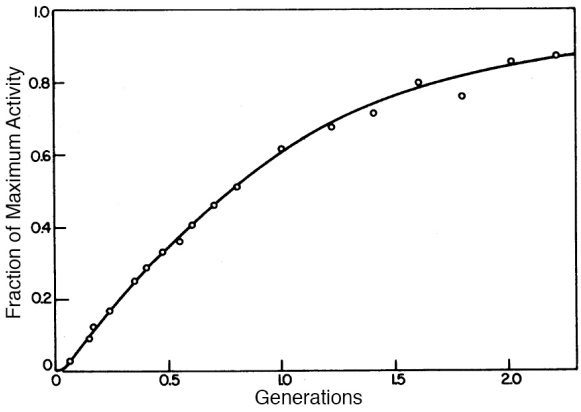
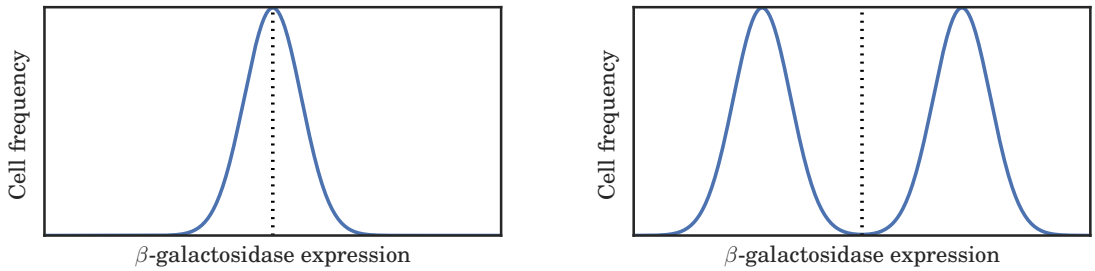


Figure 1.1: β-galactosidase activity measured in bulk aliquots taken at regular intervals from a chemostat with a fixed concentration of inducer. Reproduced from [3].

ductive. Many molecular biology laboratory techniques produce an aggregate measure for an entire population (e.g., culture, tube, plate, specimen), collapsing any cell-to-cell variation into a single point estimate (e.g., mean value). A concrete example is **lysis**, which is used commonly to extract and purify molecular components such as DNA, RNA, or protein. Lysis is used out of pragmatic necessity to obtain sufficient material for further analysis. By pooling the contents of each individual cell, the lysate physically produces an average quantity and renders the cellular heterogeneity of the sample inaccessible. The experimentalist who needs lysis for his or her protocols would prefer that the cellular heterogeneity it destroys is negligible.

On the contrary, cellular heterogeneity is far from negligible and this fact is not new, though it may have been forgotten. Ironically, one of the clearest examples of cellular heterogeneity and the importance of preserving single-cell resolution comes from an early study of the *lac* operon, a fundamental model system of molecular biology [4]. The *lac* operon is a genetic regulatory system of the bacterium *Escherichia coli* that controls production of several enzymes in response to lactose, including β-galactosidase, which enable its use as an energy source. In their 1957 study, Novick & Weiner [3] investigated the kinetics of β-galactosidase production in response to an inducer.[1] Using a method that measures aggregate β-galactosidase activity *in vitro*, enzyme production appears to rise linearly upon induction

[1] Specifically, the nonmetabolizable lactose analogue thiomethyl-β-D-galactoside (TMG).



(a) Intermediate induction (b) All-or-none induction

Figure 1.2: Alternative schematic depictions of distributions of β-galactosidase expression by single *E. coli* cells during induction of the *lac* operon as studied in [3]. The vertical dotted lines indicate the (identical) population means, measured by the aggregate enzyme activity assay. In (b), the mean identifies a cellular state that is vanishingly rare in the distribution.

and taper off as a saturation point is reached (Figure 1.1). One might be tempted to conclude that after induction, each cell increases enzyme production until all cells are producing at full capacity. However, as the authors observed:

> Whenever kinetic experiments are performed using bacterial cultures, the question must be raised whether the results obtained represent the events occurring within the individual cell or some average of a heterogeneous population. (p. 559)

They went on to show that every *E. coli* cell exists in one of two discrete states: **induced** (producing β-galactosidase at the maximum rate) and **uninduced** (producing essentially no β-galactosidase). The appearance of intermediate production rates is nothing but an artifact of averaging over a heterogeneous distribution of cellular states by the aggregate assay. In this case, the aggregate measurement was highly misleading, implying that the majority of cells exist in a state that is in fact vanishingly rare (Figure 1.2).

1.1.2 A century of single-cell analysis

Novick & Weiner used the term **single-cell analysis** in reference to the dilution experiments that allowed them to infer the induction state of individual cells in a parent culture. At the time, those methods and the corresponding term were rarely used in biology (Figure 1.3). Instead, the majority of research in cell biology was conducted by microscopy, which

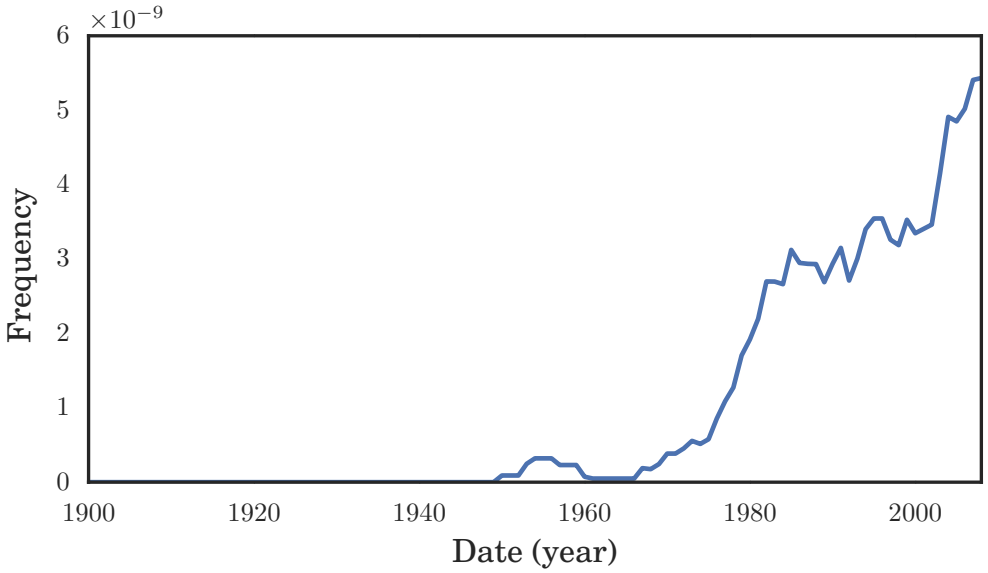


Figure 1.3: Frequency of the 3-gram "single cell analysis" in the corpus of books digitized by Google (`https://books.google.com/ngrams`).

inherently preserves single-cell resolution.

Microscopy

The invention of the microscope was the founding event of cell biology, usually attributed to van Leewenhoek and Robert Hooke in the seventeenth century. In the nineteenth century, biologists such as Theodor Schwann turned the microscope to animal tissues. At the start of the twentieth century, the study of individual cells under the microscope was the state of the art for biological science.

The term **cell type** emerged from the descriptive analyses of early microbiologists, who began to classify cells based on a variety of features observed under the microscope including morphology[2], staining reactivity[3], and functional behaviors.[4] It is worth noting that a great deal of early descriptions of cell types come from studies of the immune system, perhaps because this system is the interface of the multicellular organism with the microbial world. To cite an ordinary example, a 1905 edition of the Journal of the American Medical Association

[2]Often nuclear morphology, e.g., megakaryocytes [5]
[3]e.g., Ehrlich's neutrophil granulocytes
[4]e.g., Metchnikoff's phagocytes

included an article describing the human immune response to pathogen exposure, which began with this summary:

> Cell Types—If leucocytes [*sic.*] consisted of but a single variety of cells, with common functions, the study of leucocytosis would be reduced to a simple process. It would then be sufficient simply to ascertain the degree of leucocytosis by estimating the total number of leucocytes to a cubic millimeter of blood. It is well known, however, that the white blood corpuscles consist of several more or less distinct varieties. ([6])

The author went on to emphasize (and demonstrate anecdotally) that the *composition* of the immune response—the proportions of the various cell types induced by the pathogen—can indicate prognosis in human disease.

As is well known, in the mid-twentieth century biological research was dominated by studies of the molecular components of the cell, using biochemical techniques to reveal their properties. As mentioned previously, these biochemical techniques typically traded single-cell resolution for experimental tractability, pooling together the molecular pieces of disintegrated cells in order to obtain sufficient material. Hence the need later to explicitly designate "single-cell" methods as such. In the broad history of biological sciences in the twentieth century, single-cell analysis began at the forefront and, after receding for several decades, reemerged in the 1970s after several methodological developments led to the era of **immunophenotyping**.

Immunophenotyping

The reemergence of single-cell analysis was spurred by the advent of **flow cytometry** in the 1960s, an ingenious introduction of fluidics into cell biology. Flow cytometry preserves the integrity of single cells by injecting them into a narrow laminar fluid stream, allowing individual measurement as each cell passes serially through a detection chamber. In the late 1960s, the technology was elaborated to incorporate fluorescence optics such that the staining intensity of a fluorescent dye could be quantified in each cell. The method was further extended by Herzenberg *et al.* in the early 1970s to include electromagnetic control of the fluid stream,

permitting the physical sorting of cells based on their measurements—a technology known as fluorescence-activated cell sorting (FACS) [7]. FACS technology was immediately deployed to characterize cellular heterogeneity in the immune system, using functional assays of sorted populations to characterize cell types and obtaining the composition of peripheral blood from the cell counts recorded by the cytometer [8]. Thus, a striking continuity exists from the work of early microscopists to the immunologists of the 1970s, with the taxonomy of hematologic cell types and the cellular composition of blood being constant objectives.

In order to use flow cytometry or FACS to any advantage, it is necessary that a biologically interesting feature can be coupled to fluorescence intensity. For example, staining DNA by bromodeoxyuridine incorporation or Hoechst dye provides a means to couple fluorescence intensity to DNA content [9]. To uncover the diversity of cell types more comprehensively, probes with greater specificity are required—ideally, probes that can label the distinct proteins by which distinct cell types manifest. Precisely this capability was provided by the advent of hybridomas in 1975 [10], which transformed the methodological possibilities for single-cell analysis. Created by fusing antibody-producing B cells with myeloma cells, hybridomas are essentially monoclonal antibody factories, an unlimited source of protein-specific tags that can be chemically conjugated to detectable stains such as fluorophores. The availability of hybridomas effectively turned the immune systems of allospecific animals into laboratory reagents.

It was quickly appreciated that monoclonal antibodies specifically label different subsets of cells whose functional distinctions could be demonstrated by *in vitro* assays following FACS. By 1984, an international protocol for antibody nomenclature was established (using the "cluster of differentiation" [CD] system), formalizing the use of antibodies as laboratory reagents [11].

"Immunophenotyping" refers to the use of antibodies as mediators of protein-specific staining and thereby measuring the phenotypes of individual cells. An immunophenotype is usually reported in terms of the CD nomenclature: for example, a $CD3^+/CD8^+$ cell is "positive for" (i.e., expresses) the antigens to which CD3 and CD8 antibodies bind. From functional studies,

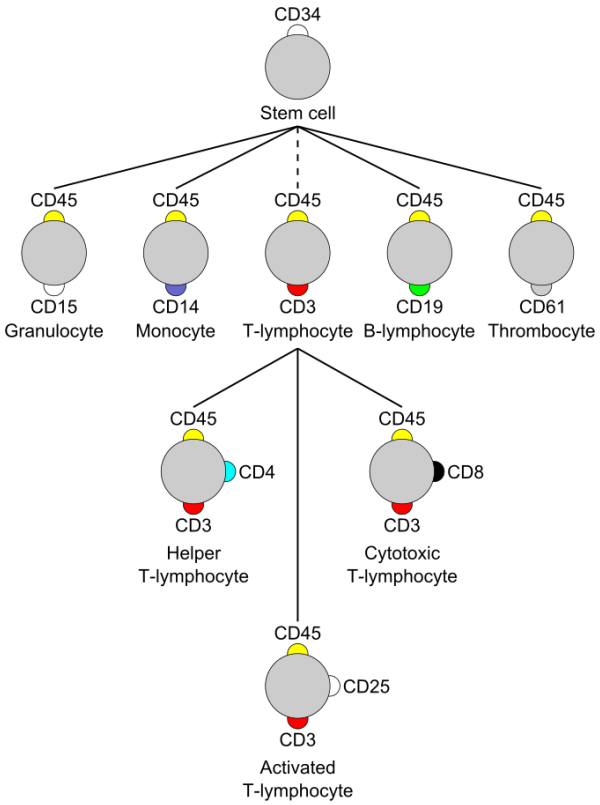


Figure 1.4: Schematic depiction of some immune cell types and CD antigens that distinguish them. Source: `https://commons.wikimedia.org/wiki/File:Cluster_of_differentiation.svg`

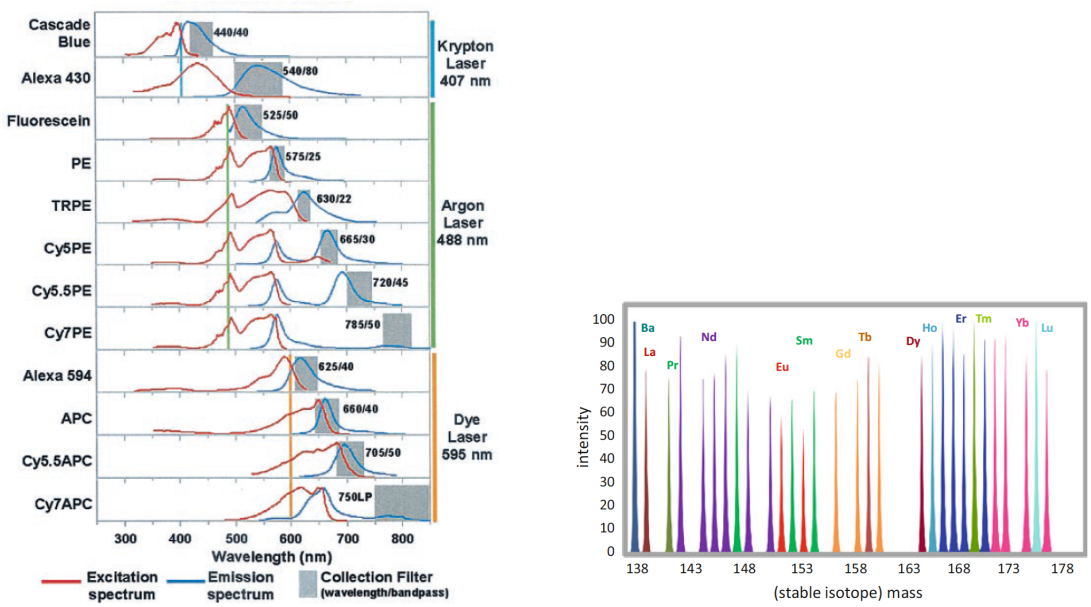
we know that a cell with this particular immunophenotype is a cytotoxic T cell. A very basic schematic of immune cell types, highlighting characteristic CD immunophenotypes, is shown in Figure 1.4. Note that 10 different CD markers are used to distinguish these cell types; some markers (CD4, CD8, CD25) provide fine distinctions within a parent group defined by other, more broadly expressed markers (CD3, CD45).

Fluorescence-based cytometry and CD reagents co-developed for three decades, dramatically expanding knowledge about cellular phenotypes. As the "bank" of designated CD reagents grew, so too did the repertoire of fluorescent tags that could be used to detect them. As different fluorophores have different excitation and emission spectra, the use of multiple lasers (for excitation) and filters (for detection) allowed cytometers to measure multiple CD antigens simultaneously in single cells. By the end of the twentieth century, cytometers were capable of measuring 11 distinct fluorophores simultaneously [12]. The addition of more and more **dimensions** to the technology was driven by the recognition that each dimension reveals another facet of phenotypic complexity, each with the potential to disclose vital information

about the system under study. For example, De Rosa *et al.* [*ibid.*] showed that naive T cells (defined by functional assays) could be weakly enriched by a single marker (CD45RA) and were only completely purified when a combination of 5 markers was used. In general, the number of resolvable subpopulations grows geometrically with the number of measured dimensions.[5] Obtaining fine-grained resolution of the population structure by measuring cells in multiple simultaneous dimensions may be critical for getting an accurate "census" of a multicellular population [13]. As mentioned above in reference to Figure 1.4, measurements of CD4, CD8, and CD25 in addition to CD3 are critical for establishing even a coarse-grained understanding of the T cell composition in a hematologic sample.

The availability of multiple fluorophores and the ability to mix and match them with different antibody probes allows the design of multivariate staining panels. Such panels spurred massive growth in the popularity of flow cytometry and FACS for research as well as clinical disease diagnostics and monitoring [14]. Fluorophores have been the blessing and the curse of antibody-based protein detection: making multivariate detection possible while also setting a restrictive upper limit on the number of proteins that can be measured simultaneously. The characteristic wavelengths at which a given fluorophore absorbs and emits light are not precise values but rather broad distributions over the electromagnetic spectrum (Figure 1.5a). Loading a panel with multiple fluorophores inevitably leads to **spectral overlap**, which can make it difficult or ultimately impossible to determine which fluorophore generated the light detected at a given wavelength: the signal is confounded. The problem can be ameliorated somewhat by **compensation**, a linear transformation of the data that removes the correlation between fluorophores expected from spectral overlap. After proper compensation, the number of simultaneous features that can be measured is nevertheless limited by the number of windows in the electromagnetic spectrum where the signal from one fluorophore is greater than the sum of emissions from the remaining fluorophores on the panel. Flow cytometry is generally limited to ~ 12 fluorophores in practice, with higher numbers being possible but requiring that any given cell express a sparse subset of target proteins and that these sparse

[5] Provided they are not redundant, each additional dimension splits the cells into at least two additional subpopulations.



(a) Fluorescence spectra for 12-color FACS [13] (b) Mass spectra for 30-isotope mass cytometry [15]

Figure 1.5: Fluorescence versus mass spectra. While spectral overlap limits the number of colors that can be used in fluorescence cytometry, mass spectra are generally non-overlapping peaks, eliminating the "crowding" problem associated with fluorescence-based methods.

combinations be known in advance so the panel can be designed accordingly.

Mass cytometry

While 12-color fluorescence cytometry is useful, the limitation begs the question whether other technologies can pick up where fluorescence leaves off, in terms of dimensionality. Indeed, in recent years a technology was introduced that uses transition element isotopes (not normally found in biological samples) as chelated antibody tags in place of fluorophores. Like fluorescence cytometry, **mass cytometry** uses a fluid stream to feed single-cell droplets serially into a detector. Instead of an optical detection chamber, mass cytometry uses a time-of-flight (TOF) mass spectrometer to quantify elemental ions present in each single-cell droplet after ionization by a 5500 Kelvin plasma [16]. This technology takes advantage of the high sensitivity of mass spectroscopy for isotopic analysis, resulting in essentially no overlap between the mass tags. Therefore, unlike the "crowding" of fluorophores in the electromagnetic spectrum, atomic mass tags do not run out of space (Figure 1.5b). Instead, the dimensionality of mass

cytometry is bounded by the range of masses for which enriched stable isotopes are available and that can be detected with comparable sensitivity: this range has extended from 30 to 45 in practice and has been forecast to reach 100 [15, 17].

While mass cytometry has at least threefold greater dimensionality than fluorescence cytometry, it is worth extra emphasis to note that the absence of spectral convolution changes the practical nature of collecting multivariate single-cell data. Designing a polychromatic fluorescence panel can be an arduous task, requiring many stages of optimization, due to the complex excitation and emission spectra of each fluorophore. It is recommended that fluorescence cytometry experiments include both positive and negative controls for each acquired feature. Positive controls measure each fluorophore in isolation in order to quantify its independent spillover into each acquisition channel—such controls provide the empirical "spillover matrix" that is used to compensate the data. Unfortunately, this is already an imperfect control because the spillover coefficient is not constant across the dynamic range of the instrument [18]. To account for imperfections in compensation and further spectral contaminations caused by dye-dye interactions and cell-dye interactions, negative controls should also be collected, in which every marker *except one* is included. These fluorescence-minus-one (FMO) controls should be generated for each marker—thus, for a D-dimensional fluorescence panel, $2 \times D$ controls should be collected every time data are acquired. The amount of work required to optimize a particular panel configuration is a significant barrier to panel redesign. On top of the experimental burden, there is no established way to use the FMO controls quantitatively—they are typically used to draw manual "positive/negative" gates as part of a fluorescence cytometry data analysis workflow that is almost always qualitative.

On the other hand, mass cytometry not only provides increased dimensionality but also produces data that are more appropriate for quantitative analysis. There is no equivalent of autofluorescence and interaction between mass tags is not a significant factor in panel design. These facts eliminate the need for marker-specific controls. Further, mass cytometry allows the inclusion of pure-isotope beads that can be used to standardize measurements acquired at different times [19]. Finally, the dimensionality allows for sophisticated barcoding systems

that enable multiplexed acquisition, which further increases quantitative comparability and reduces reagent usage [20]. Thus, mass cytometry panel (re-)design is more straightforward and flexible compared to fluorescence cytometry. These features encourage high-throughput production of single-cell data that are quantitative and multivariate, increasing both the need and the potential yield of computational analysis.

Over the last century, scientific descriptions of cellular phenotype have evolved from qualitative descriptions to quantitative high-dimensional measurements. As the data type changes, so too do the analytical possibilities for answering perennial questions relating to the types and composition of multicellular populations.

1.2 Learning cell types from experiment

1.2.1 Immunophenotyping normal hematopoiesis

It had been known before the development of immunophenotyping techniques that human bone marrow contains colony-forming cells capable of giving rise to progeny of differentiated lineages [21, 22]. In the 1980s, it was discovered that human bone marrow progenitor cells express the membrane antigen CD34 [23]. Using 3-color FACS, it was discovered that immature bone marrow cells begin to express markers of lineage commitment only after $CD34^+$ cells begin expressing another marker, CD38, implying that progenitors with a $CD34^+/CD38^-$ phenotype might represent the apex of hematopoietic development [24]. Indeed, using the sorting capability of FACS, it was shown that $CD34^+/CD38^-$ bone marrow cells are more efficient at forming colonies *in vitro* and display morphological features consistent with an undifferentiated state.

To study cell types by FACS, cells are sorted by setting rules that define ranges of values for each dimension, dividing the population into **gates** jointly satisfying each rule. For example, Figure 1.6 shows the gates used in [24] to enrich for hematopoietic stem cells (HSCs). All FACS experiments employ a "guess and check" design: Gates must be defined at the time of sorting (guess) and functional validation is performed after sorting (check). Thus,

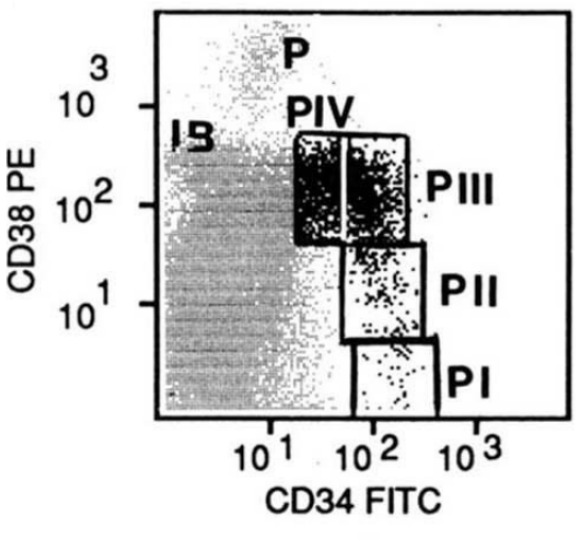


Figure 1.6: An example of FACS gates used to identify hematopoietic stem cells in human bone marrow [24]. The box labeled "PI" represents a region of the bivariate space where cells are $CD34^+/CD38^-$. Cells falling into each gate were separated by FACS and submitted to functional analysis.

learning about cell types from FACS experiments requires that the system under study produce a predictable array of phenotypes, allowing the **prospective** definition of gates based on prior knowledge. Indeed, normal bone marrow displays remarkably predictable phenotypes, allowing prospectively defined gates to isolate the same cell type universally across individuals. Progress in learning about cell types from FACS experiments has relied heavily on this predictability.

1.2.2 Immunophenotyping malignant hematopoiesis

As was the case for healthy human bone marrow, it had been known before the introduction of immunophenotyping techniques that leukemic bone marrow also contains colony-forming cells capable of generating differentiated progeny. Unlike the colony-forming cells of normal marrow, however, leukemic marrow tends to form cells of a specific lineage and this can actually be used to *define* the various types of leukemia. For example, acute myeloid leukemia (AML) is marked by an overabundance of undifferentiated leukocytes (also called "blasts") that resemble myeloid progenitors and can produce differentiated myeloid progeny, with the extent and type (monocytic, granulocytic) of differentiation depending on the individual case.

AML as a deregulated tissue

It has been hypothesized since at least the 1970s that AML arises from a disruption in normal myelopoiesis—a "differentiation block" causing cells which would be otherwise destined for myeloid fates to accumulate in an undifferentiated state. Indeed, there is experimental evidence that at least forms of AML are caused by disruption of the mechanism regulating the passage of cells from an immature to a mature cell type. For example, some AML blasts can be induced to differentiate into macrophages and granulocytes (both *in vitro* and *in vivo*) by cytokines called colony-stimulating factors (CSF); following exposure to CSF these leukemic blasts follow the same series of morphological events as observed in normal myeloid maturation, including terminal exit from the cell cycle [25, 26]. Unlike their normal counterparts, these leukemic blasts proliferate rather than die in the absence of CSF, linking the differentiation block to enhanced proliferation.

The demonstration that all-*trans*-retinoic acid (ATRA) induces terminal differentiation of acute promyelocytic leukemia (APL; a subtype of AML) and the subsequent clinical transformation of APL from the most fatal to the most curable form of acute leukemia must be regarded as one of the greatest successes in the history of translational medicine [27]. It proves that at least one form of AML is caused by a differentiation block that, when lifted, allows cells to stop proliferating and resume normal development. Unfortunately, the clinical success of ATRA for APL relies on a specific fusion protein involving the ATRA receptor, which occurs in the overwhelming majority (95%) of APL cases [28]. Thus, the clinical benefit of ATRA is not applicable to other forms of AML, unfortunately.

A "molecular lever" to control cellular differentiation in other forms of AML has been much more elusive. However, genetic and cancer-genomic studies have revealed a common molecular pathway that provides mechanistic insight to normal myelopoiesis and AML pathogenesis. Specifically, the family of transcription factors CCAAT/enchancer binding protein (C/EBP) have been implicated as major regulators of lineage commitment and differentiation in normal hematopoiesis [29]. It has been shown in genetically engineered mice that disruption of $C/EBP\alpha$ blocks the transition from common myeloid to granulocyte/monocyte

progenitor, resulting in accumulation of myeloid blasts in the bone marrow [30]. $C/EBP\alpha$ is also clearly a target of the genetic alterations that drive AML: the gene itself is mutated in a significant number of cytogenetically-normal AML genomes, and other common genetic abnormalities of AML—such as the AML-ETO1 fusion [t(8;21)] and $FLT3$-internal tandem duplication (ITD) mutation—result in downregulated expression of $C/EBP\alpha$ transcript or decreased post-translational activation of C/EBPα protein [31]. Finally, it has been shown that another member of this gene family, $C/EBP\beta$, is strongly activated by ATRA in APL cells and that this activation is required for ATRA-mediated differentiation to occur [32]. A hallmark of a true cancer driver is that different cases utilize different mechanisms to realize the same outcome at the pathway level. It is clear that AML uses a variety of proximate causes to prevent myeloid differentiation via inhibition of C/EBP transcription factor activity [28].

Leukemic "cell types": The CSC model

Perhaps because AML is a deregulated myelopoietic process, it retains much of the hierarchical structure of normal hematopoiesis despite its malignant behavior. It has been recognized for decades that most cases of AML are composed of phenotypically distinct subpopulations that vary in their potential for self-renewal, proliferation, and differentiation [33–35]. Given the similarity exhibited by AML to the structural organization of cellular potential in normal bone marrow, it is logical to hypothesize that AML immunophenotypes reflect the functional features of analogous subpopulations in normal bone marrow.

It was in the context of AML that the cancer stem cell (CSC) model was developed. In 1997, it was first shown *in vivo* that $CD34^+/CD38^-$ human bone marrow cells are dramatically enriched for HSCs, which are capable of engrafting and reestablishing the immune systems of nonobese diabetic/severe combined immunodeficient (NOD/SCID) mice [36]. That same year, it was reported by the same group [37] that $CD34^+/CD38^-$ leukemic marrow cells are similarly enriched for cells that engraft and propagate AML in NOD/SCID mice. This demonstration that only a subset of leukemic cells are capable of "seeding" new malignancies

implied that these cells are leukemic stem cells (LSCs). Though it was initially reported that *only* $CD34^+/CD38^-$ cells are LSCs, subsequent studies using more sensitive assays identified $CD38^+$ [38] and $CD34^-$ [39] LSCs, overturning the initially proposed LSC model.

While the CSC model has been questioned wholesale in some cancer types, it remains well supported that most cases of AML do exhibit a functionally differentiated hierarchy—i.e., only a subset of blasts can propagate the malignancy and these cells give rise to progeny that both possess and lack this capability [40]. Furthermore, it has been shown that estimates of LSC frequency are correlated with chemoresistance and poor overall survival [41, 42], corroborating not only the existince of LSCs but also their clinical significance. However, a consistent immunophenotype that can prospectively isolate this subset of cells across patients has not been forthcoming [43].

It should perhaps be expected that LSCs do not present the same immunophenotype across patients. After all, most CD antibodies target membrane-bound proteins that are correlated with—but not causal for—cellular function. Many common CSC markers, including CD34, are primarily involved in cell adhesion[6] rather than functions more closely related to stem cell identity [44]. CD34 is certainly not necessary for HSC function in general, given that murine HSCs are specifically negative for mCD34 [45]. That these surface markers may be used as proxies for cellular function in normal hematopoiesis—universally, across individuals—is a testament to the exquisite coordination of protein expression in healthy tissues. Such coordination should not be expected in malignancy, driven as it is by deregulation.

Thus, passive monitoring of surface antigens is not sufficient to identify the important aspects of intratumor heterogeneity. Reporters of functional potential should be preferred, as each unique malignancy may locate functional potential in different phenotypic compartments.

The question then arises what to use as a functional reporter. One option is the xenotransplantation assay, which requires sorting every tumor into prespecified gates via FACS and estimating LSC frequency via limiting dilution assays, which requires injecting tumor cells into several mice *per gate*. In fact, this approach was taken by one group, whose results

[6] Which raises legitimate concerns about biases imparted by the xenotransplantation assay, but which is not the main issue I wish to stress here.

will be discussed at length elsewhere [43]. Another option is to measure the quantitative activation of proteins that participate in cellular signaling cascades, using antibodies that specifically target phosphorylated epitopes [46]. It stands to reason that the **signaling profile** of a cell—the activation status of its various signaling pathways—is intimately connected with its functional potential [47].

1.2.3 Immunophenotyping in higher dimensions

It is worth considering the importance of dimensionality for learning about cell types, especially in less characterized tissues. The human proteome generates cellular diversity through the regulated expression and modification of thousands of protein species across tissues and developmental stages [48–50]. As mentioned previously, the reliability of low-dimensional phenotypes such as $CD34^+/CD38^-$ for defining cell states is a result of the exquisite regulation of protein expression that couples these surface markers to functional state in healthy tissues. Even cells defined by this bivariate phenotype are a heterogeneous population: Only a fraction of $CD34^+/CD38^-$ cells are HSCs (for example, [24] report that 25% of these cells form primitive colonies *in vitro*). There are probably other, currently unknown, phenotypic features (possibly not surface antigens), that if added to form a higher-dimensional phenotype would define HSCs with greater (possibly 100%) accuracy. Similarly, adding dimensionality to the definition of LSC enhances the possibility of producing a definition that can be applied across individuals.

Critically, the benefit of increasing dimensionality is not simply derived from increasing the likelihood of finding a single essential marker. On the contrary, the benefit comes from representing cell states as high-dimensional patterns of protein expression and activation, which are potentially robust to variation in any particular marker. In other words, it is very rare that a discrete biological function can be attributed to an individual molecule; rather, biological functions arise from interactions among multicomponent molecular modules whose activities would manifest as distinct patterns in high-dimensional measurements of the cell [51].

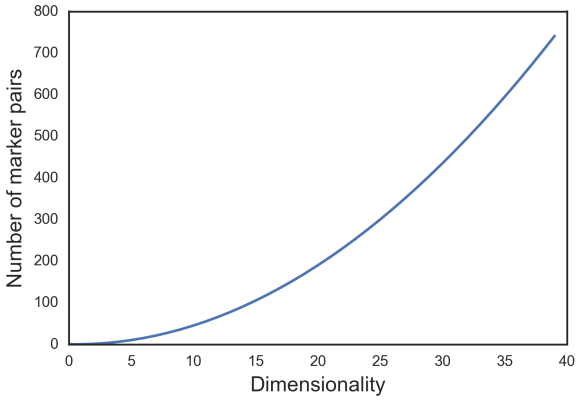


Figure 1.7: The number of possible marker combinations as a function of measurement dimensionality.

1.3 Learning cell states from data

Computational biology was born of necessity not only to organize the very high-dimensional measurements generated by new genome wide technologies at the start of the twenty-first century, but to actually extract knowledge from these overabundant data. Because DNA sequence is (typically) unchanged between individual cells within a sample, it is safe to regard each cell as an interchangeable source of DNA molecules, a critical mass of which is required to actually generate genomic sequence reads. Thus the earliest genomic data were generated from lysates of multicellular populations. As genomic technologies moved to quantify gene expression through mRNA abundance, the use of multicellular lysates became problematic. Unlike gene sequence, gene expression may vary drastically between cells within a sample. Yet, bulk measurement of cells through whole-sample lysis collapses this intercellular heterogeneity into a single (average) measurement (similarly, for example, to what is depicted in Figure 1.2). Though these bulk lysates are not ideal for understanding gene expression in heterogeneous populations, they were technologically necessary for many years. Only very recently have technologies emerged that provide genome wide quantification of mRNA in individual cells at a scale that can properly reveal the full distribution of gene expression in a heterogeneous population [52].

Simultaneously, experiments focused on preserving and exposing cellular heterogeneity

have gradually increased their dimensionality and only recently entered the quantitative domain, as outlined in the previous section. In early FACS experiments, population heterogeneity was characterized manually by the experimentalist through exploration of bivariate scatter plots (like the one shown in Figure 1.6). During the growth of dimensionality that occurred over the last 30 years, manual exploration of FACS data has become intractable (Figure 1.7). For 4-dimensional data, the experimentalist has $\binom{4}{2} = 6$ plots to examine. For 12-dimensional data, there are $\binom{12}{2} = 66$ marker pairs to study. For 30-dimensional data, the experimentalist now has $\binom{30}{2} = 435$ scatter plots to study, per sample. Even a very motivated individual will be unable to extract the most from the data because patterns in 30-dimensional space simply cannot be discovered by sequential examination of 2-dimensional subspaces. Thus a computational approach is needed, not only to aid in handling the overabundant data, but to extract knowledge that could otherwise never be discovered.

This section introduces concepts and methodological groundwork that will be useful for performing computational analysis of single-cell data. In particular, we explore the related concepts of dimensionality reduction and clustering, which are both relevant to the task of modeling the underlying cell states that produce the variety of high-dimensional cellular phenotypes measured by new technologies such as mass cytometry.

1.3.1 Terminology

Before proceeding, some terminological subtleties are worth addressing. These distinctions bring clarity to the discussion and may provide insight for the methods to be presented in subsequent sections.

We have already encountered the concept of **cellular phenotype** multiple times. A cellular phenotype is simply any observable trait of a cell. "*Immuno*phenotype" refers to the fact that a trait has been observed by antibody labeling.[7] A cellular phenotype can be an individual trait or a concatenation of traits: Any of {$CD45^+$, *granular*, $CD15^+$, $CD45^+/CD15^+$} is a cellular phenotype. For D-dimensional cytometric data, the vector of D staining intensities

[7]The "immuno-" prefix is often dropped because the use of antibody labeling is so common and because it is often clear from context, especially when the CD terminology is used.

generated by each cell may be regarded as its phenotype: After all, this is precisely what is known about the observable traits of that cell.

Cell state refers to something more abstract. The term has a wider scope than "phenotype" and has functional connotations as well. For example, *mitosis* is a cell state. If cell cycle markers are measured, a number of phenotypes would reflect mitotic cells. "Cell state" may also refer to functional potential, as *multipotent* describes the ability to produce daughter cells that reach different fates. Because cell states arise in different contexts and reflect the functional outcome of complex molecular systems, one can expect that a diverse but restricted set of phenotypes correspond to a cell state. From a modeling perspective, cell states are *latent variables*: unmeasured (or unmeasurable) attributes of the cell that can be inferred from patterns they generate in the observable phenotypes. For an extreme example, Novick & Weiner inferred that "lactose induction" (a cell state) is an all-or-none phenomenon, based on the pattern of β-galactosidase activity (a phenotype).

As mentioned previously (§1.1.2), the term **cell type** emerged in the qualitative descriptions of early microscopists. The term implies functional specialization associated with multicellular organization: examples include *skin cell*, *muscle cell*, *nerve cell*. Upon closer examination the term becomes slippery, as it depends on a desired resolution (e.g., *nerve cell* names a wide variety of neural cell types) and implies a discretization of developmental processes that may in fact be continuous [53]. Taking these provisos into account, "cell type" implies significant phenotypic and functional stability over time, achieved perhaps through epigenetic regulation [54]. The stability of a cell type might exceed the time scale of a cell state: For example, a T cell can exist in different states (*mitotic, activated*) while maintaining its "identity" as a T cell. The stability of a cell type may also preclude the accessibility of certain states: T cells cannot produce insulin. One can expect that a diverse but restricted set of phenotypes *and* states correspond to a cell type.

Every cell has some (set of) phenotype(s) and exists in some (set of) state(s). Whether every cell can be regarded as an instance of a cell type is perhaps an open question, especially in the context of disease.

A population is **heterogeneous** when it contains cells in one of several states (and/or of multiple types). For each state, the cells in that state form a phenotypically coherent **subpopulation**—i.e., the cells in that subpopulation are phenotypically more similar to each other than they are to cells of other subpopulations. In some contexts, it may be convenient to refer to a **subpopulation phenotype**. This abuse of terminology might be taken as shorthand for "the distinctive traits shared by most cells in a subpopulation." It may also be used in reference to a D-dimensional vector computed on the cells of the subpopulation and meant to represent their collective attributes, such as the average expression in each dimension.

1.3.2 Cluster analysis

Since cytometry became a multidimensional measurement technology, it has been apparent that subpopulations could be identified as **clusters** of cells with similar phenotypes. For example, in 1991, Terstappen *et al.* wrote:

> Multidimensional flow cytometry permits quantitative identification of distinct cell populations in the heterogeneous blood and BM [bone marrow] aspirates. In a multidimensional space created by simultaneous measurement of independent parameters, cell populations with dissimilar properties emerge at different positions as clusters of cells.[8]

Though they did not do so, this intuition can be formalized and the task of identifying subpopulations can be treated algorithmically.

First we consider a highly idealized scenario. First, suppose there is a "complete" measurement space that includes every feature that can possibly contribute to cell state: the abundance and activation of every protein species, the spatial location of every molecule, *et cetera*. Let there be $\widetilde{D}$ such features. Suppose there is a technology that measures each of these $\widetilde{D}$ features in each cell sampled from a multicellular population. Suppose that this

[8] Note that these authors use the term "population" in the sense of "subpopulation" as defined in §1.3.1

multicellular population is heterogeneous, comprising exactly K cell states each present in some fixed proportion $\phi_k, \sum_{k=1}^{K} \phi_k = 1$.

When N cells are measured, the output of the technology is a $N \times \widetilde{D}$ matrix (equivalently, a set of N $\widetilde{D}$-dimensional vectors). If the biological system is noiseless, then each cell state is present as $\lfloor \phi_k N \rceil$ identical cells. If the measurement technology is also noiseless, then the $N \times \widetilde{D}$ matrix can be equivalently represented by K sets, each containing $\lfloor \phi_k N \rceil$ identical $\widetilde{D}$-dimensional vectors. In this case, the sets of identical vectors trivially identify the K cell states.

Of course, reality is barely a crude approximation to this hypothetical. Despite the substantial advances outlined above (§1.1.2), we have not reached the asymptotic "complete" measurement space. In an incomplete measurement space, there is always the possibility that an unmeasured variable defines distinct cell states that will be irresolvable in the data. On the other hand, the number of potentially resolvable cell states increases with each added dimension, suggesting that it should always be valuable to seek new cell states in data of increasing dimensionality. Further, due to the modular organization of the cell's physical and regulatory systems [55, 56], it is reasonable to expect many dimensions of the complete measurement space to be redundant and therefore that the *intrinsic* dimensionality of the cell is less than $\widetilde{D}$ (see §1.3.3).

The hypothetical postulates regarding noise are even further from reality. Stochasticity is inherent in molecular systems [57], an inevitable result of fluctuations in transcription and translation that cells have evolved to buffer and exploit in order to control transitions between states [58, 59]. Thus, cell states should not be expected to generate sets of identical vectors in real data, even in the absence of measurement noise. Instead, cell states will generate **densities**: regions of the measurement space containing a high frequency of similar but nonidentical vectors. A population composed of K states would generate K different (but possibly overlapping) densities.

Thus, one may take a statistical view and model the heterogeneous population as a mixture of densities (also known as a mixture model). This approach has been tried by a number of

groups for flow cytometry data and will be further discussed in the next chapter (§2.1.1). Given the foregoing exposition, one might suppose that a mixture of densities would work quite well to model heterogeneous populations; however, these models suffer from overly rigid assumptions and computational instability in high dimensions.

A simpler method, which is closely related to mixture models and provides a paradigmatic example of cluster analysis, is k-means [60]. It is worthwhile to examine k-means in some detail because it exemplifies how one might formalize the intuitive concept of a cluster. Essentially, a cluster is a subset of data points that are mutually more similar to each other than they are to the other data points. A clustering solution assigns all data points to exactly one cluster such that the overall intra-cluster similarity is maximized. k-means defines both an objective function that formalizes this idea and an algorithm for optimizing it. The objective function, called the within-cluster sum of squares (WCSS), is given by:

$$\arg\min_{\mathcal{C}} \sum_{k=1}^{K} \sum_{\mathbf{x}\in\mathcal{C}_k} \|\mathbf{x} - \mu_k\|^2 \tag{1.1}$$

where $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ is a partition of the data $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ into K disjoint clusters and μ_k is the mean of the data in cluster k, $|\mathcal{C}_k|^{-1} \sum_{\mathbf{x}\in\mathcal{C}_k} \mathbf{x}$. One can see that (1.1) formally encodes the notion that a good clustering solution should maximize intra-cluster similarity: It does so by minimizing the squared Euclidean distance between the cluster's centroid and every point it contains. As discussed in the next chapter (§2.1.1), the form of (1.1) imposes a particular structure on the clusters that may not be desirable in general.

A k-means clustering solution is an arrangement of $\mathcal{C}$ that minimizes (1.1). The algorithm for minimizing (1.1) is a special case of the expectation-maximization algorithm and consists in iterative updates of $\mathcal{C}$ and μ_k. The procedure is guaranteed to converge to a *local* minimum; however, even in simple cases there are numerous local minima that yield poor clustering results [61, 62]. The problem of finding the global minimum of (1.1) is NP-hard [63]—i.e., it is computationally intractable[9] and therefore heuristics (such as the expectation-maximization

[9] More precisely, finding the global minimum of (1.1) is in the class of problems believed not to be solvable in polynomial time ($\mathcal{NP}$) and therefore requires computational resources that scale exponentially with the size of the data. This prohibitive scaling renders such problems computationally intractable.

algorithm) must be used in practice to obtain approximate[10] solutions. This difficulty is not specific to k-means; in general, formulations of cluster analysis are NP-complete at best. Thus, finding "good" clustering solutions necessarily involves designing good heuristics.

Another paradigmatic clustering technique that has been widely used (especially in bioinformatics [64]) is hierarchical linkage. Unlike k-means, the approach is entirely heuristic and provides no global objective function. Briefly, the algorithm begins with every data point in its own cluster and merges the two most similar points into a new cluster; merging is repeated hierarchically until the entire data set is merged into a single cluster. Hierarchical linkage does not explicitly produce a clustering solution but rather a sequence of binary merge decisions—i.e., a tree. The tree must be post-processed by some heuristic or other (often "by eye") in order to extract clusters.

Innumerable variants of these paradigms have been developed over the years. One reason for this proliferation is probably that—by virtue of the computational complexity—the need for heuristics introduces domain dependencies. Though no machine learning technique ought to be used "out of the box" without testing assumptions and integrating domain knowledge, this proviso is perhaps nowhere more important than in cluster analysis. A clustering method designed precisely to cope with the various demands of high-throughput high-dimensional single-cell data is the subject of Chapter 2.

1.3.3 Manifold learning

The irony of high-dimensional data analysis is that while we want to measure as many simultaneous variables as possible, we ultimately want to discard as many of these as possible. More precisely, we seek to reduce them in an information-preserving manner. There are two reasons: one motivated by the data, the other by the analysis.

From an analytical perspective, each dimension is very costly in terms of the ability to evaluate structure in the data. There are several ways to demonstrate mathematically the so-called *curse of dimensionality* (see, for example, [65] and references therein). The common theme is that the volume of the measurement space increases exponentially with the

[10] or exact but not guaranteed to be globally optimal

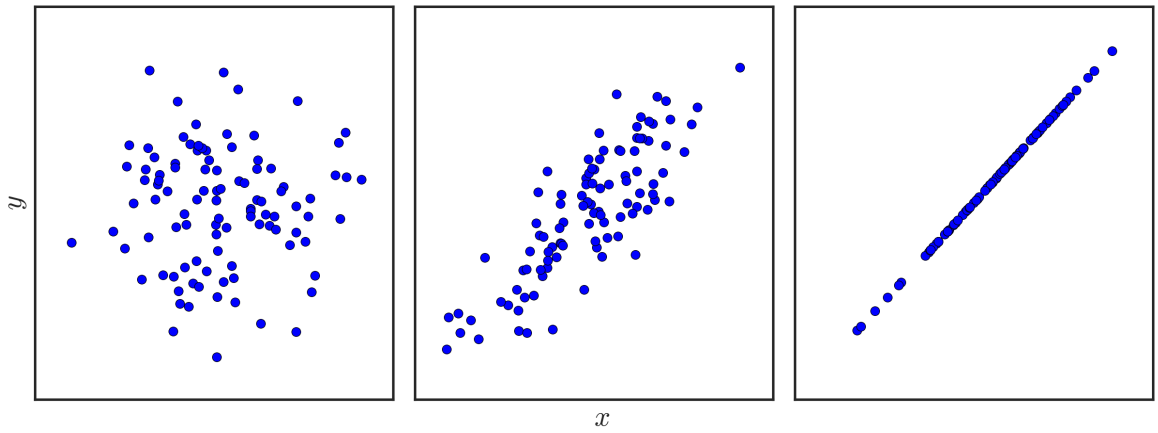


Figure 1.8: The joint distribution of two Gaussian variables x and y. In the left panel, the variables are completely independent. In the middle panel, the variables are moderately correlated. In the right panel, the variables are perfectly correlated. As the correlation brings the data onto a lower-dimensional manifold, the sparsity of the measurement space also increases.

dimensionality; therefore one would require an exponential increase in the number of data points to maintain a constant sampling rate.[11] In other words, high-dimensional spaces are unavoidably sparse and sparsity is bad for inference.

Compressing linear dependencies

Fortunately, when there are dependencies between the measured variables, the **intrinsic dimensionality** is less than the measured dimensionality. A simple example involving two Gaussian variables, x and y, is shown in Figure 1.8. When the two variables are completely independent, x provides no information about y and therefore both are required to determine a point's location: the data are truly two-dimensional. When the two Gaussian variables are perfectly correlated (i.e., completely dependent), x provides *all* information about y (and vice versa). In this case, the two-variable system has an intrinsic dimensionality of one, apparent from the restriction of the data to a linear subspace of the (x, y) coordinate system—i.e., a line. In other words, though the data are embedded in the two-dimensional space $\mathbb{R}^2$, they

[11] For example, suppose that 1,000 data points are sufficient to estimate a distribution in a square (i.e., two-dimensional hypercube) with constant edge length r. To achieve the same sampling rate in a 20-dimensional hypercube with the same edge length requires $\sim 10^8$ data points.

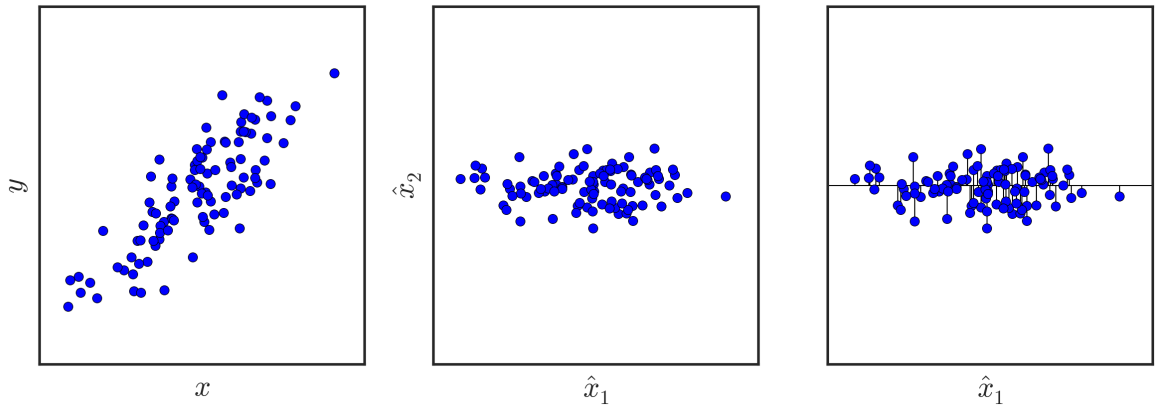


Figure 1.9: PCA extracts uncorrelated latent variables ($\hat{x}_1$ and $\hat{x}_2$, middle panel) from correlated variables (x and y, left panel) by applying a rotation. The first principal component, $\hat{x}_1$, accounts for 90% of the covariance of x and y.

can be re-parameterized by $\mathbb{R}$, the real line.

In the unrealistic case of perfect correlation, the dimensionality can be reduced simply by dropping one of the variables. In reality, variables display imperfect dependencies, reflecting either noise or true differences in the variables or both. For example, in the middle panel of Figure 1.8, neither x nor y would be the best substitute for the other. Instead, the redundancy of these variables should be handled by combining them. This is the approach taken by latent variable models such as principal component analysis (PCA). The simplest dimension reduction method, PCA assumes that the observed variables are generated by a linear transformation of (a smaller number of) latent variables and seeks to recover these latent variables by reverse engineering the linear transformation. In the case of the moderately correlated Gaussian variables of Figure 1.8, this amounts to a simple rotation such that the first latent variable ("component") accounts for 90% of the variance in the data. Projecting the data into the linear subspace spanned by this single component effectively reduces the dimension to one (Figure 1.9).

A well-known weakness of PCA is that it can compress only linear dependencies. As shown in Figure 1.10, rotation and rescaling do nothing to reduce the dimensionality of x and y when their dependence is nonlinear. Both latent variables $\hat{x}_1$ and $\hat{x}_2$ are needed: Projecting the data onto the first component $\hat{x}_1$ destroys the structure of the data, despite the fact that

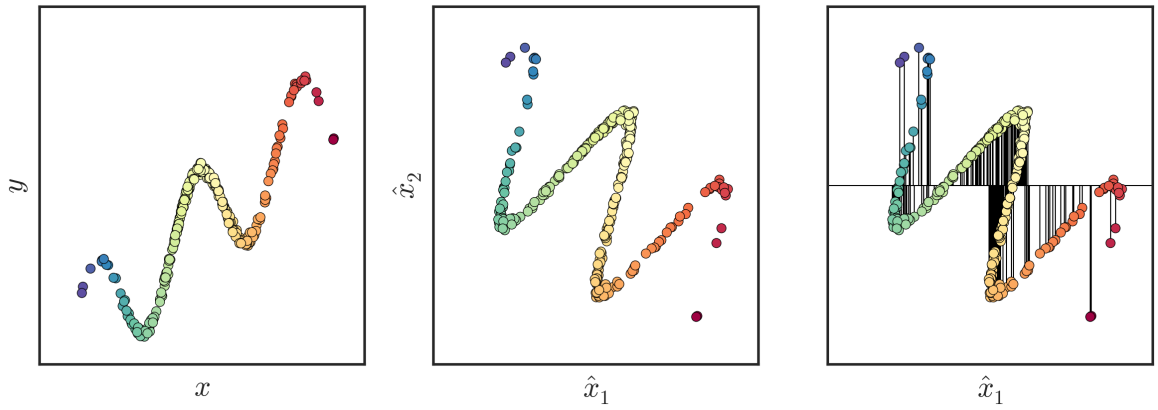


Figure 1.10: PCA is unable to reduce the nonlinear dependency between x and y. Points are colored by their order on the line. Rotation and rescaling (middle panel) and projection onto the first component (right panel) corrupt the structure of the data.

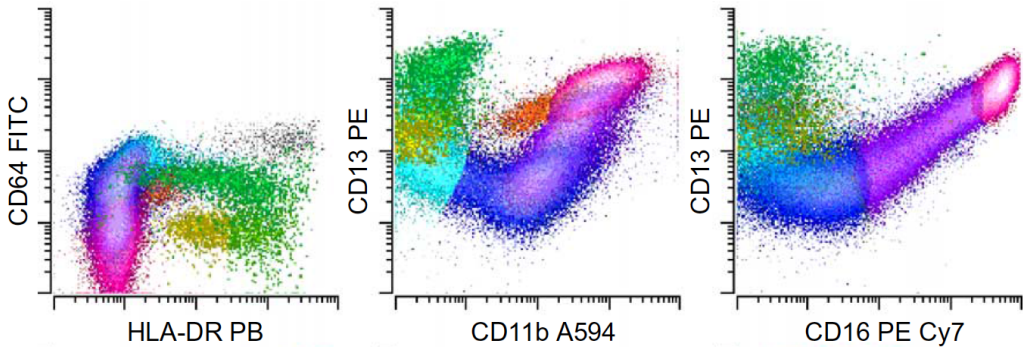


Figure 1.11: Nonlinear relationships between proteins in normal human bone marrow. Cells are colored by manual gating according to known stages of neutrophilic differentiation. Yellow and green subpopulations represent the earliest stages of maturation. From [66].

the intrinsic dimensionality of this structure is clearly one.

Compressing nonlinear dependencies

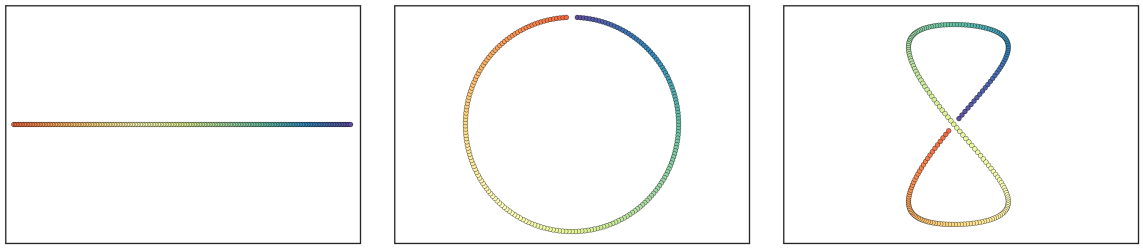
It should come as no surprise that the relationships between variables in single-cell data are generally not linear (Figure 1.11). The catalog of known biological mechanisms is full of interactions, cascades, feedbacks, and thresholds—all of which generate nonlinear relationships between the cell's molecular components. Previous work in our group has shown that linear dimensionality reduction techniques such as PCA fail to capture phenotypic heterogeneity while nonlinear methods such as t-Distributed Stochastic Neighbor Embedding (t-SNE) [67]

are very effective for cytometry data [68].

There are a number of reasons that t-SNE is successful for visualizing single-cell data. Perhaps the most important reason, and the one most relevant for this dissertation, is that t-SNE is a **distance-preserving** method. This class of method is inspired by concepts from mathematical topology. A central concept in topology is that one can make a distinction between the structure of an object and its representation in space. An object can be stretched, twisted, or deformed without fundamentally altering its structure—what matters is the *connectivity* of points on that object. Technically, this topological object is called a **manifold**, which is regarded in distinction to its **spatial embedding**. Figure 1.12 illustrates the distinction between a manifold and its spatial embedding. Intuitively, an important feature of manifolds is that their structure is *locally* Euclidean. For example, the surface of the Earth is a two-dimensional manifold (a sphere embedded in physical space $\mathbb{R}^3$) that is "flat" on small scales [65]. The **geodesic** distance[12] between two points in the same city is roughly Euclidean in $\mathbb{R}^3$, which is not true for two points on different continents. In Figure 1.12, the Euclidean distance between any pair of adjacent (i.e., similarly colored) points is virtually unchanged across the three spatial embeddings. Conversely, the Euclidean distance between non-adjacent points can change quite drastically between different spatial embeddings.

With these insights in mind, we can propose that the manifold is the true object of interest and that the spatial embedding is a necessary but ultimately dispensable way to learn its structure. Distance-preserving methods begin with the proposition that the data are generated on a manifold but are embedded in the measurement space $\mathbb{R}^D$ by an unknown and generally nonlinear mapping. From this, we can use the corollary that data points which are adjacent in the embedding space are probably generated by adjacent points on the manifold. In other words, "any manifold can be fully described by pairwise distances" [65, p. 69]. Given the set of points in $\mathbb{R}^D$ and their pairwise distances, the goal of **manifold learning** is to reconstruct the manifold by finding a model that is simpler than the data (e.g., of lower dimensionality) that preserves pairwise distances.

[12] Literally, the shortest arc between two points on the Earth's surface, this term may also refer more generally to the shortest path on any manifold.



(a) (b) (c)

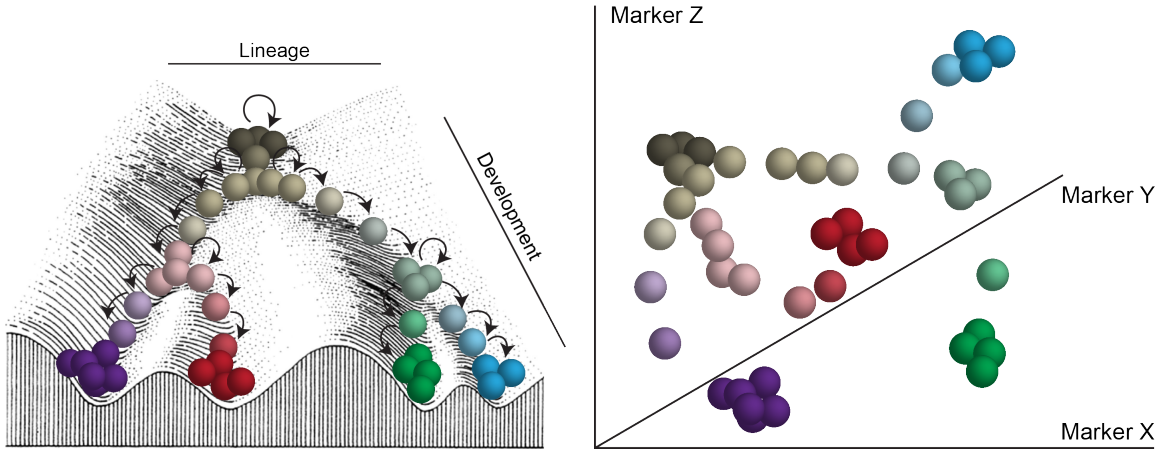
Figure 1.12: Different spatial embeddings of the same topological manifold. The real line $\mathbb{R}^1$ is (trivially) a 1-dimensional topological manifold. Colors represent the intrinsic ordering of points on the manifold. (a) A subset of the real line is shown embedded in a metric space of the same dimensionality (the vertical axis contains no information). (b) The same manifold is shown embedded in $\mathbb{R}^2$ as a disconnected circle. (c) The same manifold is shown embedded in $\mathbb{R}^2$ as a disconnected figure eight. Both (b) and (c), but not (a), are nonlinear embeddings of the manifold in $\mathbb{R}^2$.

1.3.4 The phenotypic manifold

Multicellular populations are generated by the articulated processes of division and differentiation. Because daughter cells will generally be very similar to their mother cells, the phenotypic heterogeneity of the population is established by incremental divergences. These incremental divergences create restricted paths through the space of possible phenotypes. As cells are constrained to follow these restricted paths, cellular phenotypes lie on a subspace of the full high-dimensional phenotypic space. We take the view that this subspace can be regarded as a topological object, which we call the **phenotypic manifold** (Figure 1.13). As such, it may (and does) have a nonlinear embedding in the measurement space. The task of learning cell states can be translated to the task of learning about the phenotypic manifold. Modeling the phenotypic manifold is a conceptually motivated approach, rooted in an understanding of the generative mechanism of cellular populations. As the subsequent chapters aim to show, it is also an analytically powerful approach, yielding results that are superior to comparable methods and leading to new insights about cell states in complex tissues.

1.4 Dissertation outline

This dissertation is about two things:



(a) (b)

Figure 1.13: Schematic depiction of a "phenotypic manifold." (a) Phenotypic heterogeneity is generated as cells traverse an "epigenetic landscape," conceptualized originally by Waddington [69] and subsequently repurposed by others in the context of modern biology [70]. As cells (spheres) divide and differentiate, the space of possible phenotypes is structured by regions of relative stability, represented by grooves in the landscape. The manifold depicted here exists in two dimensions representing lineage and development, respectively. (b) Cells generated on the manifold depicted in (a) and measured by a 3-dimensional single-cell technology produces a nonlinear embedding of the manifold in $\mathbb{R}^3$.

(1) Computational techniques for learning cell states from high-dimensional single-cell data

(2) Insights into normal and malignant hematopoiesis obtained from **(1)**

Chapters 2 and 3 concern different aspects of item **(1)**. Chapter 2 addresses the fundamental task of subpopulation discovery: Given nothing but single-cell measurements, what are the distinct cell states reflected in the data? Using the concepts developed above, I detail an algorithmic solution to this problem and demonstrate its superior performance to other methods.

Chapter 3 extends the work of Chapter 2 to a more structured setting. Given partial knowledge about a sample, can this knowledge be extended to categorize the remaining, unknown portion of the sample?

In Chapters 4 & 5, the methods developed in Chapters 2 & 3, respectively, are applied to mass cytometry data collected from primary human bone marrow of a cohort of leukemia patients and healthy donors. Chapter 4 deals with fundamental questions regarding the

phenotypic composition of these samples, as revealed by the data-driven method introduced in Chapter 2. Chapter 5 drills into the functional states associated with these subpopulations and applies a suite of computational techniques—including the content of Chapter 3—to generate and corroborate inferences about these functional states. Taken together, Chapters 4 & 5 indicate that the immunophenotypes of leukemic blasts often belie their functional state, the latter of which can be learned by computational techniques and correlate with survival.

Finally, Chapter 6 places the work in a broader context, summarizing what has been established in the previous chapters and suggesting future directions for learning cell states from high-dimensional single-cell data.

Chapter 2

Extracting cell states from graphs of phenotypes

Complex tissues are composed of biologically meaningful subpopulations that are phenotypically coherent despite the intrinsic variability that makes each cell unique. A fundamental challenge in quantitative analysis of single-cell data is to establish the major cell states present, enabling an efficient and meaningful profile of the tissue.

Bone marrow is a common model system for such studies due to the rich diversity of immune cells found therein, not to mention experimental tractability (e.g., dissociation of cells into suspension). The wide range of cell types present in bone marrow has been established by decades of experimental research, which has produced a taxonomy of functionally distinct lineages (lymphoid and myeloid at the broadest level) and developmentally related progenitors.

While normal immune cells are typically binned into predefined "landmark" cell subsets, this strategy is unsuitable for less predictable or under-studied tissues such as cancer, where new phenotypes have been shown to occur [66]. Thus a data-driven, unsupervised approach is needed that takes single-cell measurements as input and returns a grouping of cells into distinct subpopulations—i.e., clusters.

2.1 Related Work

Dimensionality reduction techniques such as t-distributed stochastic neighbor embedding (t-SNE) [67, 68] help visualize the data but do not explicitly partition samples into distinct subpopulations. Moreover, while a two- (or three-) dimensional projection is necessary for visualization, this amount of dimension reduction is not intrinsically desirable. Subpopulations that are not visually separable in an optimal two-dimensional projection may in fact be distinct in high-dimensional space, were this requirement to be relaxed. Therefore, clustering methods which do not seek reduced dimensionality but rather operate in the full-dimensional space may distinguish subpopulations that appear indistinct even in an optimal two-dimensional projection.

Several methods have been proposed for clustering single-cell measurements into discrete phenotypic subpopulations. A thorough comparison of available methods was conducted by the "FlowCAP consortium" in 2013 [71]. We evaluated the best-performing methods from this contest, as well as standard clustering methods, and found that they did not perform well for mass cytometry data. Details of this comparison are discussed below in Section 2.3.

2.1.1 Parametric vs. Nonparametric methods

Clustering methods can be organized into two broad classes.

Parametric methods assume that the data are generated by a mixture of parametric densities, in which case a clustering solution is obtained by finding parameter values that maximize the probability of the data. This class of methods, of which the Gaussian mixture model is the best known (both in general and among cytometry experts [72]), have conceptual appeal but tend to perform poorly on real data. The parametric densities impose strong assumptions about cluster shape (e.g., ellipsoid, convex, symmetric) which become increasingly unlikely (and difficult to verify) in high-dimensional space. Attempts to reduce the severity of these assumptions within the parametric framework do so at prohibitive cost in terms of tractability and scalability [73]. Even supposing that a parametric model correctly specifies the underlying distribution, parameter estimation is not robust to noise [74], a pervasive fea-

ture of any real-world data set. The parametric approach seems particularly undesirable when one further considers that the measurement space of a single-cell data set generally presents a nonlinear embedding of the true underlying phenotypes [68].

Nonparametric methods eliminate the assumptions of statistical models by avoiding these models altogether. Being defined by what they are not (i.e., parametric methods), this class is more heterogeneous. Typically, nonparametric methods define some objective function that expresses a 'good' clustering solution and seek an assignment of data instances to clusters such that this function is optimized. k-means is a paradigmatic example. Here, the objective function (WCSS; Sec. 1.3.2) induces a centroidal Voronoi tessellation (Figure 2.1) in which clusters are not necessarily ellipsoid or symmetric but are always convex. While weaker than the other two assumptions, convexity is still a strong assumption about cluster shape—again, especially in the context of a nonlinear embedding. As a problem, the convexity assumption can be side-stepped by specifying a large k, since any non-convex set can be broken into smaller convex subsets. A number of methods developed specifically for cytometry data use k-means (or similar) to produce a starting point with a large number of small convex sets and then apply various heuristics to build clusters from these elements by merging [75, 76]. The trouble with such heuristics is that, lacking principled motivation, they are often unstable and exhibit sensitivity to small changes in the data, giving inconsistent results even under small perturbations such as resampling (§2.3).

Spectral methods are a distinctive subclass of nonparametric clustering methods. SamSPECTRAL [77] is one such method developed for cytometry data. Rather than operate on the data directly, these methods operate on a matrix of pairwise similarities and in so doing are more closely related to manifold learning (§1.3.3). As such, spectral methods more successfully handle arbitrary shapes resulting from nonlinear embedding. The downside of spectral methods is that they can scale very poorly with the number of data points and become prohibitively expensive in terms of computation. For instance, in [77] the authors estimate that a standard spectral method, with $\mathcal{O}(N^3)$ running time and $\mathcal{O}(N^2)$ memory requirement, would take 2 years and require 5 terabytes of memory to process a 300,000-cell data set. To

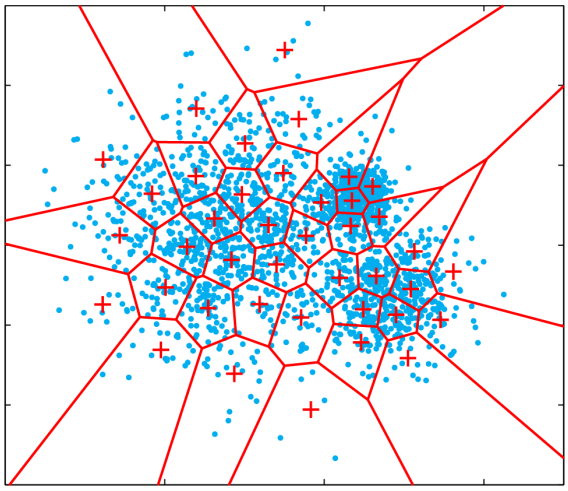


Figure 2.1: Example of a Voronoi tessellation in two dimensions, reproduced from [65]. The "+" symbols represent cluster centroids and the red lines identify the Voronoi regions induced by their locations.

address this problem, SamSPECTRAL deploys a heuristic down-sampling scheme that throws away data until a 'representative' set of $1500-3000$ cells ($0.5-1\%$ of a 300,000-cell data set!) is established and clusters are determined from operations on this set.

2.2 PhenoGraph: unsupervised subpopulation discovery

For the new era of high-throughput, high-dimensional single-cell biology, a method is needed that can robustly identify subpopulations *ab initio*. Such a method ideally possesses the following properties:

- Scales favorably with data size (N) and dimensionality (D)
- Handles nonlinear embedding and arbitrary cluster shapes
- Finds clustering solutions that are robust to user-specified input and random resampling

In this section I present a graph-based method that satisfies these requirements.

At the outset, single-cell measurements can be regarded as points in the D-dimensional space of non-negative real numbers, $\mathbb{R}^D_{\geq 0}$, where D is the number of biological features measured in the experiment. A sample of N cells can be represented by the $N \times D$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$. Unsupervised subpopulation discovery is tantamount to clus-

tering the rows of $\mathbf{X}$, i.e., defining a partition of the N cells into $K < N$ disjoint sets $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$.

Due to complex dependencies between the measured covariates, one can expect that the D-dimensional measurements $\mathbf{x}_i$ correspond to locations on a lower-dimensional topological space, $\mathcal{P}$, which we call the *phenotypic manifold* (§1.3.3). The embedding of $\mathcal{P}$ in $\mathbb{R}^D_{\geq 0}$ is not linear in general.

To learn the phenotypic structure of the sample, we should prefer to handle $\mathcal{P}$ rather than $\mathbf{X}$. For one, the lower dimensionality of $\mathcal{P}$ mitigates the analytical difficulties associated with high-dimensional spaces. More importantly, from a theoretical perspective, the distribution of points on $\mathcal{P}$ reveals the important cell states in the sample. To see this, consider each individual cell as a dynamical system that is able to traverse $\mathcal{P}$ by the coordinated changes in expression that produce the phenotypic and functional heterogeneity of the population. Cells that reach robust cell states—defined by their dynamic stability—will tend to stay in that region of the phenotypic manifold for prolonged durations. Thus, when the population is sampled, the dense regions of $\mathcal{P}$ identify the robust cell states.

We are not able to measure $\mathcal{P}$ but we can estimate it from $\mathbf{X}$. A discrete approximation of $\mathcal{P}$ can be constructed by mapping each point in $\mathbf{X}$ to a vertex in a **graph**. Formally, the graph $G = (V, E)$ is an ordered pair in which V is a set of N vertices connected to each other by edges in the set E.[1] If the edges in E only identify short distances—i.e., only the most similar points are directly connected—G becomes a discrete approximation of $\mathcal{P}$ [65, p. 100].

Thus the work of identifying cell states can be broken into two tasks: 1) constructing G that approximates $\mathcal{P}$, and 2) analyzing the structure of G to identify dense regions.

2.2.1 Representing phenotypes in a graph

Previous work in our group leveraged the graph representation of a phenotypic manifold to learn developmental trajectories in B cell lymphopoeisis [53]. In that setting, the task was not density estimation but rather dimension reduction (to a one-dimensional developmental

[1] Please note some terminological equivalences that can be found in the literature: {*graph* ⇔ *network*; *vertex* ⇔ *node*; *edge* ⇔ *arc* ⇔ *link* }

axis). In the present setting, the task is clustering and this should motivate the way in which the graph is constructed. Indeed, G is a dramatically different representation of $\mathbf{X}$ and the quality of the mapping $f : \mathbf{X} \to G$ determines how closely G approximates $\mathcal{P}$.

How does one convert a set of multidimensional coordinate vectors to a graph? The bijective mapping of cells to vertices is straightforward: $\mathbf{x}_i \to v_i, \forall i \in \{1, \ldots, N\}$. The question is really how to design the set of edges such that paths through the graph represent available paths through $\mathcal{P}$—not more and not less. If the graph contains too many edges, distant regions of $\mathcal{P}$ become connected by *short circuits* and the graph loses its advantages over the original measurement space. The problem of short circuits is seriously exacerbated by the presence of noise in the measurements, another challenge that must be addressed by the graph construction procedure. If the graph contains too few edges, it can fracture into spuriously disconnected components and information is lost about the structure and density of points on $\mathcal{P}$. Finally, edges can be binary (present or absent) or they can be weighted, in which case the weight reflects the strength of the connection between the vertices on either side of it. The question therefore arises whether edges should have weights and if so how those values are determined.

As a starting point, there are two simple strategies for converting coordinate data to a graph: the ϵ-rule and the k-rule. Both are motivated by the definition of manifold as a topological space that is *locally* Euclidean (§1.3.3). This leads to the concept of neighborhood: at each point $\mathbf{x}$ there is a neighborhood of local points $\mathcal{V}(\mathbf{x})$ for which the Euclidean metric is a valid distance metric. If we knew the neighborhood of each point, graph construction would be trivial: specify edges such that each vertex is connected exactly to its neighborhood. Instead, we guess by selecting each point's "nearest neighbors." By the ϵ-rule, points falling within a D-dimensional hypersphere of radius ϵ are identified as neighbors, $\mathcal{V}_\epsilon(\mathbf{x})$. By the k-rule, the k nearest points are identified as neighbors, $\mathcal{V}_k(\mathbf{x})$. Both rules require specification of a parameter that controls the neighborhood size and a distance metric that defines proximity.

The ϵ-rule is closer in spirit to the motivating concepts of manifold and neighborhood. However, in practice it is very difficult to select an appropriate value for ϵ, especially when

the significance of interpoint distances can vary across different regions of the measurement space [78], as may very well be true in single-cell data. Selecting an appropriate ϵ requires a depth of knowledge about the structure of the data that presupposes the very task for which graph construction is intended. The k-rule, on the other hand, may lack the conceptual appeal but is more effective in practice. The procedure adapts to different scales by ranking distances at each point. The number of neighbors is an intuitive quantity to select compared to the radius of a D-dimensional hypersphere and provides more direct control over the sparsity of the graph, which is important for practical applications. However, the k-rule has important limitations that are relevant for graph clustering. While the k-rule overcomes the variable-scale problem, it does so at the expense of ignoring differences in local density: every point selects k neighbors regardless of whether it is at the center of a dense cluster or outlying in a sparse region.

This problem can be addressed by considering a weighted graph. In a binary graph, the *degree* of vertex v_i, $\deg(v_i)$, denotes the number of edges in the graph that involve v_i, and this quantity can be understood as a measure of density at each vertex. This means that the degree of every vertex in a binary k-neighbor graph is at least k (note that $\deg(v_i) > k$ when $\exists j : \mathbf{x}_i \in \mathcal{V}_k(\mathbf{x}_j) \wedge \mathbf{x}_j \notin \mathcal{V}_k(\mathbf{x}_i)$). The definition of degree can be generalized for weighted graphs. Let $w : i \sim j \to \mathbb{R}_{\geq 0}$ be a weight function that assigns a non-negative weight to each edge $i \sim j \in E$ and let $\mathbf{W}$ denote the $N \times N$ weight matrix comprising elements $\mathbf{W}_{ij} = w(i \sim j)$. Note that for a binary graph $w(i \sim j) = 1, \forall\ i \sim j \in E$. In general, we can define the vertex degree:

$$\deg(v_i) = \sum_j \mathbf{W}_{ij} \tag{2.1}$$

which, in the case of a binary graph, equals the number of incident edges on v_i. For a weighted graph, $\deg(v)$ becomes a more flexible measure of density at each vertex, provided that edge weights scale with local density. If a weight function satisfies that condition, the graph combines the desirable properties of the ϵ-rule and the k-rule.

To define a weight function that scales with local density, consider the relationship between the density of a region and the k-neighborhoods in that region. Specifically, when two

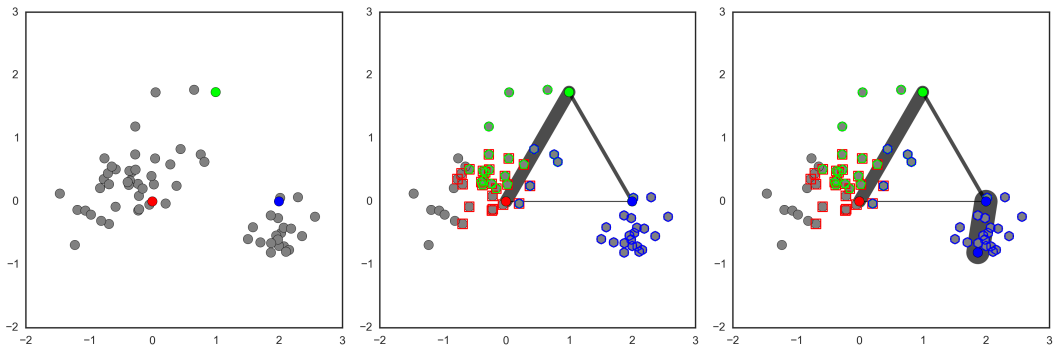
points are located in the same dense region, they tend to have neighbors in common—their neighborhoods will overlap, $\mathcal{V}_k(\mathbf{x}_i) \cap \mathcal{V}_k(\mathbf{x}_j) \neq \emptyset$. On the other hand, if two proximal points are involved in separate density structures, their neighborhoods will tend to be disjoint. Let

$$J_k(\mathbf{x}_i, \mathbf{x}_j) = \frac{|\mathcal{V}_k(\mathbf{x}_i) \cap \mathcal{V}_k(\mathbf{x}_j)|}{|\mathcal{V}_k(\mathbf{x}_i) \cup \mathcal{V}_k(\mathbf{x}_j)|} \tag{2.2}$$

denote the Jaccard coefficient between the k-neighborhoods of cell i and cell j.[2] The Jaccard coefficient quantifies the similarity between two sets and is bounded in the closed interval $[0, 1]$, reaching the upper and lower bounds when the two sets are identical or disjoint, respectively. In quantifying the overlap of $\mathcal{V}_k(\mathbf{x}_i)$ and $\mathcal{V}_k(\mathbf{x}_j)$, Eq. (2.2) reflects whether i and j participate in the same density structure (Figure 2.2). As a corollary, when a point is central to such a structure, it will have overlapping neighborhoods with a large number of points and the total edge weight involving that point, $\deg(v)$, will be large. Thus Eq. (2.2) provides the desired scaling with point-wise density (Figure 2.3).

Thus, we define the *Jaccard graph* as a weighted k-neighbor graph in which edge weights are given by Eq. (2.2), the Jaccard coefficient between k-neighborhoods. This graph requires specification of the parameter k, which provides an "initial" estimate of neighborhood size. As we have seen, the weight function relaxes the influence of k by incorporating the local data distribution to tune the effective density at each point. As a result of this tuning, the properties of the graph are robust to the choice of k, as demonstrated in a subsequent section (§2.3). Importantly, this weighting procedure distinguishes between two very different types of rare observations: outliers due to noise and genuine rare cell types. While both may occur at a similar frequency in the data (0.5%, for example), because outliers are generated by noise, they have few shared neighbors with other points and their overall influence in the graph is dampened. Conversely, the co-occurrence of genuine rare cells in close phenotypic proximity results in highly overlapping neighborhoods, which translates to a set of strongly interconnected vertices, signifying the presence of a robust cell type. Overall, the Jaccard

[2] It is interesting in the context of this section that the name given by Jaccard for this quantity was *le coefficient de communauté* [79].



(a) (b) (c)

Figure 2.2: The Jaccard coefficient between k-neighborhoods provides a similarity measure that reflects the structure of data density. (a) Two densities separated by a sparse region. The red, green, and blue points are mutually equidistant (in the Euclidean sense). (b) The open shapes label the k-neighborhoods of the three equidistant points by corresponding color ($k = 25$). The magnitude of the Jaccard coefficient between each pair of neighborhoods is represented by the width of the line connecting each colored point. Note that the lines would have equal width if similarity were quantified by the Euclidean metric alone. (c) A second point within the bottom-right density illustrates the relative insignificance for the blue point of the small neighborhood overlaps it shares with the red and green points.

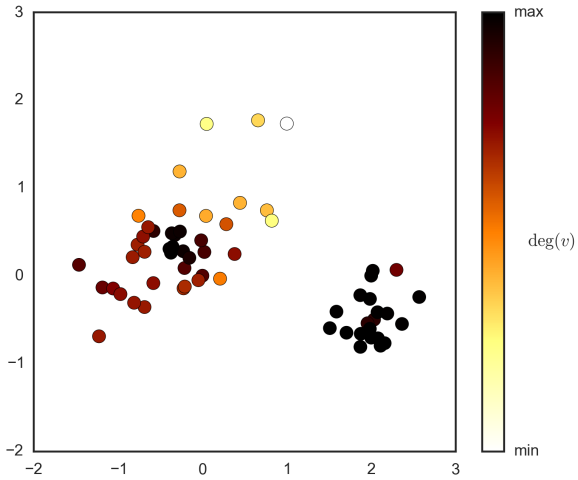


Figure 2.3: The same sample data as in Figure 2.2. The Jaccard weight is calculated between all k-neighborhoods ($k = 25$) and the vertex degree at each point i ($\sum_j \mathbf{W}_{ij}$) is represented by color. Points that are central to dense regions have the largest degree.

graph uses consensus between the perspectives of each point to penalize spurious edges and to strengthen well-supported ones.

2.2.2 Community detection in a graph of phenotypes

Once a graph, G, has been constructed, identifying subpopulations can be cast as a search for highly interconnected subgraphs of G. To this end we borrow from social network research, which has developed powerful algorithms to partition large networks into *communities* [80]. In our setting, communities represent an accumulation of phenotypically similar cells that likely reflects biologically meaningful phenotypic stability, thus revealing stable cellular states in the population. Partitioning the graph into these communities produces a dissection of the population into phenotypically coherent subpopulations.

A simple yet powerful measure on the community structure of a graph is the *modularity*. For a (weighted) graph and a set of community assignments $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$, the modularity is:

$$Q(\mathbf{W}, \mathcal{C}) = \frac{1}{m} \sum_{ij} \Big[\mathbf{W}_{ij} - \frac{\deg(v_i) \deg(v_j)}{m} \Big] \delta(c_i, c_j) \tag{2.3}$$

where $\mathbf{W}_{ij}$ is the edge weight between vertices i and j, $c_i \in \{1, \ldots, K\}$ is the community assignment of vertex i, the Kronecker delta function $\delta(u, v) = 1$ if $u = v$ and 0 otherwise, and $m = \sum_{ij} \mathbf{W}_{ij}$ is the total weight in the graph and serves as a normalization factor [81]. The modularity quantifies the total intra-community edge weight beyond what is expected by chance from the number of edges and the degree distribution. The modularity is 0 when the partition $\mathcal{C}$ is random, and reaches a maximum of 1 when edges exclusively connect vertices in the same community, i.e., when $\mathcal{C}$ identifies K disconnected components.

Given some proposed community assignments $\mathcal{C}$, Eq. 2.3 can be used as a quality measure on the partition. Alternatively, Eq. 2.3 can be used as an objective function that is maximized in the search for a good partition. In this respect, it is interesting to note that Eq. 2.3 is somewhat similar to Eq. 1.1, the k-means objective function. Given an assignment of points to discrete classes, both provide a quality measure based on the "tightness" of the points in each class. Modularity has other desirable properties, not least of which is taking into account

an expected value of tightness for random partitions.

Note that while much of the theory and algorithmic development regarding modularity have been pursued and applied in social network analysis, it is actually a more general quantity. Equation 2.3 is formally equivalent to the Hamiltonian of the Potts model of statistical mechanics; as such, it is a natural quantity describing the "correctness" (in an energy-minimization sense) of an assignment of interacting elements to a number of discrete states [82].

Finding a set of community assignments that maximize Eq. 2.3 is a combinatorial optimization problem for which exact solutions are computationally intractable but for which good heuristic approximations have been developed. In particular, the Louvain method [83] has become popular due to its efficiency on large graphs (up to hundreds of millions of vertices). The Louvain method is hierarchical and agglomerative. At the beginning of the first iteration every vertex is placed in its own cluster. At each iteration, all vertices are scanned and each vertex v_i is added to the community of its neighbor v_j that yields the greatest increase in modularity, ΔQ. If there is no v_j such that $\Delta Q > 0$, v_i is left in its current community. The process is repeated hierarchically (representing bottom-level communities as vertices in the next iteration, etc.) until no further increase in Q is obtained. An appealing property of the Louvain method is that eventually moving to a new hierarchical level does not increase Q, suggesting that a "natural" resolution has been reached. Because the same objective function is used throughout, the top level of the hierarchy is guaranteed to have a higher modularity score than a level below and the communities specified at the top level can be taken as the clustering solution. Thus, the number of clusters is determined in parallel with the communities themselves and reflect a level of detail that is optimal with respect to the modularity.

PhenoGraph

Combining the concepts discussed above, we developed PhenoGraph, an algorithmic approach to defining subpopulations in single-cell data. PhenoGraph begins by converting single-cell

measurements to a graph of phenotypes. The Jaccard graph is used, which incorporates crucial information about the data density by way of edge weights. These weights are instrumental in defining strongly interconnected communities of phenotypes according to the modularity score (Eq. 2.3). The Louvain method is used to find a partition of the graph into communities that maximize the modularity. Because the output of the Louvain method depends on a random initialization, PhenoGraph runs multiple random restarts in order to avoid poor local maxima and ensure a high-quality solution. Pseudocode for PhenoGraph is presented in Algorithm 1. Two open-source software implementations of PhenoGraph are currently available: a standalone Python version[3] and as part of the MATLAB single-cell analysis toolbox *cyt* developed in our lab.[4]

Algorithm 1 PhenoGraph Clustering

IN: Single-cell measurements $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, neighborhood size k, r random restarts
OUT: Subpopulation assignments $\mathcal{C}$
procedure GRAPH CONSTRUCTION($\mathbf{X}, k$)
  for all $i \in \{1, \ldots, N\}$ **do**
    for all $j \in \mathcal{V}_k(\mathbf{x}_i)$ **do**
      $\mathbf{W}_{ij} \leftarrow J_k(\mathbf{x}_i, \mathbf{x}_j)$
    end for
  end for
  return W
end procedure
procedure GRAPH CLUSTERING($\mathbf{W}$)
  $Q_{\max} \leftarrow 0$
  $t \leftarrow 0$
  while $t < r$ **do**
    $\mathcal{C}_t \leftarrow$ LOUVAIN($\mathbf{W}$)
    $Q_t \leftarrow Q(\mathbf{W}, \mathcal{C}_t)$
    if $Q_t > Q_{\max}$ **then**
      $\mathcal{C} \leftarrow \mathcal{C}_t$
      $Q_{\max} \leftarrow Q_t$
    end if
    $t \leftarrow t + 1$
  end while
  return $\mathcal{C}$
end procedure

[3] https://github.com/jacoblevine/PhenoGraph
[4] http://www.c2b2.columbia.edu/danapeerlab/html/phenograph.html

While the number of random restarts of the Louvain subroutine can be considered an input parameter to PhenoGraph, in practice we found that each restart tends to produce results highly similar to the others, such that good results are found within a small number of solutions. For example, the standard deviation of modularity scores obtained from 100 random restarts on test data (§2.3) was extremely small (8.75×10^{-4}). Once the graph has been constructed the computational cost of random restarts is small and this parameter can be fixed at some moderately large value (e.g., 50), or other heuristics can be used to determine that a solution is not a poor local maximum.

2.3 Validation with normal bone marrow data

Healthy human bone marrow, which is rich in distinct and well-characterized immunological cell types, presents an opportunity to evaluate the performance of computational methods for subpopulation discovery. In particular, two questions are of interest:

1. Do clustering solutions recapitulate *ab initio* the immunological subsets that are known from decades of research?

2. Does PhenoGraph perform favorable in comparison to other methods?

To address these questions, three validation data sets were assembled, comprising mass cytometry measurements of bone marrow mononuclear cells (BMMCs) from three healthy donors.

Validation Data Set 1 (VDS1) was a publicly available mass cytometry data set [17] (`http://reports.cytobank.org/1/v1`) of healthy adult BMMCs. It consisted of 167,044 cells collected from a healthy human donor ("Marrow 1" in [17]), which had been manually assigned to 24 cell types by standard immunological gating techniques in the original publication. The gating strategy for manual assignment was based on the 13 surface markers measured in this data: CD45, CD45RA, CD19, CD11b, CD4, CD8, CD34, CD20, CD33, CD123, CD38, CD90, CD3. These 13 markers were also used for all computational analyses. Manual gating assigned 49% of the cells to a known cell type while the remaining cells were not assigned to any known cell type.

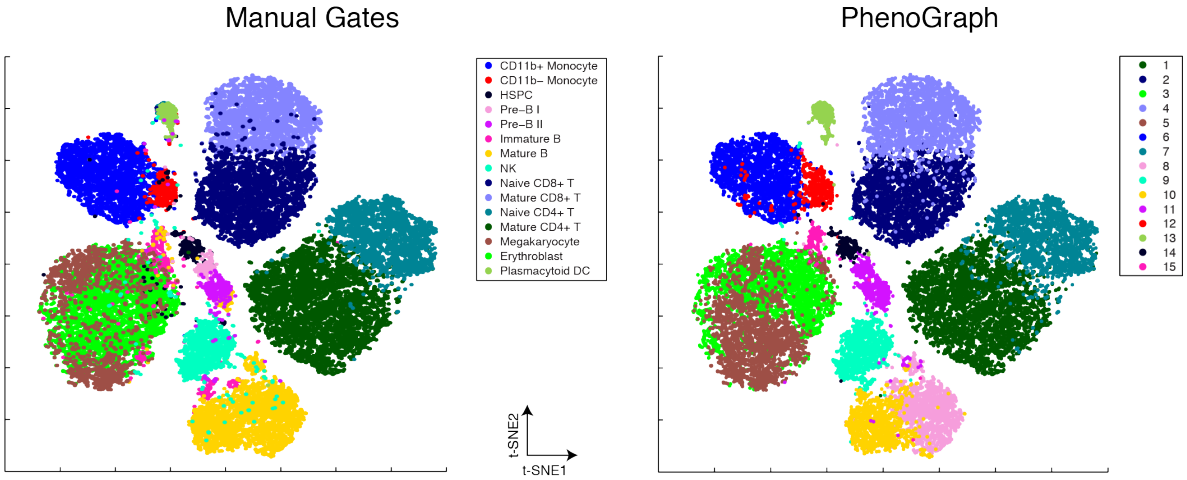


Figure 2.4: 30,000 random cells from VDS1 with manual cell type assignments, visualized with *t*-SNE. Cells are colored by cell type assignments established by manual gating (left panel) or subpopulations detected by PhenoGraph (right panel).

Validation Data Sets 2 & 3 were newly collected mass cytometry measurements of 32 surface markers on BMMCs from two healthy adult donors. Measurements from these samples were manually gated into 14 cell types based on 19 markers measured in this data: CD3, CD4, CD7, CD8, CD15, CD16, CD19, CD20, CD34, CD38, CD41, CD44, CD45, CD61, CD64, CD123, CD11c, CD235a/b, HLA-DR (note that manual gating becomes extraordinarily difficult even at this dimensionality). The full 32 marker-panel was used for all computational analyses of these data. (Data are available for download at `http://cytobank.org/nolanlab/reports`).

Before a systematic evaluation, it is instructive to visualize the performance of PhenoGraph and other methods on healthy bone marrow. For this purpose, we assembled a curated sample of 30,000 cells from VDS1 that included only cells that were assigned to a known cell type by the manual gating strategy. PhenoGraph was run on this sample to obtain a data-driven approximation to the manually-defined cell types. Figure 2.4 shows the PhenoGraph assignments in direct comparison to the manual assignments, visualized by *t*-SNE. Note that neither *t*-SNE nor PhenoGraph use the manual assignments in any way. Also note that both *t*-SNE and PhenoGraph are run directly on the data matrix and that PhenoGraph does not use any information from *t*-SNE.

A qualitative analysis of Figure 2.4 reveals that PhenoGraph correctly identifies clusters corresponding to most major healthy cell types including naive and mature $CD4^+$ T cells, naive and mature $CD8^+$ T cells, natural killer (NK) cells, $CD11b^+$ and $CD11b^-$ monocytes, pre-B cells, plasmacytoid dendritic cells (pDCs), megakaryocytes and erythroblasts. It found a single cluster of early B cells, grouping together the manually assigned Pre-B I and Pre-B II populations. It found two mature B-cell clusters, splitting the B-cell population into two—based on low and mid expression of CD123—in a manner that appears corroborated by the structure of the t-SNE map.

2.3.1 Quality measures

To quantify the agreement of clustering solutions with the manual assignments, we used two alternative quality measures: the mean F-measure and the normalized mutual information (NMI). The mean F-measure was used in the recent FlowCAP competition and we used this procedure as described in their publication for the sake of consistency [71].

For each "ground truth" subpopulation γ_i in the benchmark data and a cluster c_j returned by the algorithm, Precision (Pr_{ij}) quantifies the proportion of c_j that identifies γ_i, Recall (Re_{ij}) quantifies the proportion of γ_i that is identified by c_j, and the F-measure is defined as the harmonic mean of these quantities:

$$F(\gamma_i, c_j) = \frac{2 \cdot \mathrm{Pr}_{ij} \cdot \mathrm{Re}_{ij}}{\mathrm{Pr}_{ij} + \mathrm{Re}_{ij}} \tag{2.4}$$

The F-measure quantifies the accuracy of a binary classification, but can be extended to M classes $\Gamma = \{\gamma_1, \ldots, \gamma_M\}$ by taking the weighted average:

$$F_{\mathrm{mean}}(\Gamma, \mathcal{C}) = \sum_{\gamma_i \in \Gamma} \frac{|\gamma_i|}{N} \max_{c_j \in \mathcal{C}} F(\gamma_i, c_j) \tag{2.5}$$

where $|\gamma_i|/N$ is the proportion of the validation data in subpopulation γ_i and $\mathcal{C} = \{c_1, \ldots, c_K\}$ is the clustering solution being evaluated. The mean F-measure is bounded by the interval $[0, 1]$ with 1 representing a clustering solution that perfectly matches the ground truth (in

this case, the manually-defined subpopulations).

To avoid any possible biases inherent in the mean F-measure, we also used the normalized mutual information (NMI), another popular clustering quality measure. This score treats the true cluster assignments and the output of the algorithm each as discrete random variables and quantifies their statistical redundancy, which reflects the clustering accuracy (note that a perfect clustering result and the true labels are completely redundant—they contain exactly the same information). This redundancy is captured by the mutual information:

$$I(\Gamma;\mathcal{C}) = \sum_{\gamma\in\Gamma}\sum_{c\in\mathcal{C}} P(\gamma,c)\log\Big(\frac{P(\gamma,c)}{P(\gamma)P(c)}\Big) \tag{2.6}$$

which is non-negative but otherwise unbounded. The normalized variant

$$\mathrm{NMI}(\Gamma,\mathcal{C}) = \frac{I(\Gamma;\mathcal{C})}{\sqrt{H(\Gamma)H(\mathcal{C})}} \tag{2.7}$$

where $H(\cdot)$ is the Shannon entropy, reaches a maximum of 1 when $\mathcal{C}\equiv\Gamma$.

2.3.2 PhenoGraph outperforms leading methods

Besides the quality of the clustering solution itself, there are other desirable properties to be evaluated. These are:

Robustness. The method should produce similar results under small perturbations of the data such as random resampling

Automation. The method should not require too many user-defined parameters and should produce similar results under small perturbations of their values

Scalability. The method should be able to process large, high-dimensional samples in a reasonable amount of time

Note that automation can be viewed as another form of robustness, i.e. robustness to user input.

I evaluated PhenoGraph on all these criteria together with the three best-performing open-source methods tested by the FlowCAP consortium: FLOCK [76], flowMeans [75], and SamSPECTRAL [77]. Additionally, two commonly used "general purpose" clustering methods were included: the Gaussian mixture model and hierarchical linkage. PhenoGraph quantifiably outperformed all competing methods, as detailed below.

For all comparisons the following implementations were used: FLOCK 2.0 (FlowCAP-I version) implemented in C; flowMeans 1.18.0 (Bioconductor version 3.0) implemented in R; SamSPECTRAL 1.20.0 (Bioconductor version 3.0) implemented in R. While FLOCK and flowMeans have no user-defined input, SamSPECTRAL requires tuning of two parameters, sigma (σ) and separation factor (sf), which were tuned as recommended in the user guide for that software. For hierarchical linkage clustering and Gaussian mixture models, the MATLAB R2013b implementations were used. In principle, those methods are not comparable because they require specification by the user of the number of subpopulations. In our testing, we provided this number to these methods (which is of course known for the validation data), though this did not result in an impressive performance from either method.

Integrated quality and robustness analysis

Cluster solution quality and robustness were evaluated simultaneously in the following manner. From each validation data set, we generated 50 random subsamples of 20,000 cells each. Each method was run on each subsample and the mean F-measure and NMI were computed on each individual output. This procedure generated distributions of quality measures for each method. Ideally, these distributions should have a high center (indicating high average solution quality) and small variance (indicating robustness to random resampling). The distributions of mean F-measure for the VDS1 subsamples are shown in Figure 2.5.

I also used the random subsamples to evaluate PhenoGraph's robustness to the user-defined parameter k. Clustering solutions were generated by PhenoGraph for each subsample at four values of k spanning a four-fold range. As shown in Figure 2.5, the value of k has essentially no impact on the quality of the solution or the sensitivity to random resampling.

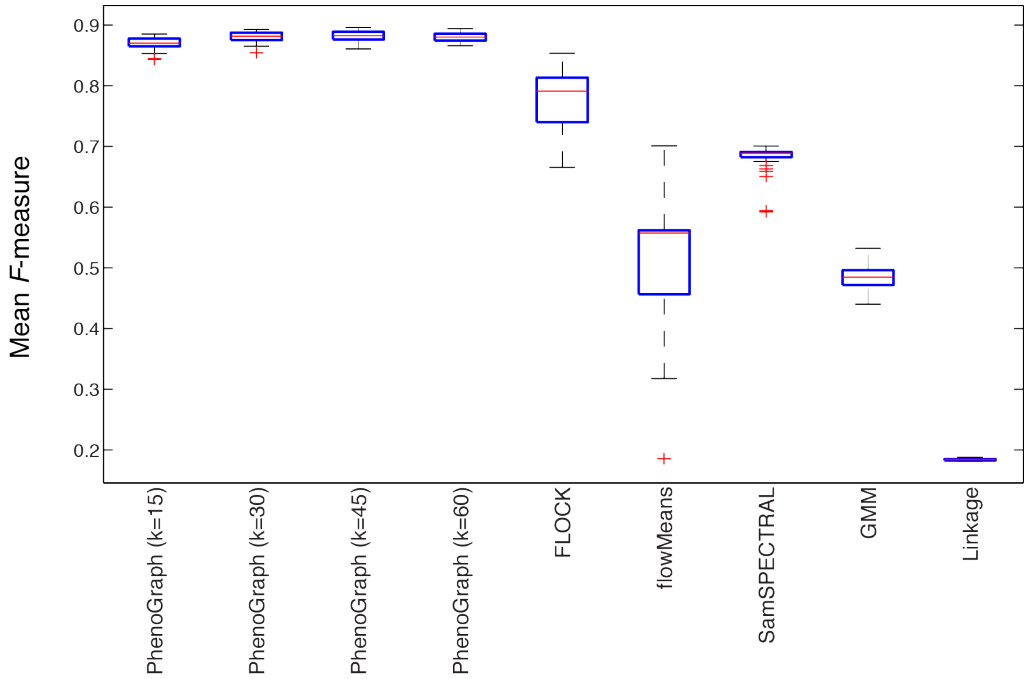


Figure 2.5: Distributions of mean F-measure obtained for each method on 50 random subsamples from VDS1.

This type of analysis could not be performed for FLOCK or flowMeans because those methods are completely automated—they take no input parameters. SamSPECTRAL takes two input parameters that require careful tuning. Instead of testing the quadratic number of possible pairs for the two parameters, these were selected as recommended in the user guide for that software.

All results using NMI were virtually identical and are omitted for space, but can be found in the supplementary material of [84].

From this analysis, it can be concluded that PhenoGraph outperforms all competing methods in terms of solution quality and robustness.

Analysis of scalability

As before, the input data matrix is $N \times D$. It is desirable that a method scale well with increases in both N and D.

PhenoGraph exhibits superior scaling with dimensionality D

The FlowCAP consortium tested performance on flow cytometry data only; methods that performed well there were probably not designed for the higher dimensionality of mass cytometry data. The 32-dimensional data of Validation Data Sets 2 & 3 makes this clear. As in the previous analysis, we generated 50 subsamples from VDS2 and VDS3 to generate quality measure distributions for each method. The results are presented in Figure 2.6. FLOCK was unable to run on these data sets because it assumes that the number of observations is vastly greater than the number of dimensions to the the extent that a 20,000-cell test set contained "'too few" observations for 32-parameter data. SamSPECTRAL was able to run, but with poor results and poor computational efficiency; flowMeans produced better results than SamSPECTRAL but at an enormous cost in terms of run time. Furthermore, the performance of flowMeans was highly unstable, sometimes producing good results and sometimes failing completely due to numerical underflow. Out of 100 runs, flowMeans failed to produce any result on 59 occasions. On the other hand, PhenoGraph continued to exhibit high quality, robust results and fast run times.

PhenoGraph exhibits superior scaling with number of observations N

Finally, we tested the scaling behavior as the data matrix contains increasing numbers of observations. Testing was conducted on 32-parameter data sampled from VDS2. At each sample size, $N \in \{20, 40, 60, 80, 100\} \times 10^3$, 5 random samples were generated. The small number of samples was motivated by the mounting inefficiency of flowMeans and SamSPECTRAL, which scale exponentially with N (Figure 2.7). PhenoGraph displays roughly linear scaling with N.

All performance comparisons were conducted on a 2.6 GHz Intel Core i7 with 16GB of RAM. At 80,000 cells, average run time was 105 minutes for SamSPECTRAL, 254 minutes for flowMeans, and 5 minutes for PhenoGraph. In the time it takes flowMeans to process 80,000 cells, PhenoGraph can process 10^6 cells.

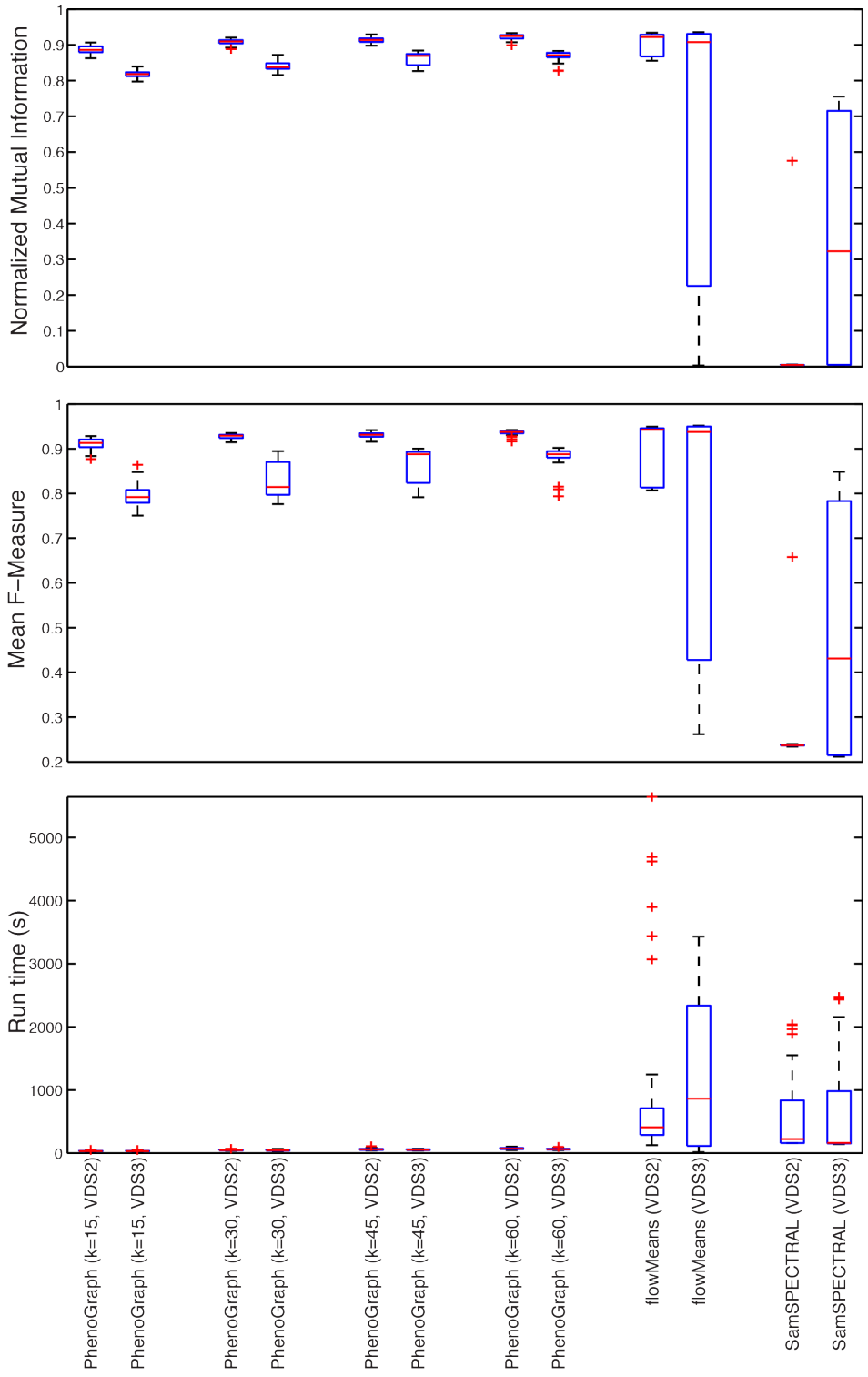


Figure 2.6: NMI, mean F-measure, and run time distributions for 50 random subsamples of 20,000 cells each from VDS 2 & 3.

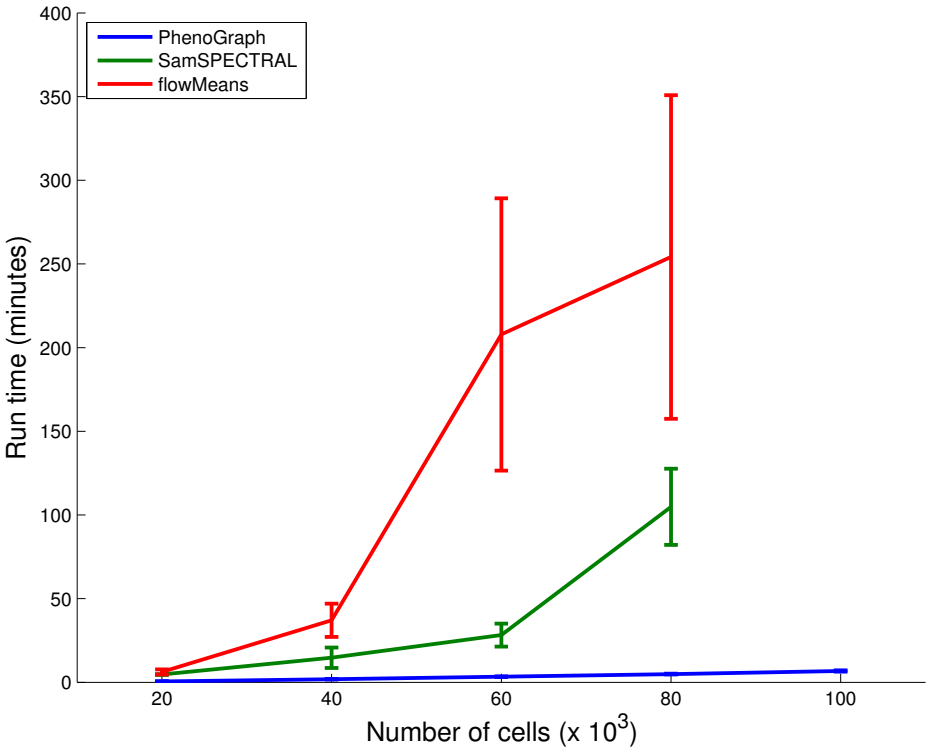


Figure 2.7: Systematic comparison of run time as a function of sample size. PhenoGraph (run here with $k = 30$) displays significantly superior computational efficiency and scales roughly linearly with the number of cells. Lines trace mean run time for 5 random samples at each sample size and error bars display the standard error.

2.4 Summary

In this chapter, we considered the fundamental problem of identifying cellular subpopulations in single-cell data. The particular features of single-cell data should be considered in designing an algorithmic solution to the problem. We used the conept of the phenotypic manifold (§1.3.4) to motivate our graph-based method, PhenoGraph. PhenoGraph uses a graph of single-cell phenotypes to approximate the phenotypic manifold and translates the task of density detection to the problem of modularity optimization in this graph.

Healthy human bone marrow contains a diversity of cell types well known through decades of research. Thus, single-cell measurements of bone marrow cells provide a good benchmark data set, where manual gate assignments provide an approximate gold standard. Using several such benchmark data sets, we showed that PhenoGraph is superior to other comparable methods in terms of all relevant metrics: PhenoGraph is computationally efficient and able to run on very large data sets without subsampling; PhenoGraph produces high quality results that closely resemble manual cell type assignments obtained through decades of experimental studies; PhenoGraph is robust, producing highly similar irrespective of the exact subsample of data points used and the exact setting of its single user-defined parameter.

The ability to reconstruct *ab initio* the known cell types of a well-characterized tissue is an important demonstration of the method's reliability. Ultimately, the data-driven approach becomes truly powerful when applied to less characterized systems, where we do not presume to know the subpopulation structure in advance. The quality of PhenoGraph's performance in the benchmark data translates to confidence in the biological significance of subpopulations it identifies in samples of unknown structure. In Chapter 4, PhenoGraph is used to identify subpopulations in leukemic marrow, where the subpopulation structure remains an open question.

Chapter 3

Classifying cells with random walks

Central to PhenoGraph is its underlying graph structure, constructed by representing data points as vertices connected by edges that capture their phenotypic and structural similarities. By partitioning this graph into maximum-modularity communities, we obtain a successful phenotypic dissection of the sample into subpopulations representing robust cell states. Theoretically, the reason for PhenoGraph's success is that the graph G is a good approximation to the phenotypic manifold $\mathcal{P}$. If so, then G should be useful for other tasks related to learning about $\mathcal{P}$. In this chapter, we consider the case where there is partial knowledge about the data and we want to extend this knowledge to uncharacterized cells. For example, we might construct a data set by concatenating measurements of known cell types together with a sample from some disease state. We could then ask, for each cell in the disease sample, what healthy cell type is it most similar to?

Another interesting application is recovering cells that are missed by manual gating strategies. Unlike data-driven approaches, manual gating labels only a portion of cells in a sample—those falling squarely within boundaries manually drawn in sequential bivariate subspaces of the data. As mentioned in Section 2.3, despite the fact that the authors of [17] were able to define 24 cell types from 13 dimensions by manual gating, 51% of cells "fell through the cracks" and were not assigned to any cell type. In this case, manual assignments can be considered as partial labels providing "seed" information about cell types. Computationally

extending the manual assignments to the unlabeled cells may not only prevent the loss of so much data, but may also reveal unappreciated phenotypic features of these cell types when they are identified in this more data-driven manner.

3.1 Problem formulation

Formally, consider a data set $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ in which $L < N$ points are assigned to one of M distinct classes. Without loss of generality, the assigned points are the first L rows of $\mathbf{X}$ and their assignments are given by the vector $\mathbf{y} = [y_1, \ldots, y_L]$, which assigns each of the first L samples in $\mathbf{X}$ to one of M distinct classes. The inference task is to compute assignments for the unlabeled points, $\mathbf{y}' = [y_{L+1}, \ldots, y_N]$. Because the task is to assign unlabeled points to classes specified in $\mathbf{y}$, this is a *classification* problem. As such, the labeled points are called the *training data* and the unlabeled points are called the *test data*.

In machine learning, this style of classification is called semi-supervised or transductive learning. It differs from supervised or inductive learning in that both the training and test data are given up front. Thus the inference procedure is permitted to use not only the features of the training data but also the structure of the entire data set to generate classifications. This is a powerful advantage indeed, largely for the same reasons that motivate the use of a graph for manifold learning. When assessing whether an unlabeled point is similar to a set of labeled points, their respective locations on a low-dimensional manifold should be at least if not more informative than metric distances in high-dimensional space.

3.2 PhenoGraph Transductive Learning

The extension of PhenoGraph to the transductive case can be thought of as a refinement of the k-nearest neighbors classifier, a classical and very simple non-parametric supervised learning technique. In that method, for each unlabeled point $\mathbf{x}_i$, the k nearest labeled points are identified and $\mathbf{x}_i$ is classified as the majority class among those k labeled points. This approach is vulnerable to many pitfalls. For example, small changes in k can easily tip the

balance of the majority from one class to another, except in the simplest cases. Because it is inductive rather than transductive, this method treats each unlabeled point independently, rather than transferring information between the unlabeled points as classifications accrue. Reformulating the problem as a transductive learning task allows relaxation of the strict value k to a more "distributed" vote that takes into account the larger-scale structure of the data.

Intuitively, each point i is allowed to be influenced by labeled vertices beyond its immediate neighborhood, connected to i through multiple edges in the graph. The classification becomes a weighted vote, averaged over the entire graph. The weights are derived from a probabilistic interpretation of the graph connectivity using *random walks*. A random walk is a path through the graph where, at each vertex i, the next vertex is selected probabilistically from the set of i's neighbors, $\mathcal{V}(v_i)$. The proximity of two vertices can be computed as the probability of a random walk passing through both—which, because it averages over all possible walks—is a powerful and robust measure of vertex similarity that takes into account the entire structure of the graph. Discussed below, this framework provides the mathematical foundation to allow a set of partial vertex labels to "diffuse" through the graph onto the unlabeled vertices, producing transductive classifications.

To extend PhenoGraph to the transductive case, we begin with the same graph construction procedure as before, using the Jaccard coefficient between k-neighborhoods as edge weights. The graph is constructed using all data and without reference to the labels. At this point, the graph is dappled with labeled vertices which act as landmarks for regions of the graph associated with the various classes (Figure 3.1).

Exploiting the connection between random walks on graphs and discrete potential theory [85], the probability of cell i being assigned each class m can be calculated by solving a system of linear equations representing the electric potential at each unlabeled node when voltage is alternatively applied to the vertices of each labeled class [86]. Each unlabeled node is then assigned to the class that it reaches first with highest probability.

Given a partially labeled data set $(\mathbf{X}, \mathbf{y})$, compute the Jaccard graph as described in Section 2.2.1. From the weight matrix $\mathbf{W}$ and its degree matrix $D = \mathrm{diag}(\deg(v_1), \ldots, \deg(v_N))$,

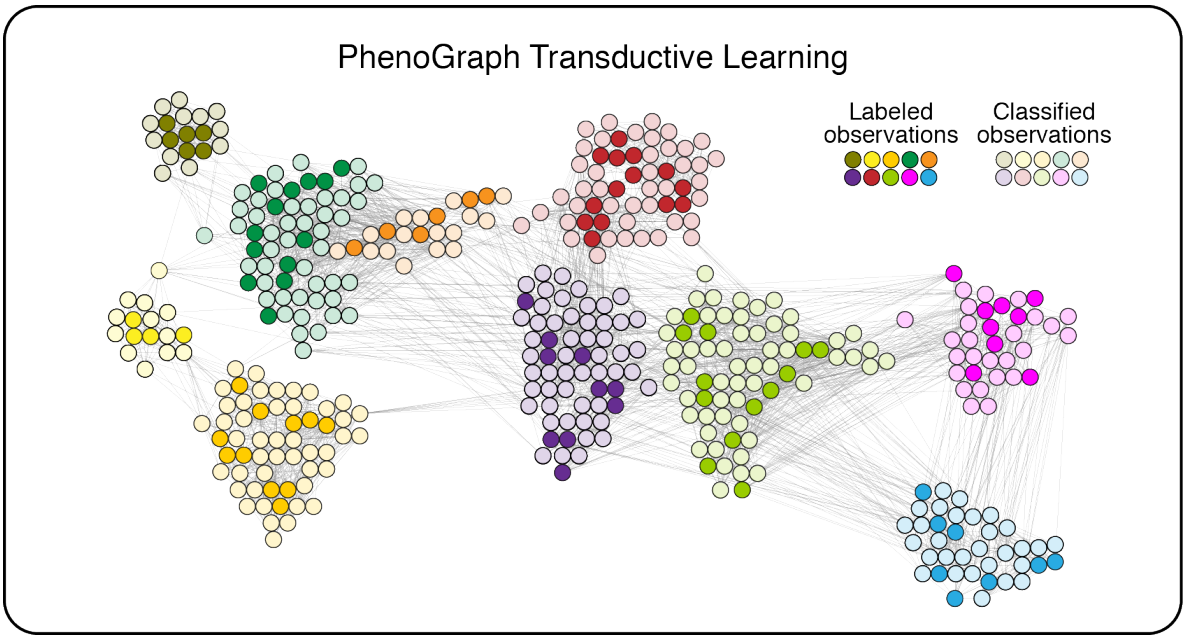


Figure 3.1: Schematic depiction of PhenoGraph Tranductive Learning. The graph is built without consideration of the class labels (bold colors) and each unlabeled vertex is classified (light colors) according to the class that is most likely reached first by a random walk in the graph.

the graph Laplacian is defined:

$$\mathcal{L} = D - \mathbf{W} \tag{3.1}$$

Note that each vertex is either labeled or unlabeled and therefore must be in one of two sets, V_L (labeled vertices) and V_U (unlabeled verticies), $V_L \cup V_U = V$ and $V_L \cap V_U = \emptyset$. Thus $\mathcal{L}$ can be arranged as a composition of submatrices corresponding to V_L and V_U as follows:

$$\mathcal{L} = \begin{bmatrix} \mathcal{L}_L & B \\ B^T & \mathcal{L}_U \end{bmatrix} \tag{3.2}$$

The graph Laplacian is used to compute the probability that random walks originating at verticies in V_U first arrive at particular classes in V_L. These probabilities can be calculated through the solution of a system of sparse linear equations (see [86] for complete derivation). Specifically,

$$\mathcal{L}_U P = -B^T Q \tag{3.3}$$

where Q is a $L \times M$ binary matrix representing the class of each vertex in V_L and P is a $(N - L) \times M$ matrix containing the desired probabilities for every vertex in V_U. In other words, each row j of P is a M-dimensional probability vector ($\sum_{m=1}^{M} P_{jm} = 1$) expressing the probability that a random walk departing from vertex j arrives first at a vertex in class m. Therefore, the matrix P is the quantity of interest, as it provides a probabilistic assignment of each point to each class (similar to the "responsibilities" of parametric mixture models). Every unlabeled data point can be assigned to the most probable class

$$y_j = \arg\max_m P_{jm} \tag{3.4}$$

or the assignments can be chosen more elaborately. For example, the entropy $H(P_j)$ provides a direct measure of the uncertainty of assigning j to any class; this can be used to rank and filter assignments to control the quality of the inference.

Psuedocode summarizing PhenoGraph Transductive Learning (PTL) is given in Algorithm 2.

Algorithm 2 PhenoGraph Transductive Learning (PTL)

IN: Single-cell measurements $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, partial labeling $\mathbf{y} = [y_1, \ldots, y_L]$, neighborhood size k
OUT: Completed labeling $\mathbf{y}' = [y_{L+1}, \ldots, y_N]$, matrix of assignment probabilities P
procedure GRAPH CONSTRUCTION($\mathbf{X}, k$)
  for all $i \in \{1, \ldots, N\}$ **do**
    for all $j \in \mathcal{V}_k(\mathbf{x}_i)$ **do**
      $\mathbf{W}_{ij} \leftarrow J_k(\mathbf{x}_i, \mathbf{x}_j)$
    end for
  end for
end procedure
procedure TRANSDUCTION($\mathbf{W}, \mathbf{y}$)
  Compute Eq. 3.1 and Eq. 3.2
  Solve Eq. 3.3 for P
  return P
end procedure
OPTION: $\mathbf{y}' \ni y_j \leftarrow \arg\max_m P_{jm}$

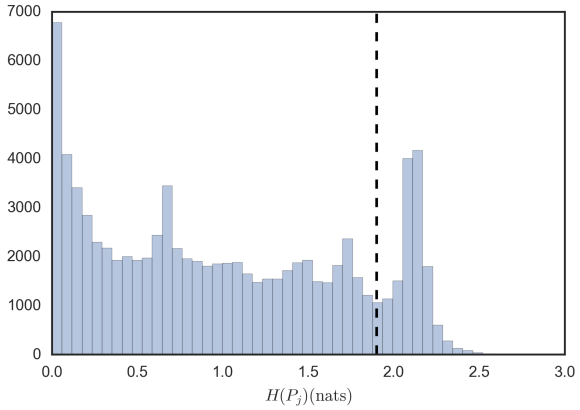


Figure 3.2: Entropy distribution of random walk probabilities for unlabeled cells. High values indicate classification uncertainty. The vertical dashed line indicates the filter used to eliminate cells with high classification uncertainty.

3.3 Using random walks to recover cells missed by manual gating

To demonstrate the capabilities of PTL, we revisit Validation Data Set 1 (§2.3), comprising $> 100,000$ healthy bone marrow cells with 13 surface marker dimensions and partial labels obtained by manual gating, which cover 49% of cells. PTL was run with $k = 30$ to obtain P and the entropy $H(P_j)$ was used to filter the classifications (Figure 3.2). This resulted in the recovery of 83% of the unlabeled data, increasing the total number of labeled cells by 86%. The number of recovered cells per cell type is shown in Table 3.1. For many cell types, the recovered cells outnumber the manually gated cells, reflecting not only the power of this approach but also the power of defining cell types in multiple dimensions simultaneously rather than sequentially.

For example, the number of common myeloid progenitors (CMPs) was doubled by algorithm. The marker distributions for the inferred CMPs are very similar to the manually-defined cells (Figure 3.3). The most notable difference is modestly elevated expression of CD4, though still at levels less than what would be considered "CD4$^+$" for a T cell. Since robust CD4 expression is a lineage marker for T helper cells, most gating strategies for progenitors include a "CD4$^-$" gate early in the sequence and these cells are lost as a result. The

	Manual Gating	Recovered
CD11b^{-} Monocyte	912	2134
CD11b^{high} Monocyte	6779	19036
CD11b^{mid} Monocyte	1278	1699
CMP	253	258
Erythroblast	12030	5327
GMP	73	1
HSC	261	96
Immature B	502	68
MEP	194	438
MPP	152	17
Mature CD38^{low} B	7796	3930
Mature CD38^{mid} B	608	482
Mature CD4^{+} T	13964	6083
Mature CD8^{+} T	7821	5171
Megakaryocyte	3684	786
Myelocyte	3025	3204
NK	3864	6355
Naive CD4^{+} T	6987	7054
Naive CD8^{+} T	9564	5010
Plasma cell	468	575
Plasmacytoid DC	293	478
Platelet	5	0
Pre-B I	240	105
Pre-B II	994	558

Table 3.1: Number of cells recovered from manual gating by PhenoGraph transduction, listed by cell type.

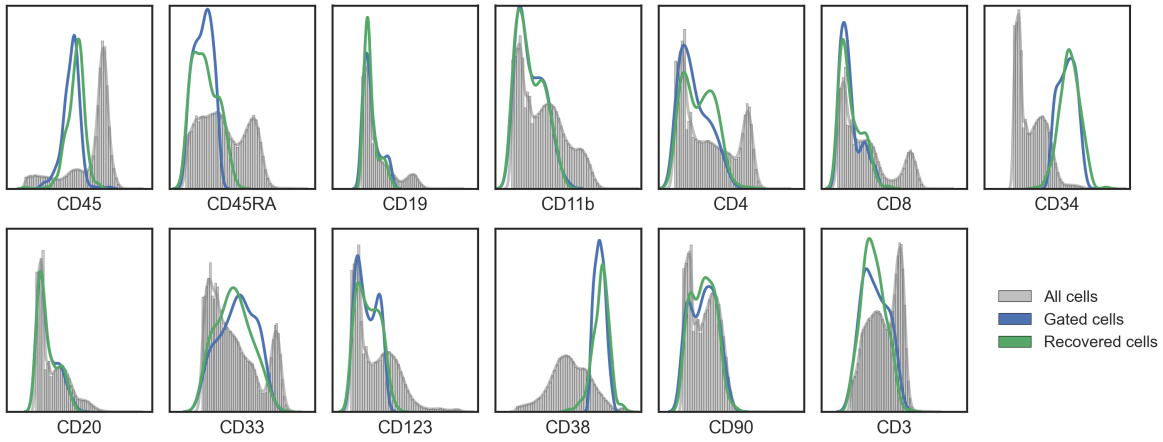


Figure 3.3: Gated (blue) and inferred (green) CMPs show similar marker distributions.

distribution of the other 12 markers, however, corroborates that these cells are CMPs.

Another interesting example are the NK cells, where the inferred number of cells more than doubles what was gated manually. Again, the marker distributions are very similar between the gated and the inferred NK cells with one notable exception: about half of the inferred NK cells express CD8 at a moderate intensity (i.e., less than $CD8^+$ T cells; Figure 3.4). In fact, it has been previously reported that about half of human NK cells do express CD8 at this moderate intensity [87]. While the manual gating strategy in [17, Fig. S5] defined NK cells as $CD8^-$, transductive learning by PhenoGraph recovered both $CD8^+$ and $CD8^-$ NK cells. Given that some of the inferred NK cells express moderate levels of CD3, it is worth noting that the $CD8^+$ inferred NK cells are indeed $CD3^-$ and are therefore not a side effect of misclassified cytotoxic T cells (Figure 3.5). The small number of $CD3^{mid}$ cells included among the inferred NK cells are $CD4^-/CD8^-$ and many express CD90 (data not shown); it is difficult to establish their identity with the limited markers available in this particular data set.

3.4 Summary

In this chapter, I introduced an extension of PhenoGraph that uses the same underlying graph structure to perform a different kind of inference task—namely, transductive learning.

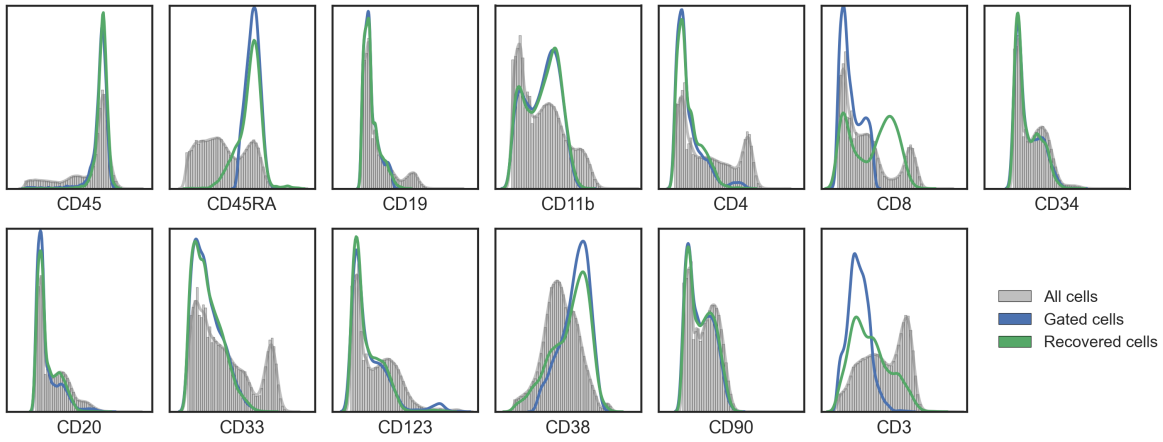


Figure 3.4: Gated (blue) and inferred (green) NK cells show similar marker distributions with the notable exception of CD8.

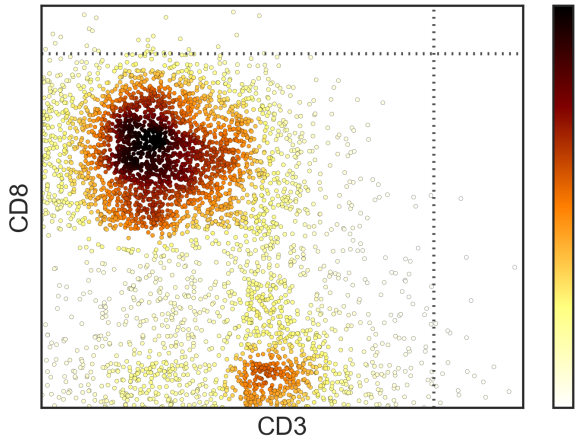


Figure 3.5: Scatter plot showing all inferred NK cells. The $CD8^+$ inferred NK cells are $CD3^-$. The dashed gray vertical and horizontal lines show average CD3 expression among all gated T cells and average CD8 expression among gated $CD8^+$ T cells, respectively. Color represents the local density at each point.

A simple performance evaluation using manually gated normal bone marrow suggests that the method successfully finds correct classifications for unlabeled data when labeled examples are available. In the next chapter, this tool will find further use as a means to characterize disease states.

Chapter 4

Data-driven phenotypic dissection of acute myeloid leukemia

A critical area of application for the data-driven methods discussed so far is human disease and especially cancer biology. Though tumors have sometimes been conceptualized as simple agglomerations of malignant cells, a parallel body of research treats cancer as a diseased tissue [88]. If tumor cells or more broadly the tumor microenvironment—comprising neoplastic cells, tumor infiltrating leukocytes, activated stromal cells—form a complex population with functional differentiation, then the structure of this population may be critical for understanding and controlling the disease.

In fact, research increasingly supports the proposition that intratumor heterogeneity is functionally and clinically significant [89]. Recent evidence implies that the pathobiology of cancer results from the actions and interactions of diverse subpopulations within the tumor. Because tumors defy the regulatory mechanisms that bring healthy tissues into such uniformity across individuals, it is difficult to know prospectively what the subpopulation structure of a tumor might be like. To uncover the unknown subpopulation structure of malignancies, it is therefore critical to obtain measurements that preserve single-cell resolution and capture as many simultaneous dimensions as possible. The high dimensionality can be exploited by data-driven techniques which, as shown in the previous chapters, can identify *ab initio* the

diversity of cellular states in the population.

4.1 Acute myeloid leukemia

Intratumor heterogeneity is pervasive in acute myeloid leukemia (AML), an aggressive liquid tumor of the bone marrow characterized by overwhelming abundance of poorly differentiated myeloid cells ("blasts"). Arising from the disruption of regulated myeloid differentiation [90], AML results in a disordered developmental hierarchy wherein leukemic stem cells (LSCs) are capable of re-establishing the disease in immunodeficient mice [37]. LSCs were first thought to display the same $CD34^+/CD38^-$ cellular phenotype as normal hematopoietic stem cells (HSCs). Subsequent studies demonstrated a disconnect between the surface phenotype and the functional state of LSCs, with both $CD38^+$ [38] and $CD34^-$ [39] LSCs having been reported. While almost all cases of AML do exhibit a differentiated hierarchy (i.e., LSCs are typically only a subset of blasts), no surface marker phenotype has been identified that consistently indicates LSCs across patients [43].

Recognizing this disconnect between functionally primitive cells (e.g., tumor-initiating, "stem-like") and their surface phenotypes, we designed experiments to simultaneously assay surface protein expression and regulatory signaling in millions of individual cells from primary AML samples. While the functional state of the cell is not directly measurable (§1.3.1), we reasoned that it may be more reliably inferred from intracellular signaling as opposed to membrane-bound protein expression. The simultaneous measurement of surface and intracellular features was enabled by the expanded dimensionality of mass cytometry (§4.2). Cells were measured both before and after several *ex vivo* molecular perturbations, which elicit intracellular responses that reflect the broader signaling network beyond what can be inferred from the unperturbed state [46]. Computational integration of these signaling responses further expanded the dimensionality of the data (§5.1.1).

4.2 High-dimensional single-cell profiling of an AML cohort

We used mass cytometry to obtain single-cell proteomic profiles of cryopreserved bone marrow aspirates from pediatric AML patients obtained at diagnosis ($n = 16$) and from healthy adult donors ($n = 5$). We performed preliminary analysis to select 16 highly informative surface markers that efficiently captured the intra- and intertumor heterogeneity in our cohort (§A.2.2). We added 14 antibody probes against intracellular phosphorylation, thus allowing simultaneous measurement of surface phenotype and signaling behavior in single cells. Each sample was subjected *ex vivo* to a battery of short-term molecular perturbations (cytokines and chemical inhibitors; see Appendix for details) to elicit functionally relevant signaling responses [17, 46]. The complete data set contained over 15 million single cells from 21 individuals measured in 31 simultaneous protein dimensions following exposure to one of 17 conditions (Table 4.1). A summary of the experimental design and first-stage analysis is provided in Figure 4.1.

4.3 PhenoGraph reveals a "landscape" of leukemic states

While healthy bone marrow is known to contain well-separated subpopulations (§2.3), it was questionable to what extent this would be true in the AML samples. We explored this question in previous work [68], where we found evidence that leukemias contain heterogeneous states that are more overlapping than what is observed in healthy marrow. These observations were corroborated by the additional 5 healthy and 16 leukemic samples investigated in the present study.

All healthy samples presented highly distinct and reproducible cell types, identifiable through their known marker combinations. Each leukemia also presented a diversity of phenotypes defined by distinct combinations of surface marker expression. For example, Figure 4.2 shows bone marrow cells from a single representative leukemia patient (SJ03) mapped into two dimensions by t-SNE. The resulting phenotypic landscape is diffuse compared to normal bone marrow, yet it is characterized by regions of distinct expression patterns.

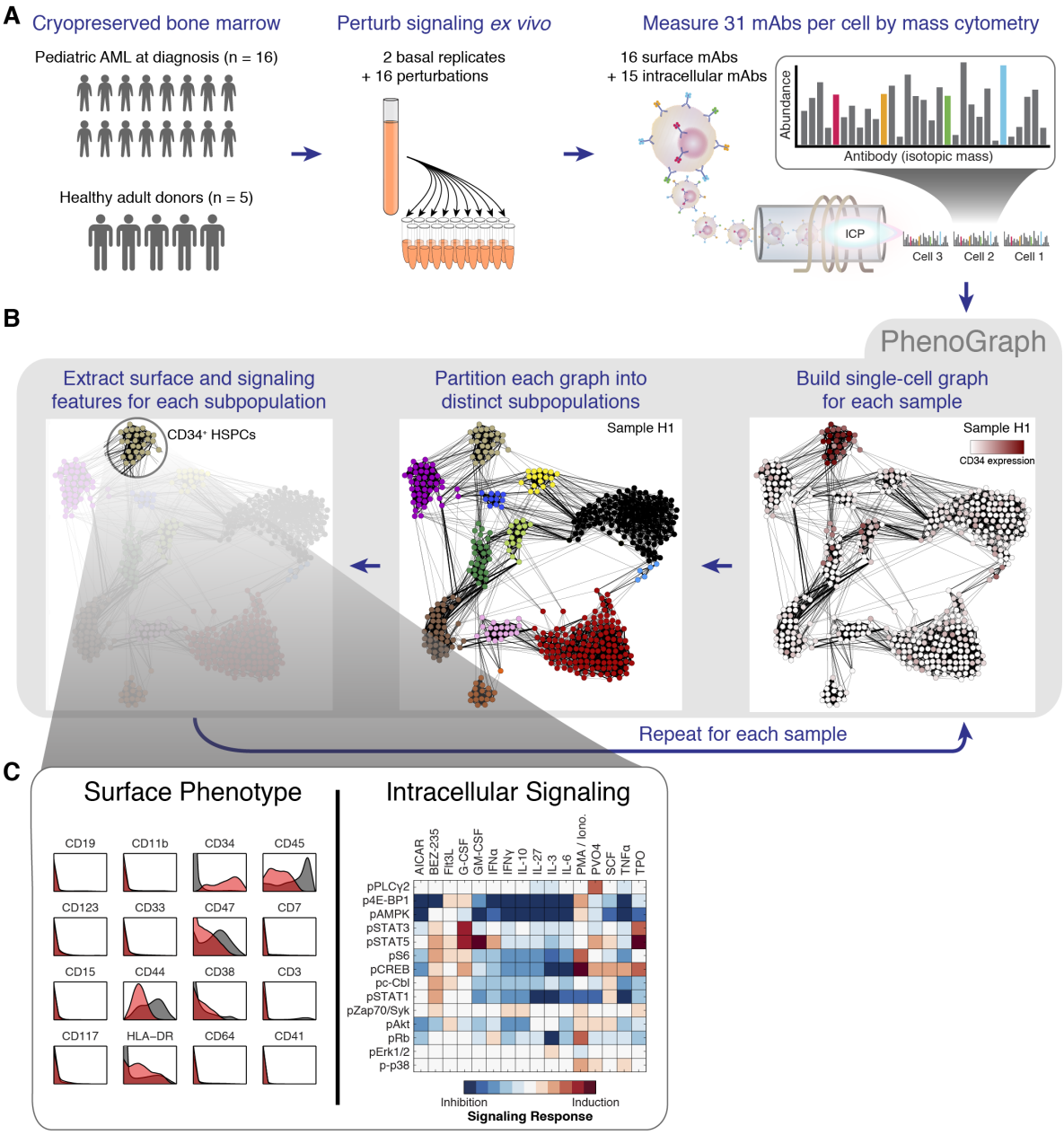


Figure 4.1: Profiling normal and malignant surface and signaling phenotypes by mass cytometry. (A) Mass cytometry profiling of an AML cohort with healthy controls. (B) PhenoGraph was used to identify the subpopulations in each sample individually. This panel shows the graph built from 500 randomly selected cells of healthy donor H1. Each vertex is colored by CD34 expression of its corresponding cell (right). $CD34^+$ hematopoietic stem and progenitor cells (HSPCs) form a densely interconnected subgraph and are assigned to a single subpopulation by community detection (middle). (C) Each subpopulation is a multifaceted data object containing surface marker distributions as well as perturbed intracellular signaling. Shown here, the HSPCs identified by PhenoGraph in donor H1 (red histograms) had a $CD34^+/CD45^{low}$ phenotype relative to the other cells in the sample (gray histograms). Each PhenoGraph subpopulation contained cells from all perturbations, permitting analysis of 224 signaling responses as shown in the heat map on the right.

Surface Markers	Signaling Markers	Molecular Perturbations
CD19	PLCγ2 (pY759)	Basal (i.e. PBS only)
CD11b	4EBP1 (pT37/46)	AICAR
CD34	AMPK (pT172)	BEZ-235
CD123	STAT3 (pY705)	Flt3L (Flt-3 ligand)
CD45	S6 (pS235/pS236)	G-CSF (granulocyte CSF)
CD33	CREB (pS133)	GM-CSF (granulocyte-monocyte CSF)
CD47	STAT5 (pY694)	IFNα (interferon α A/D)
CD7	c-CBL (pY700)	IFNγ (interferon γ)
CD15	STAT1 (pY701)	IL-10 (interleukin-10)
CD44	ZAP70/SYK (pY319/pY352)	IL-27 (interleukin-27)
CD38	AKT (pS473)	IL-3 (interleukin-3)
CD3	RB (pS807/pS811)	IL-6 (interleukin-6)
CD117	ERK1/2 (p44/42) (pT202/pY204)	PMA/Ionomycin
HLA-DR	P38 (pT180/ pY182)	PVO4 (pervanadate)
CD41		SCF (stem cell factor)
CD64		TNFα (tumor necrosis factor α)
		TPO (thrombopoietin)

Table 4.1: Markers and perturbations used for single-cell profiling. All signaling markers target phosphorylated epitopes at the amino acid residues specified in parentheses. CSF: colony stimulating factor; AICAR (5-Aminoimidazole-4-carboxamide 1-beta-D-ribofuranoside): pharmaceutical AMP-dependent protein kinase inhibitor; BEZ-235: pharmaceutical phosphoinositide 3-kinase inhibitor; PMA (phorbol 12-myristate 13-acetate): together with ionomycin induces cytokine production in many cell types. Further details can be found in Table A.1 below and in the supporting materials of [84].

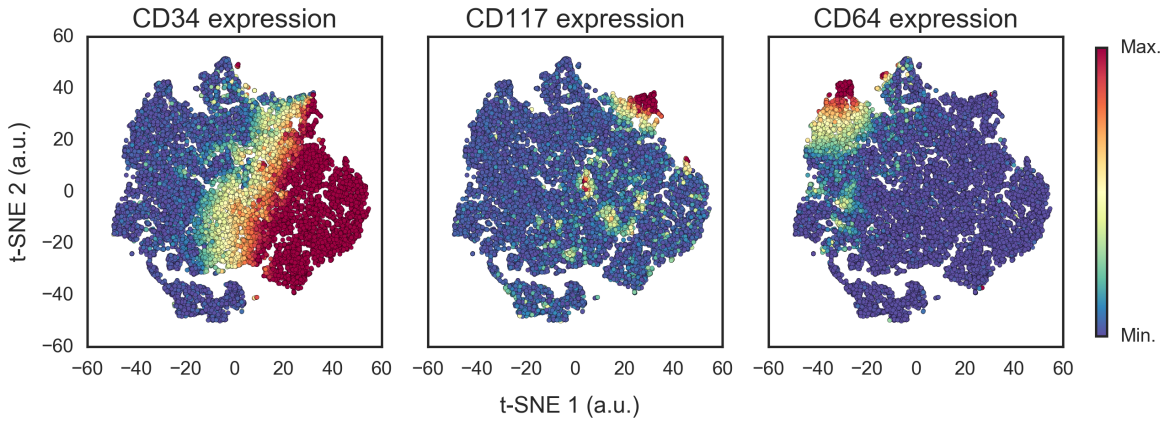


Figure 4.2: *t*-SNE map of bone marrow cells from patient SJ03. Each point represents a single cell, colored according to its measured expression of the indicated marker. All 16 available surface markers were used to generate the map; 3 are chosen as examples.

As mentioned previously (§2.1), reduction to two dimensions is required for visualization but is not necessarily optimal for identifying subpopulations in the original high-dimensional space. PhenoGraph, which circumvents the strong requirement of a two-dimensional projection, identified multiple subpopulations characterized by substantial phenotypic differences in the leukemia samples. Figure 4.3 shows clusters identified by PhenoGraph in sample SJ03 as colored labels on the *t*-SNE map of Figure 4.2. It is interesting to note that the PhenoGraph clusters accentuate subtle structures in the map that may not be obvious at first glance. A PhenoGraph cluster may comprise phenotypic "pockets" from disconnected regions of the map—such as Cluster 12 in Figure 4.3—that may reflect suboptimal splits as *t*-SNE searches for the highly compressed two-dimensional projection. In the case of Cluster 12, these two pockets are kept together by PhenoGraph because they are characterized by CD117 expression, as can be observed in Figure 4.2. An overview of subpopulations identified by PhenoGraph in patient SJ03 is given in heat map form in Figure 4.4. In addition to some non-myeloid cell types, there are clearly different leukemic phenotypes in this patient, defined particularly by distinct combinations of CD34, CD117, HLA-DR, and CD64.

To analyze the full cohort, PhenoGraph was run on each sample individually, using each cell's 16-dimensional surface marker profile to define phenotypes. This yielded an average of 28 subpopulations per sample (ranging between 17 and 48), totaling 616 subpopulations

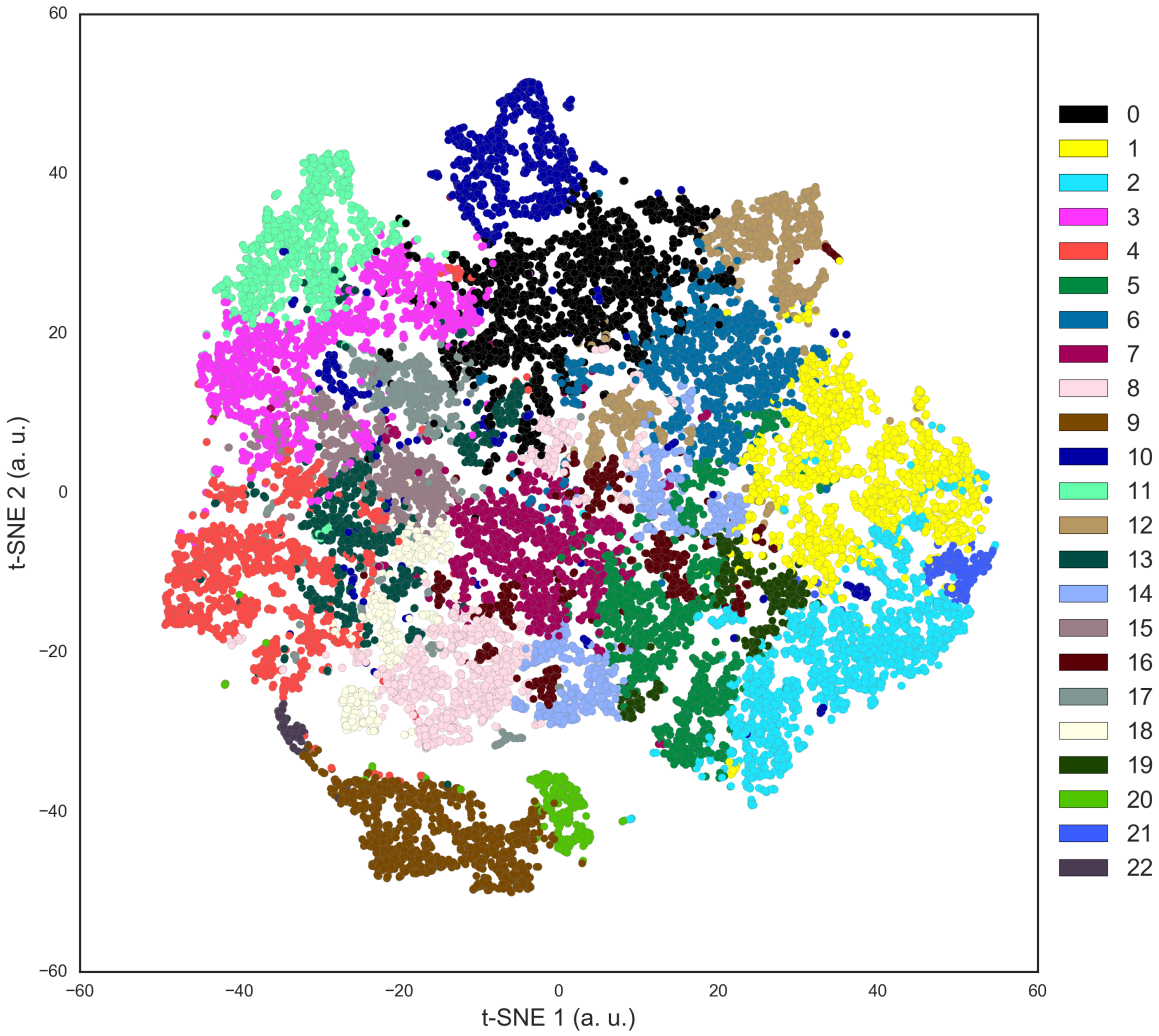


Figure 4.3: Clusters found by PhenoGraph in patient SJ03, displayed as colored labels on the *t*-SNE map shown in Figure 4.2. The clusters split the map into coherent structures that are not always immediately obvious upon inspection of the two-dimensional map.

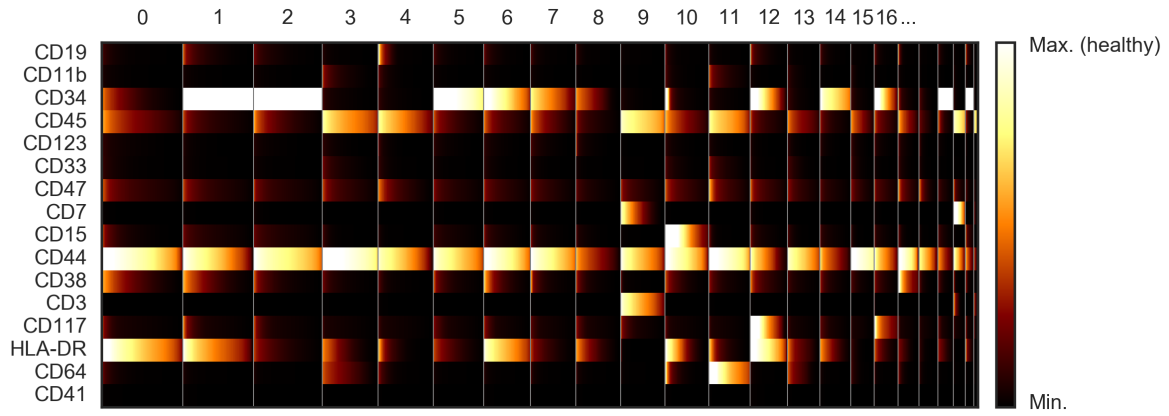


Figure 4.4: Clusters found by PhenoGraph in patient SJ03, displayed as a heat map. In this visual, clusters are separated by gray vertical lines and they are ordered from left to right by decreasing cell count. Expression values are sorted within each row and within each cluster to facilitate visualization of the intra-cluster marker distributions. Expression intensities have been rescaled as a ratio of the maximum "normal" expression intensity, determined from the 5 healthy samples.

across the entire cohort. Subpopulation size varied by orders of magnitude, from 7×10^2 to 2×10^5 cells (or .06% to 20% of a sample). For each sample, data were pooled from all perturbations before clustering, enabling characterization of subpopulation-specific signaling patterns (§5.1.1). Each resulting subpopulation was a multifaceted data object, containing information about surface phenotypes, as well as the response of each signaling marker to each molecular perturbation (Figure 4.1).

4.3.1 Patterns of intra- and intertumor heterogeneity

Each leukemia presented a diversity of surface phenotypes that seemed to share some similarities and differences across patients. We therefore sought an integrative overview that could enable direct comparison of all subpopulations simultaneously and reveal larger trends in the cohort. Toward this end, we used t-SNE, which has been so successful in revealing the "phenotypic landscape" of single cells. The difference here, of course, is that the objects being mapped are not single cells but clusters of cells. To generate a map for the entire cohort, we obtained subpopulation phenotypes by taking the 16-dimensional centroid of the surface markers of each subpopulation. The landscape generated from the subpopulation

phenotypes provided an intuitive and comprehensive overview of the major phenotypic trends of the cohort and also revealed the extent of intra- and inter-tumoral heterogeneity (Figure 4.5). Subpopulations from healthy and leukemic samples were mapped simultaneously so the healthy cell types could act as "landmarks" to aid interpretation of the leukemic subpopulations. Normal lymphoid cell types were excluded from the landscape to focus on primitive and myeloid phenotypes, "zooming in" on the myeloid lineages relevant to AML.

The AML cohort landscape organized the subpopulations into regions of phenotypic similarity, distinguished by particular marker combinations. Inspecting the structure of this landscape, we found that the vertical axis largely mimicked trends in normal myeloid development with primitive markers expressed toward the top and more mature markers toward the bottom. Healthy $CD34^{+}/CD38^{mid}$ hematopoietic stem and progenitor cells (HSPCs) provided the most primitive landmark, located at the top of the landscape. AML subpopulations in this region displayed surface profiles that resembled the HSPC phenotype. At the bottom of the landscape, the $CD11b^{+}$ healthy monocytes served as a landmark for differentiated myeloid cells, representing full maturation not observed in the leukemic samples. Between these two poles, other developing myeloid antigens—CD38, CD117, CD123, CD33—peaked and subsided, thus the vertical axis of the landscape resembled normal myeloid development (Figure 4.6). The adherence of AML phenotypes to this axis suggests that myeloid developmental programs continue to influence the phenotypic diversity of leukemic cells even after malignant transformation. The patterns of intratumor heterogeneity support this view, as most patients contained a mixture of "primitive" and "mature" surface phenotypes (Figure 4.7).

4.3.2 Metaclusters highlight inter-patient similarity

Despite the widespread phenotypic diversity observed within patients, the cohort landscape revealed a surprising conformity when comparing AML subpopulations across different patients. Multiple patients occupied each region of the landscape and no patient presented a substantially unique phenotype, suggesting that subpopulations could be matched across patients,

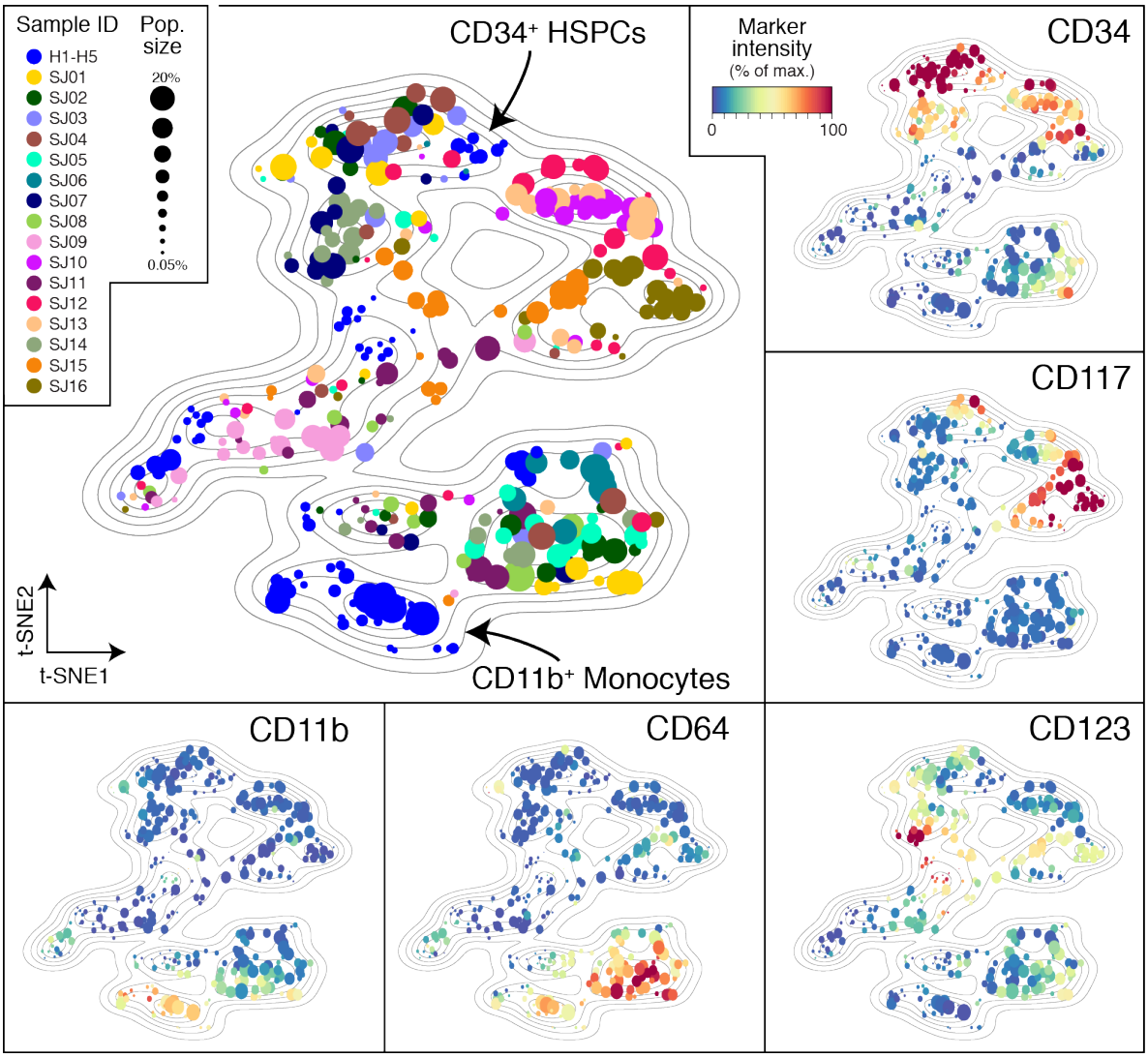


Figure 4.5: Landscape of subpopulation phenotypes generated by *t*-SNE. Each subpopulation is represented by a single point, scaled to represent its sample proportion and colored by patient identity (main panel). Normal bone marrow cell types (H1–H5; blue) provide landmarks for interpreting the phenotypes of the leukemic bone marrow samples (SJ01-SJ16). In the additional panels each subpopulation is colored by its median expression of the indicated surface marker.

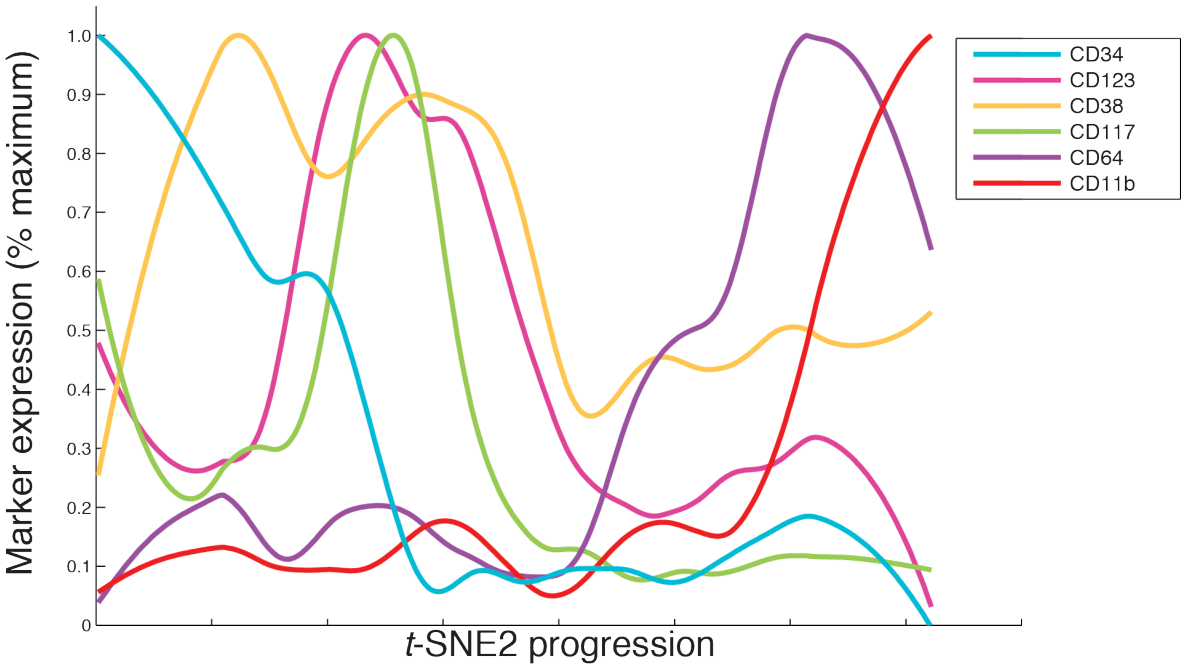


Figure 4.6: Gaussian weighted smoothing of subpopulation marker intensity displayed as a function of the second (i.e., vertical) dimension of the *t*-SNE map. The axis mimics myeloid development from left to right, reflected in the fall of CD34 expression, the transient rise of expression of CD38, CD117 and CD123, and finally the rise of CD64 and CD11b expression.

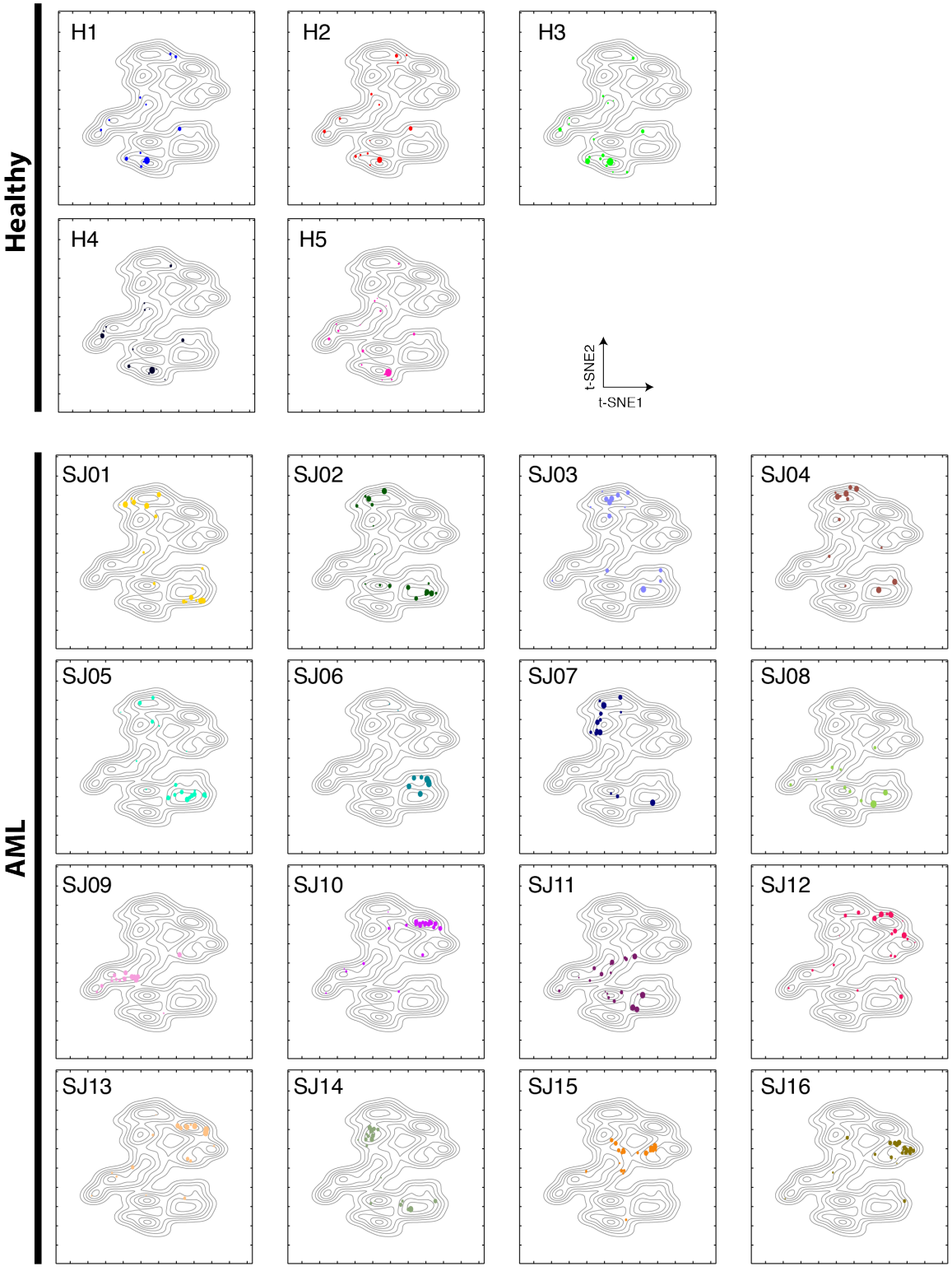


Figure 4.7: Subpopulations of each patient visualized separately in the landscape of Figure 4.5. Many patients are split between the top and bottom half of the landscape, which corresponds to primitive and mature-like surface phenotypes (Figure 4.6), suggestive of persisting developmental processes in most leukemic samples.

cohort-wide. To examine these cohort-level phenotypes further, we pursued a metaclustering approach in which subpopulations from each patient were grouped by a secondary clustering analysis [73]. Defining subpopulation phenotypes as before, we used PhenoGraph to group the subpopulations into metaclusters (MCs; Figure 4.8 A). Specifically, all 425 AML-derived subpopulations were analyzed together, leaving out the healthy-derived subpopulations to avoid bias toward normal phenotypes. PhenoGraph was run with the parameter $k = 15$, reflecting the smaller sample size compared to the initial single-cell data.

With 16 patients, the cohort is relatively small and we therefore conducted robustness analysis to determine whether the MCs were biased by the inclusion of any particular patient(s). The stability of the metacluster results to perturbations of the cohort was assessed by subsampling the patients and recomputing the metacluster assignments. Specifically, we produced 16 leave-one-out and 120 leave-two-out data sets in which all subpopulations from patient $\{i \mid 1 \geq i \geq 16\}$ (and $\{j \mid 1 \geq j \geq 16 \wedge j \neq i\}$ in the case of leave-two-out) were removed. Metacluster assignments were computed for each subsample as described in the previous paragraph ($k = 15$). Each subsample metacluster result was compared to the full-data MC assignments of the same populations (those retained in the subsample), using normalized mutual information (NMI; Eq. 2.7) to quantify similarity of the assignments. For reference, NMI was also computed for 16 random restarts of PhenoGraph using the full cohort data. While each random restart produced nearly identical results (mean NMI = 0.94), the reproducibility of the metacluster assignments was only modestly diminished by the leave-one-out and leave-two-out tests, both obtaining average NMI scores of 0.9 with very little variance (Figure 4.8 B).

The full-data metaclustering solution identified 14 MCs that represent the major cohort-wide phenotypes. Each MC had a mixed patient composition, containing subpopulations from at least 2 patients and a median of 11 patients (Figure 4.9A). The average surface marker patterns of some MCs resembled normal cell types and to varying extents (Figure 4.9B). For example, the $CD19^+$/HLA-DR^+ phenotype of MC12 matched the phenotype of mature B cells, and indeed AML patients are expected to have normal B cells. Other MCs

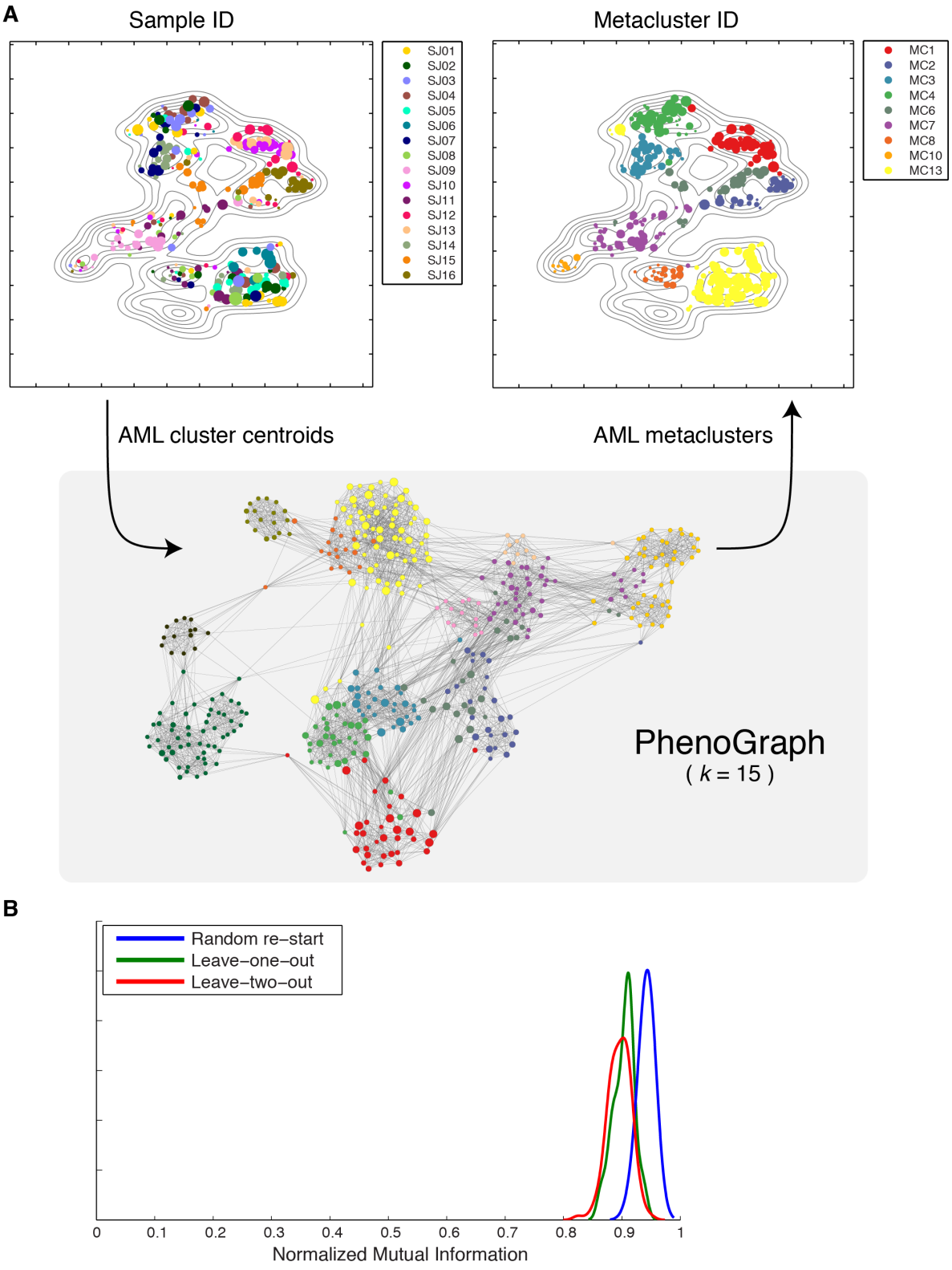


Figure 4.8: Metaclustering defines cohort-wide AML phenotypes. (A) PhenoGraph metaclusters split the AML landscape into major phenotypes, each containing subpopulations from multiple patients. (B) Reproducibility of PhenoGraph metaclusters as assessed by cross-validation.

had marker profiles that resembled progenitors in different ways. To more formally evaluate the relationship between MCs and healthy cell types, cells from the healthy samples (H1–H5) were systematically matched to MC marker profiles using linear discriminant analysis [91]. In this setting, each MC is given a multivariate normal density in 16-dimensional space with a shared covariance matrix, using the maximum likelihood estimates from the AML data. We can formally evaluate the posterior probability that each healthy cell and each MC are generated by the same source:

$$P\big(\mathrm{MC} = k \mid \mathbf{x} = \mathbf{x}_i\big) = \frac{f_k(\mathbf{x}_i)\pi_k}{\sum_{\ell=1}^{L} f_\ell(\mathbf{x}_i)\pi_\ell} \tag{4.1}$$

$$f_k(\mathbf{x}) = \mathcal{N}\big(\mathbf{x} \mid \mu_k, \Sigma_k\big) \tag{4.2}$$

for each healthy cell $\mathbf{x}_i$, where $\mathcal{N}(\cdot)$ is the multivariate normal density, μ_k is the mean surface marker expression for subpopulations in MC k, $\Sigma_k = \Sigma\,\forall k$ is the shared covariance matrix, and $\pi_k = 1/14\,\forall k$ is a uniform prior over the 14 MCs.

The posterior probability (Eq. 4.1) was evaluated for each cell in the healthy samples ($\sim 3 \times 10^6$ cells) using the `classify` function implemented in the MATLAB R2013b Statistics Toolbox. A high posterior probability for MC k indicates that cell $\mathbf{x}_i$ falls within the phenotypic boundaries estimated from the observed expression of subpopulations in MC k. Cells that were an extremely good fit[1] to one MC were considered *healthy cognate* cells for that MC.

Using this technique, 71–81% of each healthy sample was sufficiently similar to one MC to be identified as a healthy cognate. Approximately 40% of each healthy sample was assigned to MC14, consistent with the phenotype and frequency of T cells in normal bone marrow aspirates [92]. The proportion of each sample (healthy and AML) assigned to each MC was used to define a score that distinguished healthy from leukemic phenotypes. Specifically, for N cells from healthy donors and M cells from AML patients, with $\delta(c_i = k) = 1$ when cell i

[1] i.e., $P(\mathrm{MC} = k \mid \mathbf{x} = \mathbf{x}_i) > 0.99$

is associated with MC k, the score

$$s_k = \frac{\frac{1}{N}\sum_{i=1}^{N}\delta(c_i = k)}{\frac{1}{M}\sum_{j=1}^{J}\delta(c_j = k)} \tag{4.3}$$

quantifies whether cells associated with MC k were more frequently observed in healthy samples than in leukemic samples. MCs for which $s_k > 1$ were more abundant in healthy samples relative to leukemic samples and these included MC 5, 8, 9, 10, 11, 12, and 14. Examination of the expression patterns in these MCs revealed interpretable normal cell types: Immature B cells (MC5), myeloid dendritic cells (MC8), erythroblasts (MC9), granulocytes (MC10), NK cells (MC11), mature B cells (MC12), T cells (MC14). While large numbers of healthy cells were assigned to MC13 ($\sim$ 14% of normal marrow cells), these were outnumbered substantially by counts of MC13 cells in the leukemic samples. Given the monocytic phenotype of MC13, this is consistent with the histopathology of AML.

For the remaining MCs (1–4, 6, 7, 13), the score $s_k < 1$ identified phenotypes that were overrepresented in AML, indicating malignant expansions. Intriguingly, rare healthy cognate cells were identified in each normal marrow for each of these MCs, suggesting that leukemic phenotypes do not depart radically from cells that occur in normal hematopoiesis. The malignant MCs displayed phenotypes that resembled primitive and progenitor phenotypes with a myeloid bias. While each malignant phenotype comprised multiple patients, only MC13 contained subpopulations from every patient. Occupancy in MC13 varied substantially between patients (0.8%–77%), consistent with a model of AML as a block in myeloid differentiation with variable severity [90].

Samples were evaluated quantitatively in terms of their proportional occupancies of the 14 MCs (Figure 4.9 C). As expected, the 5 healthy samples were similar to each other and distinct from AML. Interestingly, MC occupancies organized the AML samples into subgroups that were significantly correlated with other molecular biomarkers. For example, patients with core binding factor translocation [t(8;21) or inv(16)] had large numbers of cells in MC4 and MC13, placing them in a group enriched for this clinical annotation ($P = 0.0014$, hypergeometric test). Patients with nucleophosmin mutations displayed a different phenotypic distribution—

occupancy of MC2, MC7 and MC13—forming another distinct patient group ($P = 0.0083$). Finally, the 3 patients characterized by large occupancies of MC1 were all cytogenetically normal ($P = 0.018$). Taken together, each leukemia, although unique, appears to be formed from a limited palette of possible phenotypes. Remarkably, the specific composition and relative proportion of MCs was determined in part by genetic background, demonstrating a genetic influence on the distribution of phenotypes observed in each patient.

4.4 Discussion

Tissues are complex populations of cells residing in phenotypically and functionally diverse states. A key challenge is to dissect the high-dimensional structure of these complex populations into components that can be studied individually and collectively. In AML, where the relationship between phenotypic and functional heterogeneity has been elusive, we used mass cytometry to profile both surface and signaling features simultaneously in millions of leukemic cells.

Using graphs of cellular phenotypes, PhenoGraph dissected each leukemic marrow into discrete subpopulations that displayed distinctive immunophenotypic as well as functional profiles. Representing each sample as a composition of subpopulations enabled a comprehensive view of the phenotypic landscape of the entire cohort. The landscape resembled normal myeloid development, but with aberrations resulting from malignant accumulation of cells and neoplastic divergence from normal phenotypes. Surprisingly, the landscape of AML immunophenotypes was restricted to a limited variety of expression patterns. These patterns occurred across different AML genetic subtypes, yet genetics had a detectable influence on the phenotypic composition of each patient. These observations suggest the persistence of developmental mechanisms that control the available repertoire of phenotypes even in the context of genetic dysregulation associated with cancer.

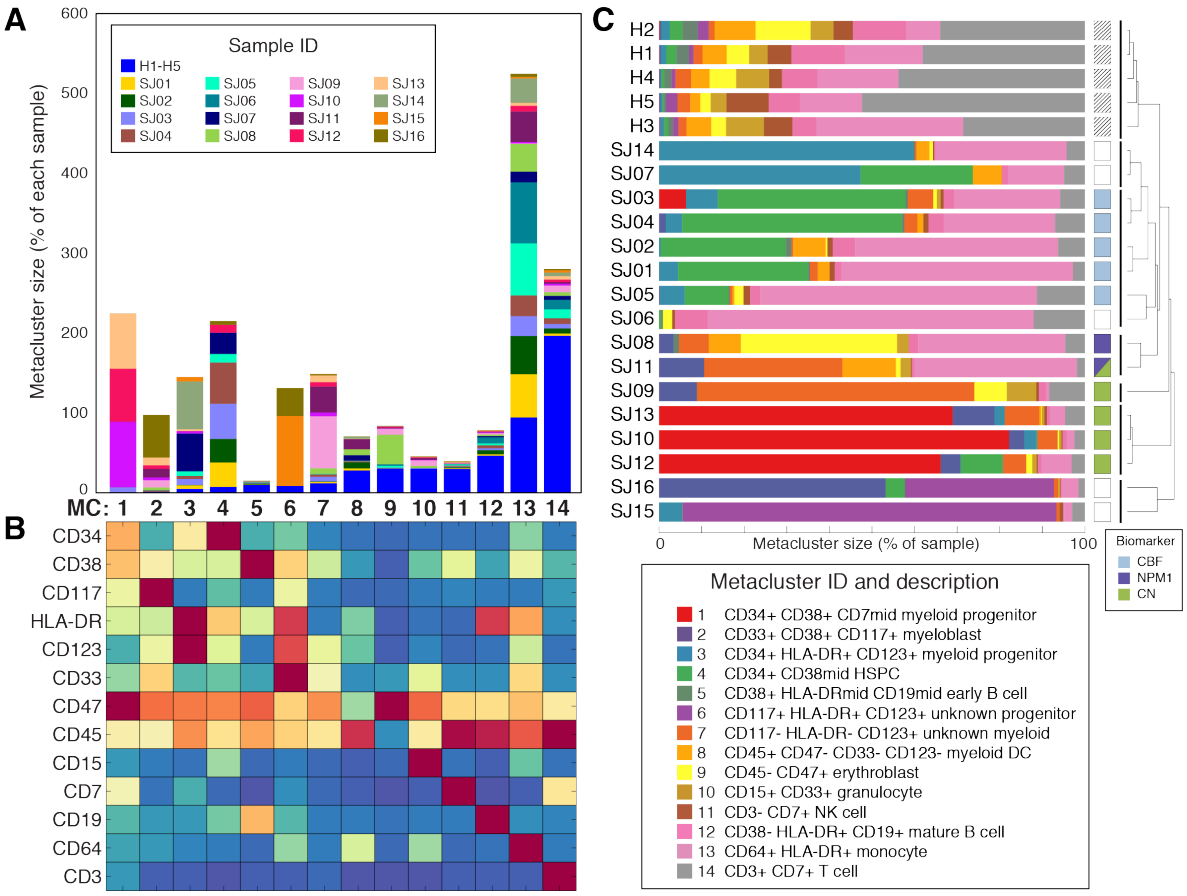


Figure 4.9: Metacluster (MC) analysis of the cohort's phenotypic composition. (A) The stacked bars indicate the contribution made by each patient to each MC. The blue segments represent the proportion of healthy samples assigned to each MC by linear discriminant analysis, as described in the main text. (B) Average surface marker expression of each MC. Colors represent expression intensity as in Figure 4.5 and elsewhere. Columns are matched with the stacked bars of (A). (C) Intrapatient heterogeneity is represented by one horizontal bar for each sample, in which segment lengths reflect the sample's proportional composition. Segment colors correspond to the MC descriptions in the legend below. Square boxes on the far right indicate molecular biomarkers that were significantly correlated with patient composition.

Chapter 5

Data-driven functional profiling of leukemic subpopulations

5.1 Signaling phenotypes reflect subpopulation function

Surface markers have become standard tools for clinical diagnosis and monitoring of blood neoplasia [93]. In normal bone marrow, cell surface markers identify stem and progenitor cell populations with distinct lineage potential and intracellular signaling behaviors [17]. However, in AML, no surface marker phenotype has been established that consistently distinguishes the more primitive blasts universally across patients (§1.2.2; [38, 39, 43]).

We hypothesized that intracellular signaling might be a better surrogate of the underlying functional potential and therefore the mass cytometry panel included 14 intracellular signaling markers, selected to represent pathways known to be functionally and clinically relevant in AML—including JAK/STAT, PI3K/AKT and MAPK. Following previous work in AML [46, 47], we supposed that not only the basal (unperturbed) activation of these pathways matters, but also their *potentiation* upon exposure to biologically relevant stimuli. Cells were therefore collected under one of 16 short-term molecular perturbations (Table 4.1), which—at 15 minutes—trigger intracellular signaling cascades without causing significant changes to surface marker expression.

Because these perturbations did not alter surface marker expression and because only surface markers were used to compute the PhenoGraph subpopulations, each subpopulation contained cells from all perturbations. It was therefore possible to compute the effect of each perturbation on each intracellular signaling molecule within each individual subpopulation. With 14 intracellular markers and 16 perturbations, this allowed the computation of 224 signaling responses per subpopulation (described below, §5.1.1).

Each of these 224 signaling responses reveals a different facet of the underlying network that controls cellular function. For each subpopulation, a concatenation of these responses can be interpreted as a quantitative, high-dimensional **signaling phenotype** that represents functional state. This phenotype is necessarily computed at the subpopulation level, meaning it reflects the behavior of cells that display similar surface marker expression in a given patient. However, the signaling phenotype itself contains only information about perturbation response and therefore provides a measure of subpopulation function that is independent of surface marker expression. We were therefore able to use the surface and signaling phenotypes as two alternative characterizations of each subpopulation, enabling direct comparison of these two kinds of phenotype.

5.1.1 Computing signaling phenotypes from molecular perturbation data

It has been shown in diverse biological systems that cellular response to environmental cues is a stochastic process and that population-level changes induced by stimulation are often mixtures of discrete single-cell responses [3, 94, 95]. In such cases, a shift in the population average is secondary to a change in the underlying distribution of cellular and molecular states. When data are available that record the states of individual cells, methods that compare distributions rather than point estimates will be more sensitive and more accurate.

A simple approach to quantify signaling response would be to subtract the average intensity of a phosphoprotein under stimulation from its average in the basal state. However, this approach has key shortcomings. First, averaging collapses the rich, single-cell data into a single point estimate and discards any variation in the response. For example, it is often

observed that surface markers enrich but do not purify functionally relevant subpopulations. In such cases, a functionally important response may change the shape of a distribution while having a minimal effect on the average. Another limitation of average difference is that it provides no measure of significance. Sample sizes inevitably vary and influence the reliability of signaling response estimates—e.g., small samples can easily exaggerate the magnitude of a response by random fluctuation. To address these concerns, we developed SARA (Statistical Analysis of Response Amplitude), represented schematically in Figure 5.1.

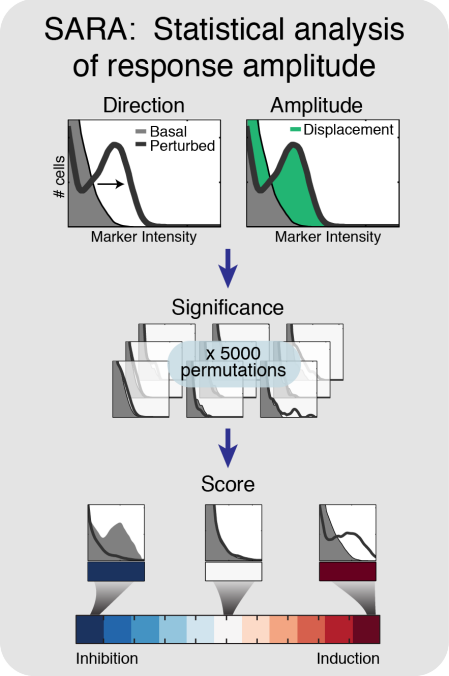


Figure 5.1: Schematic depiction of SARA.

SARA examines the entire single-cell distribution of phosphoprotein intensities to detect meaningful changes between two conditions. Each phosphomarker ϕ is treated as a random variable whose distribution depends on two other observed variables: cluster membership $\mathcal{C}$ and environmental condition $\mathcal{Z}$. We are interested in comparing two distributions:

$$P_b(\phi \mid \mathcal{Z} = b, \mathcal{C} = c) \tag{5.1}$$

$$P_s(\phi \mid \mathcal{Z} = s, \mathcal{C} = c) \tag{5.2}$$

where b denotes measurement in the unstimulated ("basal") condition and s denotes measurement under a stimulated condition. In all cases, we compare basal and stimulated distributions within the same cluster. The quantity of interest for comparing the two distributions is the cost converting one to the other, known equivalently as the Earth Mover or Mallow's distance [96] and for the one-dimensional case is the L_1 norm between the empirical cumulative distribution functions F_b and F_s:

$$\text{EMD} = \sum_{\Phi} \left| F_b(\phi) - F_s(\phi) \right| \tag{5.3}$$

where Φ is a fine grid over the support of $P(\phi)$.

Because the quantities $F(\phi)$ are empirical distributions, they are vulnerable to sampling error. Therefore, a measure of statistical significance for the EMD is introduced, based on permutations. A null distribution is built by computing EMD for a large number (5000 in practice) of permutations of $\mathcal{Z}$. The null distribution captures the differences between P_b and P_s due to sampling imbalances between the basal and stimulated conditions, taking into account the shape of $P(\phi)$. The p-value is computed as the proportion of the null distribution greater than or equal to the true EMD. Rather than impose a significance threshold, the p-value is integrated into the final SARA score as an inverse weight that dampens the magnitude of less significant responses.

Finally, while EMD is strictly nonnegative, stimulation response should be a signed quantity reflecting induction or inhibition with respect to the basal condition. The sign is incorporated by comparing the centers of P_b and P_s. Random noise may cause spurious changes in sign when the stimulation has an insignificant effect, in which case the significance penalty will cause these values to be distributed near 0. There are cases for which incorporation of direction is not appropriate (for example, a significant, evenly diverging response), but this case was never observed in practice.

The final score given by SARA is

$$\text{score} = \text{sgn}\left(\mathbb{E}[\phi_s] - \mathbb{E}[\phi_b]\right) \cdot (1 - \text{p}) \cdot \text{EMD} \tag{5.4}$$

where $\mathbb{E}[\cdot]$ is the expected value of a random variable.

To facilitate comparability across samples and signaling phenotypes, SARA scores were converted to z-scores. The dynamic range of SARA scores varied substantially between conditions. For example, the chemical perturbation pervanadate produced much more dramatic responses than biological stimulations such as IL-3. Additionally, subtle sample-specific biases in these dynamic ranges were noted before normalization, likely due to inevitable differences in handling of primary human samples from day to day. Therefore, the SARA scores were standardized within each sample and condition. Thus, each value in the signaling phenotype

represents the relative magnitude of the response within the context of the given sample and condition. Supporting the use of standardization, we expect that most conditions do not affect every phospho-marker in every subpopulation; and indeed, SARA scores induced by each condition had a single peak near zero. Thus, the use of z-scores enhances interpretability by aligning the mean response with zero and highlighting the most significant responses

In summary, SARA was used to define signaling phenotypes that, by hypothesis, could be treated as functional analogues to surface marker subpopulation phenotypes. SARA produced a quantitative signaling response for each phosphoprotein marker under each of the 16 stimulation conditions, resulting in a 224-dimensional signaling phenotype for each subpopulation. Together, PhenoGraph and SARA distilled high-dimensional data for 15 million cells into a single matrix of subpopulations and their signaling phenotypes, revealing a rich variety of signaling potential across subpopulations (Figure 5.2).

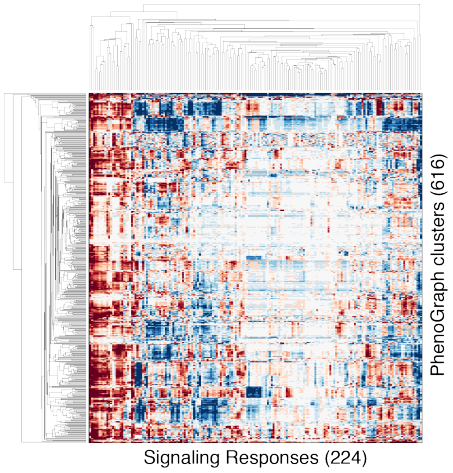


Figure 5.2: SARA generated $14 \times 16 = 224$ signaling phenotypes for each subpopulation across the cohort.

5.1.2 Signaling phenotypes are decoupled from surface markers in leukemia

Within the healthy samples, surface and signaling phenotypes were tightly coupled, consistent with previous reports [17, 97]. Hierarchical clustering of a curated set of progenitor- and lineage-associated signaling features produced a complete separation of primitive ($CD34^+$) and mature ($CD34^-$) cell types among the healthy samples (Figure 5.3; $P = 2.0 \times 10^{-52}$, Student's t test). In the leukemic samples, the same procedure produced a similar stratification of signaling phenotypes, including a set of subpopulations that recapitulate the signaling profile of healthy primitive cells. However, this stratification of primitive (PS) and mature (MS) signaling had no association with CD34 expression ($P = 0.83$, Student's t test; Fig. 4D). Decoupling of surface and signaling phenotypes in the leukemic samples is consistent with

evidence that surface markers are unreliable proxies of cellular function in AML [38, 39, 43, 97]. We therefore sought to use signaling phenotypes rather than surface phenotypes as alternative proxies for functional state.

5.2 Transductive inference of leukemic maturity

PhenoGraph and SARA yielded two alternative representations for each subpopulation: a 16-dimensional surface phenotype, and a 224-dimensional signaling phenotype (Figure 5.4). We asked if there was a characteristic signaling phenotype of undifferentiated healthy cells that could act as a high-dimensional generalization of the CD34/CD38 surface phenotype, which more faithfully captures the functional aspect of the primitive state.

Harnessing the tight coupling between surface and signaling in the healthy system, we grounded our analysis in a characterization of healthy subpopulations. To explicitly define the healthy cell types in the data, we used PhenoGraph to metacluster the surface phenotypes of the 191 subpopulations from the five healthy samples, analogously to way AML MCs were defined, as described above (§4.3.2). The analysis produced 20 healthy metaclusters (HMCs), which generally displayed recognizable surface marker phenotypes corresponding to known cell types, such as monocytes (HMC1) and HSPCs (HMC9). The explicit cell type assignments enabled identification of signaling responses that were significantly associated with cell type. Specifically, the HMC assignments formed a categorical variable that could be used to stratify each signaling response in order to assess significance by analysis of variance (ANOVA). A large number of signaling responses were strongly associated with cell type. Many of these were induction responses specific to undifferentated cells, including G-CSF→pSTAT3 ($Q = 6.4 \times 10^{-42}$) and SCF→pAKT ($Q = 1.0 \times 10^{-9}$), as previously reported [97]. The 25 most significant type-associated signaling responses are given in Table 5.1. These responses together with the surface phenotypes of the same subpopulations are shown for 4 selected cell types in Figure 5.5.

We then asked whether signaling responses were entirely sufficient to distinguish healthy cell types, rendering the surface phenotypes dispensable for characterizing the subpopulations.

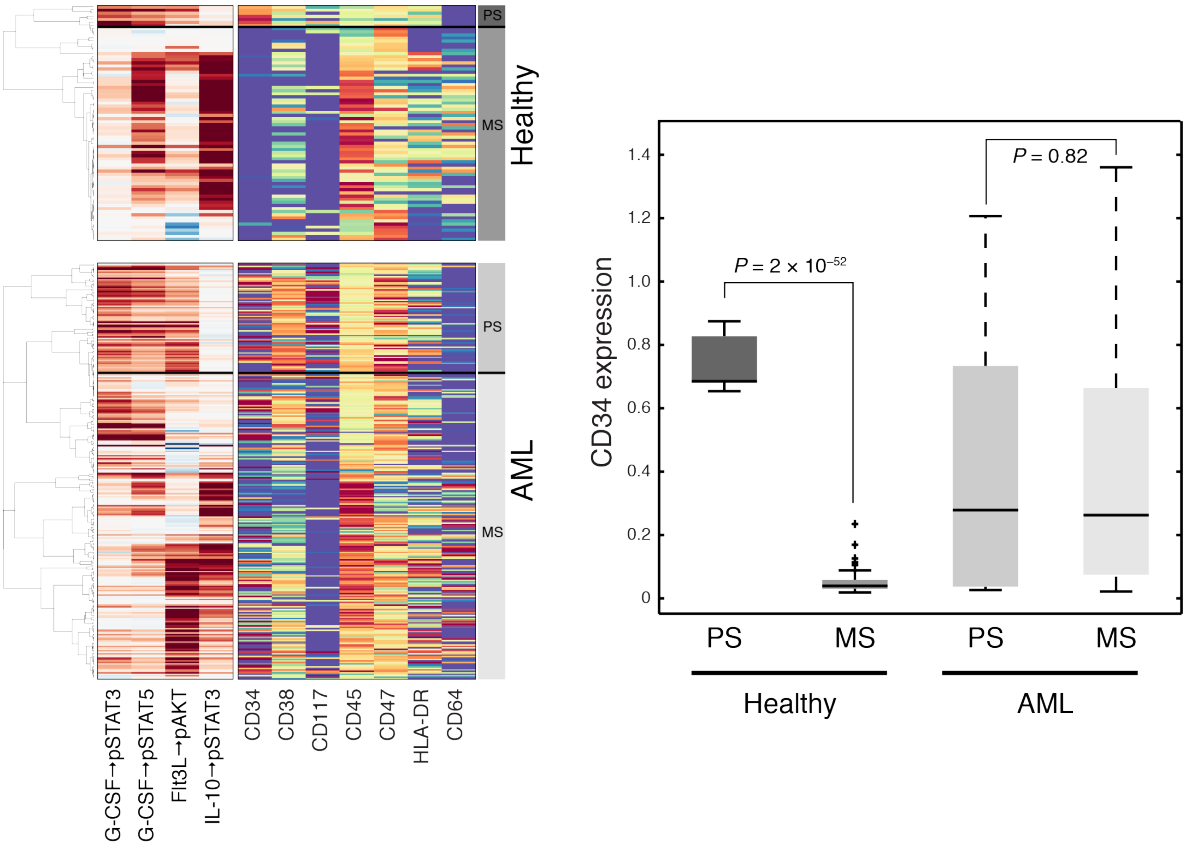


Figure 5.3: Surface and signaling phenotypes are decoupled in AML, compared to normal bone marrow. Left Panel: Linkage sorting of 4 developmentally-relevant signaling responses in the healthy samples identified patterns of primitive signaling (PS) and mature signaling (MS) correlated with expression of CD34 and CD45, in the healthy samples. Linkage sorting of the same signaling responses in the AML samples identified a cluster of subpopulations that recapitulated the primitive signaling pattern, but displayed no consistent surface phenotype. Colors as in Figure 5.4 and elsewhere. Right Panel: Box plots comparing CD34 expression between PS and MS groups, stratified by disease status. CD34 expression was significantly associated with primitive signaling only in the healthy samples.

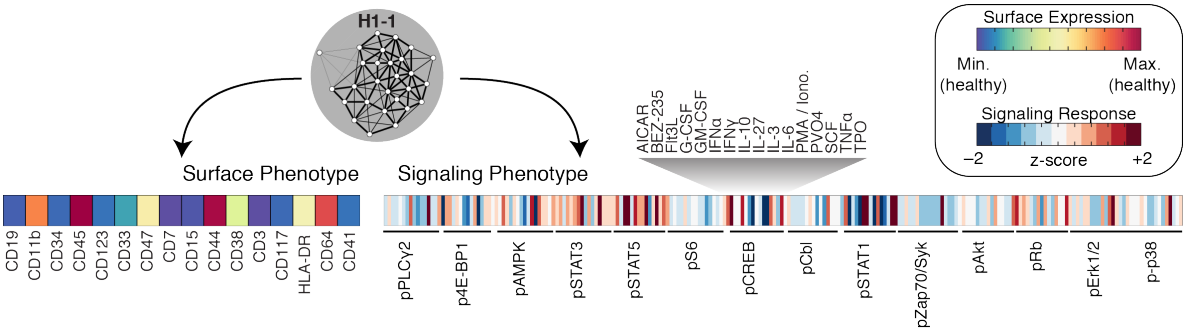


Figure 5.4: Each subpopulation has two alternative phenotypes: one reflecting surface marker expression, the other reflecting the configuration of the intracellular signaling network.

Signaling Response	FDR Q	ANOVA P
G-CSF→pSTAT3	6.40E-42	7.93E-44
IL-3→pSTAT3	2.13E-30	5.28E-32
IL-3→pSTAT5	1.88E-27	6.98E-29
PVO4→pZap70-Syk	4.49E-26	2.23E-27
PVO4→pP38	9.06E-23	5.62E-24
G-CSF→pSTAT5	1.11E-20	8.24E-22
PVO4→pErk1-2	1.28E-20	1.11E-21
GM-CSF→pSTAT5	3.13E-20	3.10E-21
PVO4→pSTAT5	1.10E-19	1.23E-20
PVO4→pPLCg2	1.12E-17	1.38E-18
PVO4→pSTAT3	3.37E-16	4.60E-17
PVO4→pS6	1.90E-15	2.82E-16
GM-CSF→pSTAT3	3.36E-13	5.42E-14
GM-CSF→pCREB	1.47E-12	2.54E-13
PMA/iono→pP38	3.47E-12	6.45E-13
PVO4→pAKT	1.00E-11	1.99E-12
GM-CSF→pS6	4.62E-11	9.75E-12
FLT3L→pAKT	8.47E-10	1.89E-10
SCF→pAKT	1.04E-09	2.44E-10
PVO4→pc-Cbl	1.60E-08	3.96E-09
IL-3→pS6	2.32E-08	6.03E-09
IL-10→pSTAT3	3.01E-08	8.21E-09
GM-CSF→pErk1-2	4.47E-08	1.28E-08
PVO4→pCREB	1.22E-07	3.62E-08
G-CSF→pCREB	1.31E-07	4.05E-08

Table 5.1: The 25 signaling repsonses most significantly associated with healthy cell type, as determined by ANOVA. Q-values were computed using the procedure introduced by [98] and implemented in the MATLAB R2013b Bioinformatics Toolbox.

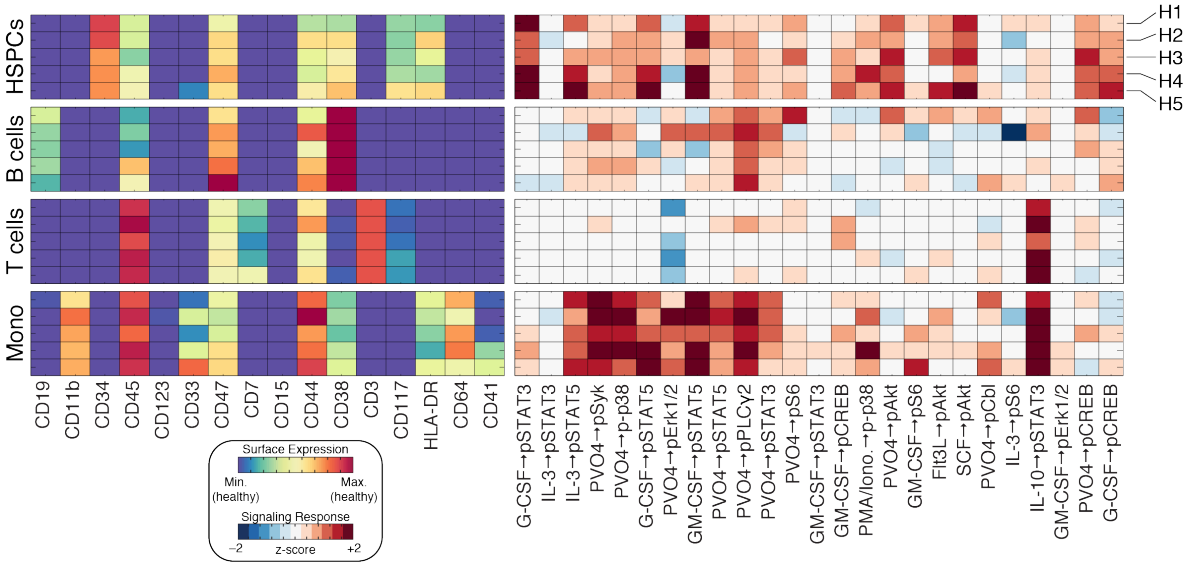


Figure 5.5: Four representative healthy cell types, identified by the HMC analysis. Each row (across left and right panels) represents each cell type in each healthy sample. The left panels display surface marker expression (which was used to name these groups). The right panels displays SARA scores for signaling responses, in descending order of significance from left to right as determined by ANOVA.

To test this idea, we used PhenoGraph Transductive Learning (PTL; §3.2) to classify subpopulations based on either their surface or their signaling phenotypes. First, we conducted testing on the 191 healthy subpopulations to see if PTL could successfully recover "held out" cell type labels (i.e., the HMC assignments) using a graph derived from surface phenotypes. Performance was evaluated using the cross-validated correct classification rate (CVCCR) as follows:

1. Build a single graph of 191 vertices using $k = 15$
2. Providing cell type labels for 4 of 5 healthy samples to PTL, classify the cell types for each vertex in the held-out sample
3. Compute the correct classification rate (CCR): the percentage of cells in the held-out sample for which PTL recovers the correct HMC label
4. Repeat (1–3), each time withholding a different healthy sample; average CCR over the repetitions to obtain the CVCCR

Considering that the graph which generated the HMC assignments is virtually identical[1] to

[1]In this case the graphs are not absolutely identical because a weighted Euclidean distance was used, see the next paragraph.

the graph used by PTL, one should expect very good performance from the surface phenotype graph and indeed the CVCCR was 99.42%. Such ideal performance might not be expected when, instead of surface phenotypes, PTL is given the signaling phenotypes. However, performance was quite good in that case, with CVCCR = 94%. Examining the results, we found that the classification errors involved distinguishing mature lymphoid cell types for which characteristic signaling phenotypes had not been measured. Focusing instead on the more significant task of distinguishing immature (i.e., HSPCs, HMC9) from mature cell types, we found that the signaling phenotypes were equivalently powerful to the surface phenotypes (CVCCR = 99.85%).

Before proceeding, a technical note. With 224 signaling responses, the space of signaling phenotypes is quite high-dimensional indeed and full of redundancies (which are there by construction: every dimension reflects the same phosphoprotein as 14 other dimensions and the same condition as 16 other dimensions [Figure 5.4]). To maximize the usefulness of the signaling phenotypes, we used a feature reweighting strategy. Each feature (i.e., signaling response) was reweighted by its importance for distinguishing healthy cell types, as quantified by the ANOVA p-values discussed earlier in this section. Specifically, for subpopulations $\mathbf{x}$ and $\mathbf{y}$ with D-dimensional phenotypes, the weighted Euclidean distance

$$d_{\mathrm{w}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{d=1}^{D} -\log\left(\mathrm{p}_d\right)(x_d - y_x)^2} \tag{5.5}$$

was used to define the k-neighbors for the first step of graph construction. In other words, the k-neighbor graphs were constructed in a space that emphasized the features that were important for distinguishing cell types in the healthy samples. In the interest of comparing performance between the surface and signaling phenotypes, the same reweighting strategy was applied to both surface and signaling phenotypes for PTL.

Considering that signaling phenotypes were sufficient to distinguish healthy primitive cells, we hypothesized that the functional state of AML subpopulations could be inferred by direct analysis of their signaling phenotypes. Using the 191 healthy subpopulations as training

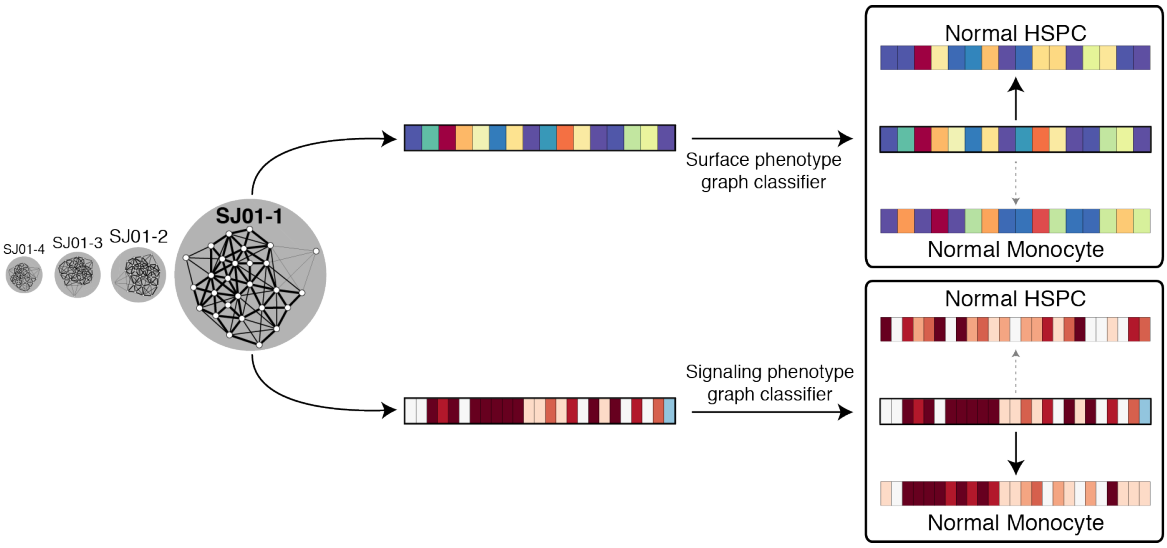


Figure 5.6: Each AML subpopulation was given two alternative classifications, one according to its surface phenotype and one according to its signaling phenotype.

data, we ran PTL on the full data set to infer maturity for each AML subpopulation (e.g., HSPC-like or monocyte-like). Because there were two alternative phenotypic profiles for each subpopulation, we performed two separate classifications—once using surface phenotypes and once using signaling phenotypes (Figure 5.6).

5.2.1 Inferred functional maturity diverges from surface phenotypes in AML

The classifiers identified primitive subpopulations within each patient sample, reflecting the heterogeneous nature of the samples. At the cohort level, each classifier labeled $\sim 25\%$ of subpopulations as primitive, but only 16% were identified as primitive by both classifiers simultaneously (Figure 5.7). In many cases (32/99), subpopulations with primitive surface marker phenotypes exhibited signaling that resembled mature cells. Conversely, many subpopulations displayed primitive signaling in the absence of primitive surface marker expression (51/118).

We denote cells labeled primitive by the surface phenotype classifier as Surface-Defined Primitive Cells (SDPCs) and cells labeled primitive by the signaling classifier as Inferred Functionally Primitive Cell (IFPC). For each patient, the sample proportion assigned to

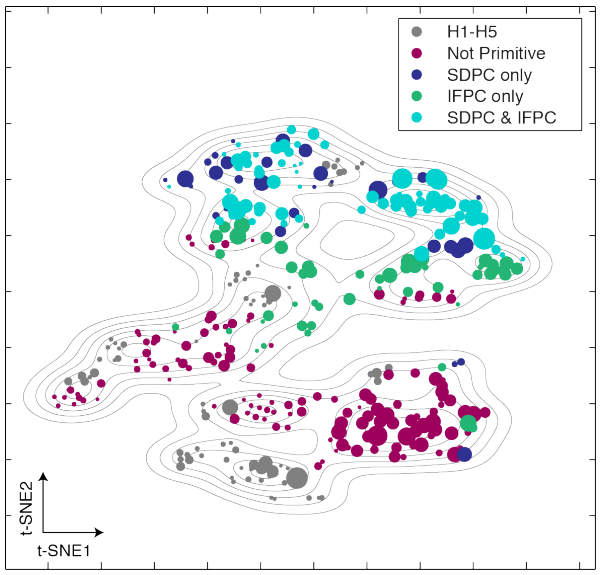


Figure 5.7: Results of the surface and signaling based classifications on the t-SNE map of Figure 4.5

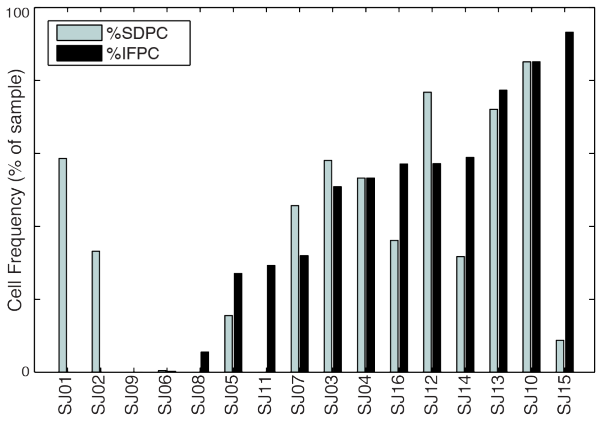


Figure 5.8: Frequencies of primitive cells in each patient as determined by the two alternative definitions.

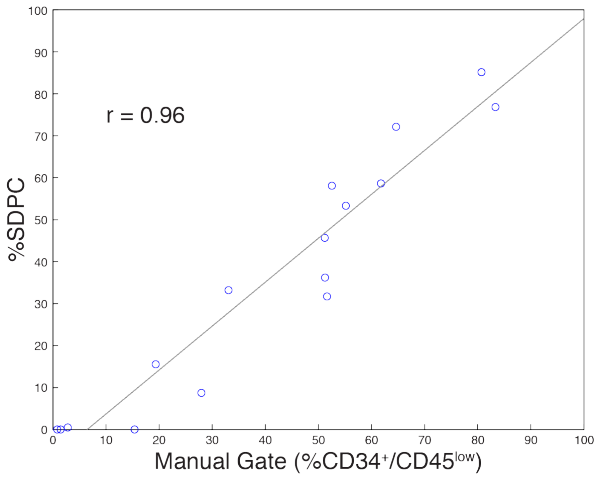


Figure 5.9: The data-driven estimates of primitive cells based on surface marker profiles (%SDPC) is highly correlated with an independent estimate based on a standard immunophenotypic procedure for blast enumeration. This indicates that PhenoGraph clustering and transductive learning are highly accurate while also implying that CD34 and CD45 effectively summarize the maturity-related information of surface marker profiles.

each of these labels produced two alternative measures of maturity (%SDPC or %IFPC; Figure 5.8). This is similar to summarizing the degree of maturation by the enumeration of $CD34^+/CD45^{low}$ blasts, a practice often used in the clinical diagnosis and classification of leukemias [93]. Indeed, we found that %SDPC was highly correlated with this standard manual gating procedure, which was performed independently (Figure 5.9; Pearson's $r = 0.96, P = 4.4\times10^{-9}$). Conversely, %SDPC was only weakly correlated with %IFPC (Pearson's $r = 0.5; P = 0.05$), demonstrating that these two metrics are not redundant in AML. Instead, examination of signaling phenotypes in AML often revealed a different degree of maturation than was indicated by the surface phenotype. We noted that the degree of discordance between IFPC and SDPC assignments was not constant across patients, indicating that the tendency of IFPCs to express canonical LSC markers was itself a variable patient feature. For example, the IFPCs in patient SJ05 were well represented by the $CD34^+/CD38^{mid}$ phenotype (Figure 5.10, left column). In other cases, IFPCs were found exclusively in the $CD34^-$ fraction, even when $CD34^+$ blasts were abundant (e.g., SJ16).

Differences in signaling patterns between primitive and mature leukemic subpopulations reveal the responses most important for these classifications. To quantify the importance

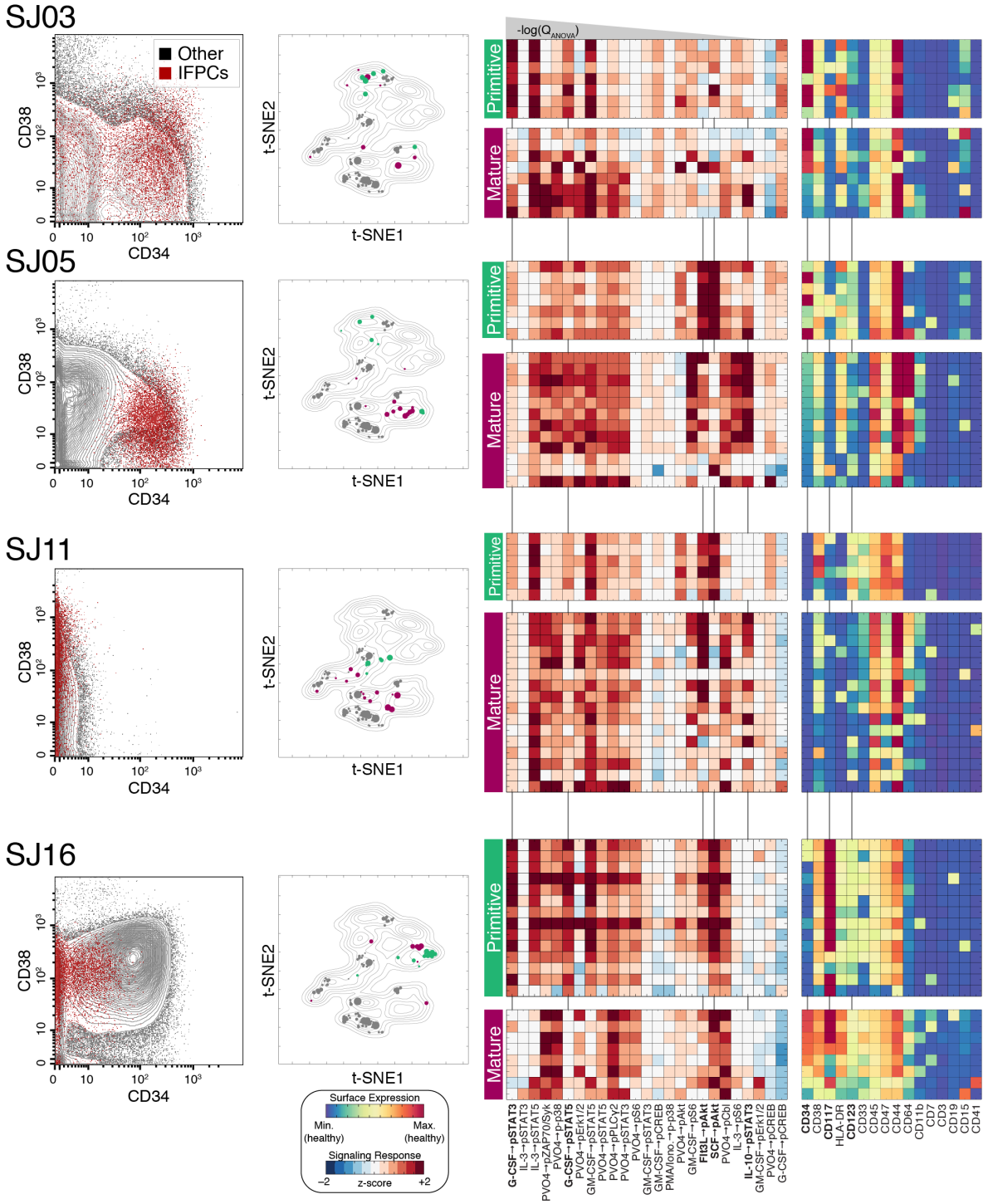


Figure 5.10: Detailed surface and signaling phenotypes of IFPC subpopulations in 4 representative samples. Biaxial dot plots (left) show the CD34/CD38 phenotype of IFPCs (red) in each sample. IFPCs displayed the canonical CD34/CD38 phenotype of primitive cells in only a subset of samples, here best exemplified by SJ05. The central panels show the placement of IFPCs on the cohort landscape (Figure 4.5; IFPCs in green, non-IFPCs in maroon, H1–H5 in gray). On the right, heat maps display the signaling and surface phenotypes of all non-lymphoid subpopulations of each sample, stratified by IFPC classification (indicated by green and maroon bars). Signaling responses are ordered as in Figure 5.5. Signaling responses marked in bold with vertical lines were especially distinctive of IFPCs.

of each signaling response, we used canonical variates analysis (CVA), a multi-class, multi-dimensional generalization of Fisher's linear discriminant [99]. CVA is similar to principal components analysis (PCA) in that it seeks a linear projection of the high-dimensional data into a low-dimensional space, but whereas PCA is unsupervised, CVA uses class labels to find a projection that maximizes the separability of these classes in the target space. This method allows visualization of the linear separability of the classes in low-dimensional space and simultaneously identify the features that are important for obtaining that separability by examining the projection matrix.

In this case, the input data were the 224-dimensional signaling phenotypes and the class labels were binarized into IFPC and non-IFPC. CVA revealed that the linear separability of these classes could be maximally preserved in a single dimension (Figure 5.11). The signaling responses most important for class separability were identified as the entries of the projection matrix with the largest (absolute) magnitudes. In this way, CVA performs feature selection implicitly. Re-running CVA with small numbers of top-ranking features produced nearly equivalent separability. We found that that the majority of discriminative power could be attributed to 5 responses: G-CSF→pSTAT3, SCF→pAKT, G-CSF→pSTAT5, FLT3L→pAKT, and IL-10→pSTAT3. Primitive subpopulations displayed strong activation in the first four of these responses, which have all been previously implicated in the biology of HSPCs [97, 100, 101] and in the pathobiology of AML [46, 102, 103]. Additionally, attenuation of the IL-10→pSTAT3 response—a response exhibited by mature immune cells [104]—was also a distinctive feature of IFPCs.

Evaluating the ability of surface markers to identify IFPCs, it was clear that no surface phenotype could be applied universally across patients. CD34 was often an important label for IFPCs, but only in a subset of cases. CD34 marked both primitive and mature subpopulations in patient SJ03, where HLA-DR was a more specific marker of IFPCs ($P = 0.0007$ vs. $P = 0.003$, Student's t test). In SJ05, where CD34 expression was tightly associated with IFPCs ($P = 7.4 \times 10^{-8}$), the multidimensional surface measurements revealed that CD123 was also an important marker ($P = 4.4 \times 10^{-6}$), whereas CD123 did not identify IFPCs in SJ03.

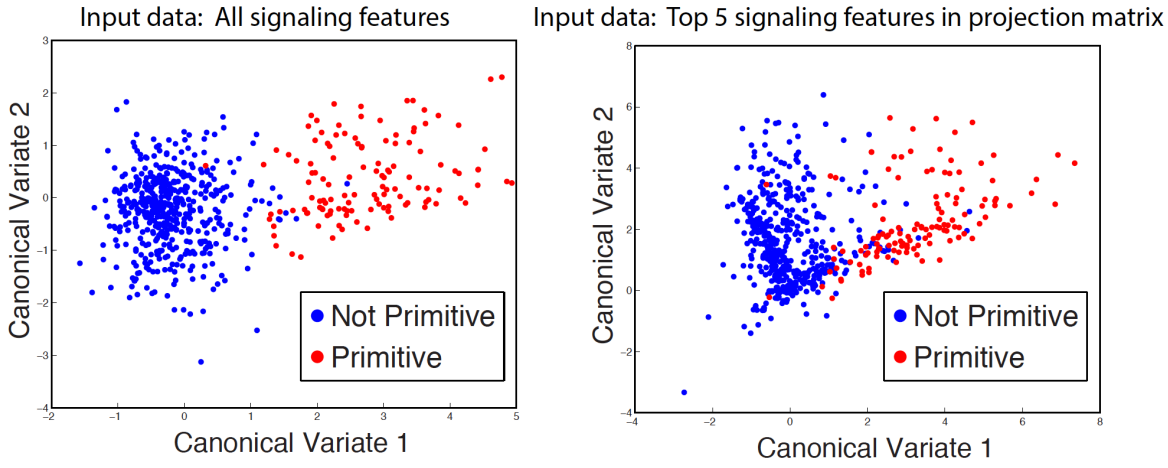


Figure 5.11: Canonical variates analysis identifies the signaling features that most effectively separate IFPCs from non-IFPCs. LEFT PANEL: IFPCs and non-IFPCs can be almost perfectly separated using the first canonical variate. RIGHT PANEL: Almost the same separability is achieved when only the top 5 features are used to compute the projection.

Patient SJ11 lacked CD34 expression almost entirely, as expected for this nucleophosmin-mutated case [39]. In this patient, IFPCs were distinctly labeled by elevated expression of CD47 ($P = 7.1 \times 10^{-6}$) and CD123 ($P = 3.4 \times 10^{-5}$). Surprisingly, we found that CD34 expression can be strongly anti-correlated with primitive signaling, as in patient SJ16, where CD34 expression was higher in mature cells ($P = 0.0027$) and IFPCs were marked instead by elevated expression of CD117 ($P = 0.0026$).

5.3 Signaling phenotype identifies clinically prognostic gene expression signature

Ultimately, intratumor heterogeneity is import insofar as functionally distinct subpopulations influence clinical outcomes, especially patient survival [42]. While our mass cytometry cohort was too small for survival analysis, genome-wide expression arrays for 15 of our 16 patients were available from a previous study ([105] and §A.4), providing a link to larger cohorts for which gene expression and survival data were available. Because our samples displayed a wide range of IFPC frequencies (Figure 5.8), we reasoned that this variance could be exploited to

identify genes whose expression covaried with these frequencies by *in silico* expression deconvolution [106]. As IFPC frequency varies across samples, genes expressed specifically by these cells should be detectably more or less abundant in the bulk gene expression measurements, thereby providing an estimate of %IFPC in independent samples from the level of this gene signature, measured in bulk.

5.3.1 Gene expression deconvolution

We developed a deconvolution method based on linear regression and cross-validation. The basic assumption of *in silico* deconvolution is that a subpopulation of interest expresses certain genes at constant rates; therefore changes in bulk expression measurements of these genes will track with changes in subpopulation size. This can be formulated as the general linear model

$$Y = X\beta + \epsilon \tag{5.6}$$

where Y is a $N \times G$ matrix of G mean-centered gene expression values, X is a $N \times 1$ vector of subpopulation frequencies (e.g., %IFPC) for N samples, and β is a $1 \times G$ vector of regression coefficients. β can be obtained from the least squares solution and represents, for each gene, its estimated "expression level" in the subpopulation.

Because gene expression data are noisy and our data contained arrays for only 15 patients, we developed a cross-validation scheme to reduce overfitting and spurious associations. Specifically, we used leave-two-out cross-validation: The patients were split into 105 unique combinations of 13 from 15 and β was solved for each of these 13-patient data subsets. For each solution of β, genes were assigned to percentile bins. Genes were added to the signature if they were placed in the top percentile more often than any other bin and had a standard deviation across data subsets of less than 5 percent. The entire strategy was performed to identify two subpopulation-associated gene signatures: an IFPC-associated set of 49 genes and a SDPC-associated set of 42 genes (Table 5.2).[2] To characterize these signatures, we queried

[2] 14 genes appeared in both signatures. These may be interesting in their own right but form too small of a group to analyze and were generally excluded for comparisons of the other signatures.

IFPC	SDPC	Both
PROM1	CD34	B4GALT6
RRAGD	GIT2	CD69
GPHN	ITM2C	EMP1
CPA3	PMAIP1	GNG7
FOSB	PTGER4	IL1RAP
ATF3	EHD3	JUP
CD96	MIR4448	KIT
IL8	MRC1	LOC282997
ATP1B1	HGF	LTBP3
SIK1	CD200	PDGFC
TPBG	MAN1A1	PIK3C2B
PTGS2	XPA	SORL1
NR4A2	BAALC	SPRY2
SLC2A5	NRXN2	TNFRSF21
ZSCAN5A	DHRS3	
CLMN	CLEC2B	
GNA15	MDFIC	
FCHO1	PELI2	
GPR125	ANPEP	
DDIT4	GAB2	
AKR1C3	HIST1H2BH	
SPINK2	PMP22	
RUFY3	MN1	
LENEP	DAGLA	
CTNNBIP1	NCAPH2	
KCNJ8	APOL2	
CLDN6	9-Sep	
LHX6	GATM	
CST7	EGFL7	
CLIP2	AGTPBP1	
ENPP2	PTOV1	
CLN6	WASF1	
FOCAD	SEC31B	
LZTFL1	TGFB1I1	
MEGF6	FLNB	
NFIL3	DAPK1	
MEX3C	DEPTOR	
SEMA6D	NARFL	
ENTPD6	ZCCHC14	
GLS	SIDT1	
DUSP10	CLIC4	
IGFBP2	FCGRT	
ADA		
ITPR2		
PLXNC1		
SOCS2		
ANKRD28		
DNMT3B		
SMAD1		

Table 5.2: Gene signatures obtained from microarray deconvolution.

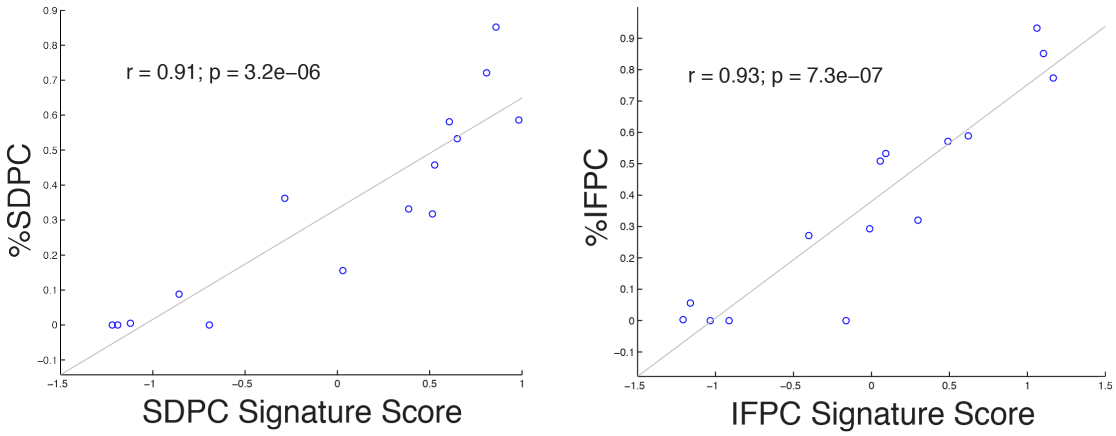


Figure 5.12: The mean expression of the gene signatures identified by deconvolution are strongly correlated with the subpopulation frequencies they are supposed to represent. This is a necessary condition to use the signature means as proxies in other samples.

the Molecular Signatures Database [107] for annotations significantly overlapping with each. The SDPC signature—which contained CD34 as its top-ranking gene—was highly enriched for gene sets associated specifically with $CD34^+$ AML (e.g., cases with the AML-ETO1 fusion [108], $Q = 7.4 \times 10^{-11}$). Alternatively, the most significant annotation for the IFPC signature was a set of genes upregulated in $CD133^+$ hematopoietic stem cells [109] ($Q = 5.5 \times 10^{-8}$). CD133 marks healthy stem cells that are possibly more primitive than $CD34^+$ HSCs [110] and has been linked to cancer stem cells in multiple cancer types [111, 112]. The mean expression of each signature was highly correlated with its corresponding subpopulation frequency (Figure 5.12), indicating that the signature mean was an appropriate proxy for these frequencies in independent clinical cohorts for which single-cell measurements were not available.

5.3.2 Survival analysis

The signatures were tested in two independent cohorts of adult AML for which both gene expression and survival data were available for a total of 242 patients [113]. For Kaplan-Meier analysis and log-rank statistics, each patient was assigned to one of two groups based on whether its signature score (i.e., the mean expression of genes in the signature of interest) was above or below the cohort median.

While the SDPC signature was associated with survival in one cohort, this was not replicated in the other. Alternatively, the IFPC signature was predictive of poor survival in both cohorts (Figure 5.13). Combining the patients into one large cohort, the IFPC signature was highly predictive of poor survival ($P = 4.8 \times 10^{-6}$, Hazard Ratio [HR] = 3.4), while the SDPC signature formed a less significant predictor ($P = 0.005$, HR = 1.6). To test these signatures against each other, we placed them together in a bivariate Cox regression model. In this setting, the IFPC signature retained its predictive power ($P = 8.2 \times 10^{-5}$, HR = 3.0), while the SDPC signature became completely uninformative for survival ($P = 0.29$, HR = 1.2).

We also examined the relationship between the IFPC signature and three signatures reported by [43], which were also developed to capture primitive gene expression programs in AML. For each Eppert signature, we found that the reported correlations with survival in the data of [113] were reproducible. To assess the prognostic value of the IFPC signature when these other signatures were known, we tested three bivariate Cox regression models in which each of the Eppert signatures was used as a predictor alongside the IFPC signature. The IFPC signature proved to be a stronger predictor of survival than any of the Eppert signatures. In each model, the IFPC signature retained significance ($P < 0.005$), while each Eppert signature became statistically insignificant ($P > 0.07$). In a multivariate Cox regression model containing all signatures (IFPC, SDPC and the Eppert signatures), only the IFPC signature retained significance ($P = 0.012$, HR = 2.4).

5.4 Discussion

In Chapter 4, we saw that the surface marker phenotypes of leukemic cells are restricted to a limited set of patterns that somewhat resemble myeloid development. This resemblance suggests that the functional maturity of leukemic cells might be reflected in their surface marker expression patterns. In this chapter, we have shown that the apparent maturity suggested by surface marker expression often belies a different functional state as reflected in the underlying behavior of the cells' signaling networks. Whereas surface and signaling phenotypes displayed tight coregulation in normal bone marrow, this coregulation was broken

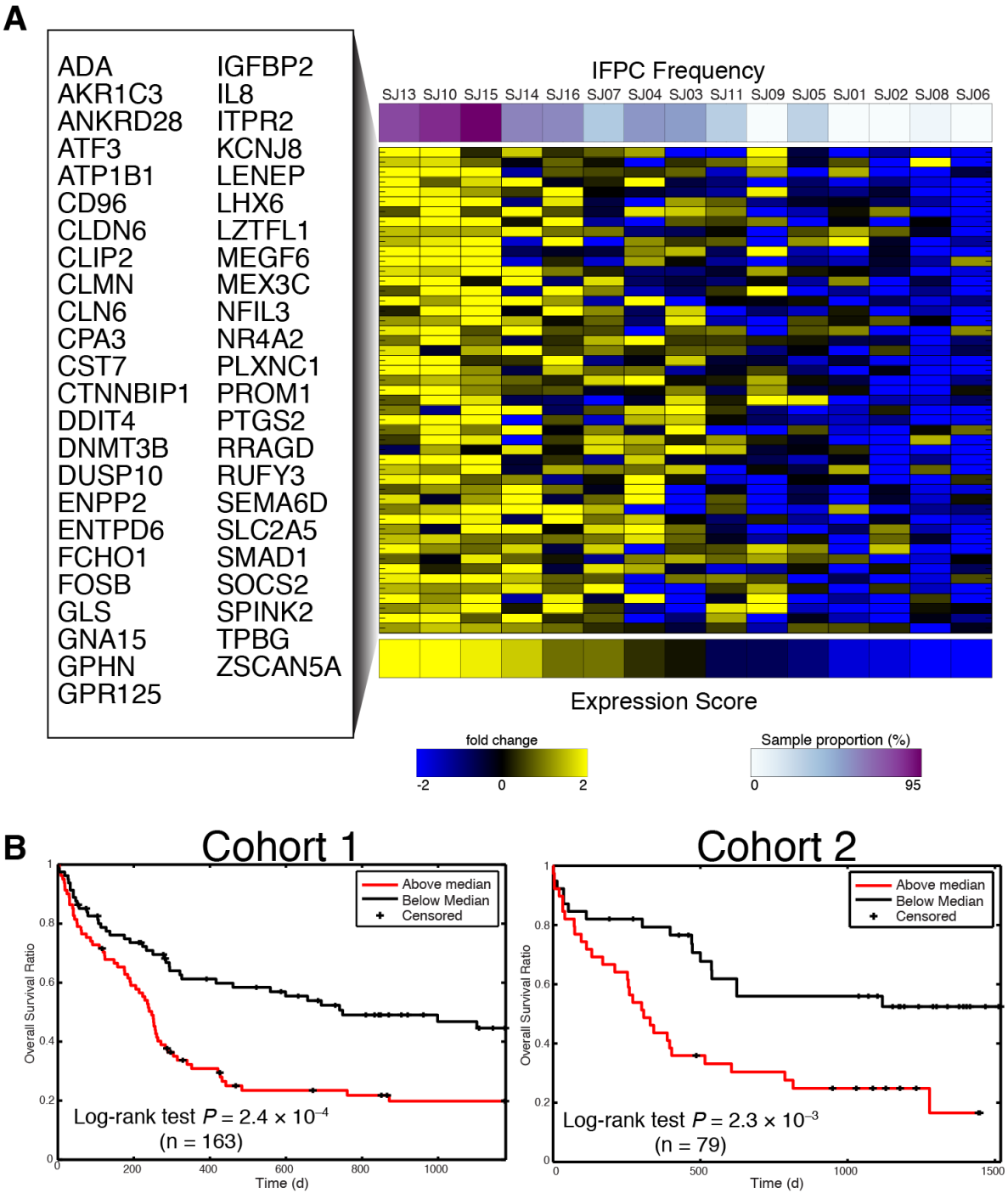


Figure 5.13: IFPC frequency identifies a gene expression signature that predicts clinical outcome. (A) IFPC gene signature identified by deconvolution of bulk expression data. The heat map displays expression of each gene in bulk measurements. Rows are alphabetically ordered; columns are ordered by the mean expression of the genes in the signature. (B) The mean of the IFPC signature forms a clinically significant prognostic indicator of overall survival in 2 independent cohorts of adult AML [113]. Patients were stratified for Kaplan-Meier analysis by whether their IFPC expression score was below or above the cohort median.

in AML. The substantial decoupling of surface and signaling phenotypes in the leukemic cells renders the surface markers typically used in diagnostics unreliable proxies of cellular state and function in AML.

Using PhenoGraph Transductive Learning (Chapter 3), we identified leukemic subpopulations that displayed signaling phenotypes similar to the immature hematopoietic stem and progenitor cells (HSPCs) of normal bone marrow. This provided a data-driven classification of functionally primitive leukemic subpopulations irrespective of their surface phenotypes. We found that these inferred functionally primitive cells (IFPCs) displayed signaling responses consistent with known properties of immature hematopoietic cells, such as phosphorylation of STAT3 on exposure to G-CSF [97].

The IFPC phenotype was found to occur in most AML samples at varying frequencies and with variable surface phenotypes, often with low or absent CD34 expression. While no universal surface phenotype captured IFPCs across patients, within each patient IFPCs displayed homogeneous expression in certain markers—markers whose importance was neither universal nor unique. Our results suggest that a subset of leukemic cells maintains a conserved, progenitor-like signaling program that phenocopies the regulatory state of normal HSPCs, regardless of surface marker expression and underlying genetic mutations.

Deconvolution analysis of microarray data identified a gene expression signature associated with the IFPC phenotype that can serve as a proxy for the frequency of this phenotype in a given sample. This gene expression signature was enriched for annotations related to primitive hematopoietic cells and included genes—such as *PROM1*, *SOCS2*, and *CD96*—that have been previously associated with healthy and/or leukemic stem cells [114, 115]. Importantly, this gene expression signature predicted survival in independent AML patient cohorts, suggesting that this signaling-based definition describes a clinically relevant cellular phenotype.

It was previously demonstrated [43] that functional characterization by physical sorting and xenotransplantation could be used to identify genes correlated with patient survival. Our analysis is conceptually related, but instead of differential expression between sorted cells, we

used *in silico* deconvolution to identify genes, based on the measured cellular frequencies of the IFPC phenotype. Ultimately, both approaches seek to identify primitive cells by means that emphasize functional over surface phenotypes, and to test whether the predominance of primitive cells——approximated by expression of a gene signature——is associated with poor survival.

The signaling-based definition of primitive cells warrants further investigation as it may indicate pathways that influence the maturation of leukemic cells and could be leveraged therapeutically to block survival or direct differentiation. More broadly, this molecular interrogation approach could be used to characterize primitive cells in any cancer where a cognate healthy primitive cell type is available to serve as a reference point.

Chapter 6

Perspectives and conclusions

6.1 Dissertation summary

Single-cell analysis has resumed the spotlight in biological science owing to technological developments that allow these primary units of the organism to be measured with high enough dimensionality that their true phenotypic and functional diversity is revealed. In particular, this dissertation has explored the diversity of cell states revealed by mass cytometry. Mass cytometry is capable of measuring dozens of protein features per cell, including phosphorylation of intracellular signaling molecules, and allows barcoded multiplexing of cells measured under different molecular perturbations. This experimental design produces complex data that benefit from computational methods, which integrate and extract knowledge from them.

Experimental approaches to immunophenotyping have hit an analytical ceiling because they rely on the predictability of the underlying system that cannot always be assumed—especially in disease states such as cancer, where molecular deregulation can cause different cell states to emerge from patient to patient. In leukemia, a particularly important cell state—the leukemic stem cell—has been shown to exist in most patients, yet it is known to display different immunophenotypes in different patients. This presents a challenge that experimental approaches which require *prospective* gate definitions, like FACS, are ill-equipped to deal with. In this dissertation, *data-driven* computational approaches present a fruitful alternative direction. By assuming less about the cancer samples, more can be learned.

To take a data-driven approach, we begin by modeling each sample as a composition of subpopulations. Subpopulations are defined algorithmically, based on quantitative high-dimensional measurements, rather than prospective gating based on assumptions from knowledge of normal hematopoiesis. The algorithm is motivated by a theoretical construct representing the process that generates a multicellular population—the phenotypic manifold. Taking insights from topology, we develop an approximation of the phenotypic manifold that is based on a graph. The graph models the cellular population by modeling *connectivity* between cells, which reflect phenotypic similarity. Reasoning that stable cellular states will result in regions of high local density, we equate densely interconnected subgraphs with biologically meaningful subpopulations. Borrowing from network analysis, we use the concept of modularity to partition the graph into densely interconnected subgraphs, revealing the biologically meaningful subpopulations. This method was given the designation PhenoGraph (Chapter 2).

Applying PhenoGraph to an AML cohort established that each patient's leukemic blasts occupied multiple phenotypic states. Surprisingly, these states fell into a restricted set of expression patterns that resembled myeloid development. As the number of dimensions increases, the number of possible marker combinations increases geometrically—if the markers are independent. Instead we found that the number of leukemic states was less than the modest number of patients in our cohort, suggesting that strong constraints on the phenotypic landscape of AML blasts persist after disease initiation. Interestingly, the distribution of each patient into these limited expression patterns was significantly associated with some genomic biomarkers, indicating a genetic influence on the subpopulation structure of leukemia.

A majority of leukemias contained blast subpopulations that resembled less and more differentiated myeloid cell types, suggesting a developmental continuum present in each patient. This is consistent with other evidence that AML is a disease caused by dysfunctional myelopoiesis. However, while many of the blast subpopulations displayed immunophenotypes resembling normal hematopoeitic stem and progenitor cells (HSPCs), it is also known that leukemic stem cells do not always display this phenotype. We therefore undertook an inde-

pendent functional characterization of each subpopulation based on molecular perturbation data.

Most functional assays, such as colony formation or xenotransplantation, require prospective isolation of cells by FACS prior to functional characterization. We took an alternative, *in silico* approach to functional characterization, which was compatible with the high-dimensional, data-driven method used for identifying subpopulations. To functionally characterize each subpopulation, we made use of the short-term molecular perturbations, which—at the order of 15 minutes—trigger intracellular signaling cascades without causing significant changes to surface marker expression. Therefore, because subpopulations were defined using only surface markers, each subpopulation contained cells exposed to one of 17 distinct environmental conditions. Making the assumption that the responsiveness of signaling cascades to these biologically relevant perturbations (e.g., G-CSF) reflects the regulatory state of the intracellular machinery that controls cell function, we took these perturbation data as a starting point for functional characterization. Quantifying the change in phosphorylation of each signaling protein caused by each perturbation resulted in 14 (intracellular markers) $\times$ 16 (17 - 1 for the basal condition) = 224 signaling responses per subpopulation. Thus, in addition to its 16-dimensional surface marker profile, each subpopulation had a 224-dimensional signaling profile that reflects the regulatory state of cells it contains.

To interpret these two sets of phenotypes, we integrated knowledge from normal hematopoiesis. All analyses described above were also applied to our 5 normal bone marrow controls. As expected, each normal bone marrow was partitioned into virtually identical compositions of distinct subpopulations, reflecting the regulatory precision and predictability of healthy tissues. It was straightforward to interpret the healthy subpopulations and identify them as various lymphoid, myeloid, or progenitor cell types. These labels made it possible to set up a transductive learning problem (Chapter 3) in order to categorize the leukemic subpopulations according to their phenotypic or functional relatedness to the healthy cell types. This approach allowed inference from features of healthy progenitors to features of leukemic subpopulations in the high-dimensional measurement spaces of either surface or signaling

phenotypes—a quantitative extension of previous attempts to infer leukemic maturity from the expression of one or two markers (CD34 and CD38).

As expected, transductive learning from surface phenotypes yielded results very similar to manual gating based on CD34. On the other hand, transductive learning from signaling phenotypes yielded results that sometimes varied substantially from the predictions based on surface phenotypes. Leukemic subpopulations that look like HPSCs on the surface do not always display HSPC-like signaling, and leukemic subpopulations that display HSPC-like signaling do not always look like HSPCs on the surface. Whereas the percentage of cells with a surface marker profile resembling HSPCs (%SDPC, surface-defined primitive cells) yields a more "conventional" characterization of leukemic maturity (similar to %$CD34^+$), the percent displaying HSPC-like signaling (%IFPC, inferred functionally primitive cells) provided an alternative estimate of the frequency of functionally undifferentiated cells. In other words, these characterizations might be taken as alternative estimates of the severity of the differentiation block for each patient.

It has been previously shown that the estimated frequency of immature leukemic blasts is correlated with clinical variates such as overall survival. We therefore undertook similar analyses to test whether %IFPC or %SDPC were significantly associated with survival. Because the primary mass cytometry cohort was too small for survival analysis, we developed a bridge to test these features in larger cohorts. Because bulk gene expression (i.e., microarray) data were available for the samples in the primary cohort, we were able to extract sets of genes, via *in silico* gene expression deconvolution, that were highly correlated with either %IFPC or %SDPC across the primary cohort. Gene set enrichment analysis supported the hypothesis that these genes were actually expressed in the IFPC or SDPC states. We therefore hypothesized that other patients with high frequencies of IFPCs or SDPCs should have detectably higher expression in these respective gene sets if their marrow is also subjected to bulk expression measurement. Using two independent cohorts, we found that the IFPC signature in particular was significantly predictive of overall survival in AML patients.

The analyses presented here represent a large-scale, data-driven, quantitative and mul-

tivariate study of cellular phenotypes in healthy and diseased tissues. In possessing these qualities, the work can be considered an answer to the call for a new brand of quantitative cell biology that utilizes the advantages of computational inference [1]. We have demonstrated the power of data-driven techniques for identifying diverse cell states and learning about their features. In particular, we have shown that the phenotypic composition of human acute myeloid leukemias is diverse but restricted to a palette of progenitor- and myeloid-like phenotypes whose frequencies are influenced by genetic background. Our data-driven analysis of subpopulation function demonstrates that phenotypic diversity can indicate functional diversity but in ways that are unpredictable across patients, a finding that has broader implications for studies of cancer stem cells and intra-tumor heterogeneity.

6.2 Contributions of this work

Computational biology was initially inspired by the emergence of genome wide measurement technologies. As those technologies are emerging at single-cell resolution, other methods which measure protein expression at single-cell resolution have reached sufficiently high dimensionality that they can need computational methods, too.

Phenotypic heterogeneity in cancer is a popular research topic, and rightly so. To our knowledge, few efforts have been made to approach this problem from a computational perspective, yet the subject is ripe for the benefits this brings. Precisely because cancer is unpredictable, efforts to understand phenotypic heterogeneity using the apparatus of conventional immunophenotyping have been stifled by inconclusive results. On the other hand, computational methods that can discover the known population structure of normal tissues such as bone marrow *ab initio* are ideal for studying the composition of tumor samples as they can recapitulate the efforts of decades of experimental research in the context of a single sample. Thus, the potential uniqueness of a cancer sample is met by an approach that is indifferent to that feature which is so problematic for prospective approaches.

The idea of identifying subpopulations computationally in single-cell data is not itself novel, but rather suggests itself naturally to many who have worked with this type of data.

A number of methods have been developed to this end, especially in recent years. Unfortunately, we found that these methods did not deliver quality results at the scale of mass cytometry data, in terms of both cell count and dimensionality. We believe it is vital to accommodate current and future scaling up of single-cell measurements and to do so without sacrificing quality. With the concurrent aims of scale and quality in mind, we developed a novel approach that extracts subpopulation structure using a graph-based representation of phenotypes. The use of the graph is conceptually motivated and handles features of single-cell data that thwart other methods, such as non-linear relationships between protein expression and subpopulations that form arbitrary non-convex shapes in high-dimensional space. The graph-based implementation also provides access to computationally efficient techniques which can accommodate very large samples without resorting to subsampling, which substantially improves the robustness of the results.

The work presented here does not simply aim to show that phenotypic heterogeneity exists, but rather takes it as a starting point for several further analyses. To our knowledge, no previous work presents a quantitative comparison of intratumor heterogeneity across multiple cancer samples. The resulting *cohort landscape* simultaneously demonstrates not only that intratumor heterogeneity is widespread, but how the subpopulations of different individuals relate to each other. The landscape shows that patients can differ both in the phenotypes that occur and in the frequencies with which they occur. The complementary metacluster analysis showed that these frequencies are influenced by the patient's genetic background. This demonstrates that a genetic lesion need not produce a single malignant phenotype but may cause a particular distribution of phenotypes, highlighting the interplay between genetic and epigenetic mechanisms in generating intratumor heterogeneity.

The computational analysis of subpopulation function provided new insights to the controversial relationship between phenotypic and functional heterogeneity. The *in silico* functional estimates were computed for each of ~ 30 subpopulations per leukemic sample, representing an order of magnitude higher resolution than can be achieved by FACS gating, which furthermore suffers from the prospective gating requirement as mentioned several times pre-

viously. At this level of resolution, it was clear that functionally primitive subpopulations often but far from universally express CD34. Similarly, many $CD34^+$ subpopulations did not display a functional profile similar to HSPCs. Even at 16 dimensions, no universal pattern of surface marker expression identified functionally primitive subpopulations. On the other hand, the signaling profile itself presented a high-dimensional pattern that more closely reflects the maturation of leukemic subpopulations, supported by the gene expression and survival analyses. The discriminative power of these signaling patterns could largely be attributed to 5 features: G-CSF→pSTAT3, SCF→pAKT, G-CSF→pSTAT5, FLT3L→pAKT, and IL-10→pSTAT3. Compared to the various observed surface marker patterns, a definition of functional maturity based on these signaling features is more universal across patients.

From a methodological perspective, PhenoGraph is a novel technique for learning cell states from data, which displays unprecedented accuracy and computational efficiency. Its accuracy is attributable to its representation of cellular phenotypes as a graph that approximates an underlying phenotypic manifold. By operating on neighborhoods instead of distances themselves, the Jaccard graph encodes the connectivity between points together with their local density, which enhances edges within continuous structures and penalizes edges that span sparse or noisy regions. The quality of PhenoGraph's results can also be attributed to the selection of modularity as an appropriate objective function for density detection in the Jaccard graph. The efficiency can be attributed to parallelization of nearest-neighbor computations as well as to the availability of modularity optimization routines that are designed for large-scale social network analysis.

In this dissertation, PhenoGraph was introduced together with a demonstration of its utility for investigating the pathophysiology of a particular cancer—acute myeloid leukemia. PhenoGraph is general and scalable both in terms of dimensionality and sample size, making it suitable in a wide range of settings for which single-cell population structure is of interest, including other cancers or healthy tissues, and for use with other emerging single-cell technologies such as single-cell RNA-seq. Many such cases are presented by the tumor microenvironment, including drug-resistant tumor subpopulations, infiltrating immune cells, and

reactive stromal components. PhenoGraph is also applicable to healthy tissues, within which a large diversity of cell types remains uncharted.

6.3 Future directions

PhenoGraph addresses probably the most fundamental question posed by single-cell data: What are the subpopulations present in the sample? In this dissertation, PhenoGraph answered this question in the context of normal and malignant bone marrow. PhenoGraph requires only the simplest experimental design: Measure cells. Given nothing but these measurements, PhenoGraph extracts groups of phenotypically coherent subpopulations. Because PhenoGraph does not consider additional "metadata"—such as condition or time point—it (and in fact any clustering method) is an *unsupervised learning* technique. On the other hand, the PhenoGraph Transductive Learning (PTL) algorithm takes a more structured data set, in which metadata are available—in the form of class labels—for a subset of measurements. An algorithm like PTL, which extends information from known to unknown observations, is a *semi-supervised learning* technique. PhenoGraph and PTL use the same underlying data model—a graph of phenotypes—yet they operate on that model differently in order to answer different kinds of questions.

Indeed, the graph-based representation of phenotypes is a powerful and flexible approach to single-cell data, which can be deployed in different ways to target different experimental questions. For example, in previous work our group used graphs to model cellular differentiation as a continuum of transitional states, a method called Wanderlust [53]. While PhenoGraph analyzes the density structure of the graph, Wanderlust and PTL consider a *dynamic process* on the graph: a random walk.

As we saw in Chapter 3, graphs possess a deep connection to probabilistic potential theory and, as we explore presently, to the general mathematical framework of Markov chains. The graphs used in Wanderlust, PhenoGraph, and PTL are only a simple matrix operation away from explicitly representing a Markov chain. When this connection is made, powerful analytic tools developed for Markov chains become available for the analysis of single-cell data.

Consider the weight matrix $\mathbf{W}$, discussed in Chapters 2 & 3 (e.g., §2.2.1), in which the entry $\mathbf{W}_{ij}$ represents the phenotypic similarity between cells i and j. Drawing again on the phenotypic manifold concept (§1.3.4), we might assume that phenotypic similarity implies a dynamical relationship—specifically, if cell j is similar to cell i it is likely that j is derived from i or was recently in a state identical to i. This relationship can be formalized as a **transition probability**:

$$P(X_{t+1} = j \mid X_t = i) \tag{6.1}$$

where X is a random variable representing the phenotype of the cell at an arbitrary time t.[1] Intuitively, if i and j are phenotypically very similar, then there is a strong dynamic coupling between them: A cell in state i is likely to move to j and the transition probability (Eq. 6.1) is high.

It is reasonable to assume that transition probabilities obey the first-order Markov property, i.e., that transitions depend only on the current state, not on a previous state k:

$$P(X_{t+1} = j \mid X_t = i) = P(X_{t+1} = j \mid X_t = i, X_{t-1} = k) \tag{6.2}$$

and so on for states prior to k.

All entries of $\mathbf{W}$ are strictly nonnegative. If the rows of $\mathbf{W}$ are normalized such that they sum to 1, then each row $\mathbf{w}_i$ can be interpreted as a vector expressing the probability (Eq. 6.2) that a cell with phenotype X_i at time t moves to phenotype X_j at $t+1$, for all j. Thus, the right stochastic matrix:

$$M = D^{-1}\mathbf{W} \tag{6.3}$$

with $D = \mathrm{diag}(\deg(v_1), \ldots, \deg(v_N))$ (as in §3.2) is easily obtained from the weight matrix of the graph and can be subjected to numerical techniques in order to learn about the underlying cellular system.

The power of the Markov chain formalism can be illustrated with a simple example. While

[1] Here, "time" is an index of the discrete dynamical process, not a measured value as in a time course experiment.

PhenoGraph is designed to identify cells that are most similar, an alternative experimental design may seek to identify cells that are most different from each other. For example, suppose two samples are collected before and after a treatment, and the objective is to identify cells that are induced by the treatment—in other words, phenotypes that are specific to the post-treatment sample. This question can be answered in the Markov chain framework using a quantity known as the **hitting time**, H_{ij}, the expected number of steps a random walker starting at i takes before reaching j, averaged over all possible random walks emanating from i. Let A denote the set of pre-treatment cells and B the set of post-treatment cells. Consider one cell, $\beta \in B$ and the hitting times from all other cells to β. If β is unlike most pre-treatment cells then it will take a long time for random walks beginning at $a \in A$ to reach β—the hitting times from A will be large compared to the hitting times from B. Thus the divergence between the two hitting time distributions, $H_{A\beta}$ and $H_{B\beta}$, would quantify the specificity of β to the post-treatment sample. As hitting times can be computed from spectral analysis of M (Eq. 6.3) [116], placing the problem in the context of Markov chains provides a mathematical framework for extending the graph-based representation of phenotypes to answer new questions.

The approaches presented in this dissertation, which build and operate on graphs of phenotypes, can be seen as particular instances of this more general framework. The recognition of this framework is important as it provides foundations from which other extensions and specializations can be derived, suggesting a way forward for future projects in computational cell biology.

References

1. Liberali, P. & Pelkmans, L. Towards quantitative cell biology. *Nature Cell Biology* **14,** 1233–1233 (2012).
2. Shapiro, J. & Dworkin, M. *Bacteria As Multicellular Organisms* (Oxford University Press, 1997).
3. Novick, A. & Weiner, M. Enzyme Induction as an All-or-None Phenomenon. *Proceedings of the National Academy of Sciences of the United States of America* **43,** 553–566 (1957).
4. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3,** 318–356 (1961).
5. Howell, W. H. Observations upon the occurrence, structure, and function of the giant cells of the marrow. *Journal of Morphology* **4,** 117–129 (1890).
6. Holmes, A. M. The nature and significance of leucocytosis. *The Journal of the American Medical Association* **44,** 257–263 (1905).
7. Bonner, W. A., Hulett, H. R., Sweet, R. G. & Herzenberg, L. A. Fluorescence activated cell sorting. *Review of Scientific Instruments* **32,** 404–9 (1972).
8. Bobrove, A. M., Strober, S., Herzenberg, L. A. & DePamphilis, J. D. Identification and quantitation of thymus-derived lymphocytes in human peripheral blood. *Journal of Immunology* **112,** 520–527 (1974).
9. Taylor, I. W. & Milthorpe, B. K. An evaluation of DNA fluorochromes, staining techniques, and analysis for flow cytometry. I. Unperturbed cell populations. *Journal of Histochemistry & Cytochemistry* **28,** 1224–1232 (1980).
10. Köhler, G. J. F. & Milstein, C. Continuous cultures of fused cells secreting antibody of predifined specificity. *Nature* **256,** 495–7 (1975).
11. Bernard, A. & Boumsell, L. The clusters of differentiation (CD) defined by the First International Workshop on human leucocyte differentiation antigens. *Human Immunology* **11,** 1–10 (1984).
12. Rosa, S. C. D., Herzenberg, L. A., Herzenberg, L. A. & Roederer, M. 11-color, 13-parameter flow cytometry: Identification of human naive T cells by phenotype, function and T-cell receptor diversity. *Nature Medicine* **7,** 245–248 (2001).
13. Herzenberg, L. A. *et al.* The History of Future of the Fluorescence Activated Cell Sorter and Flow Cytometry: A View from Stanford. *Clinical Chemistry* **48,** 1819–1827 (2002).

14. Van Dongen, J. J. M. *et al.* EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia* **26,** 1908–1975 (2012).

15. Tanner, S. D., Baranov, V. I., Ornatsky, O. I., Bandura, D. R. & George, T. C. An introduction to mass cytometry: fundamentals and applications. *Cancer Immunology, Immunotherapy* **62,** 955–965 (2013).

16. 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. *Analytical Chemistry* **81,** 6813–6822 (2009).

17. Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science* **332,** 687–696 (2011).

18. Roederer, M. Spectral compensation for flow cytometry: Visualization artifacts, limitations, and caveats. *Cytometry* **45,** 194–205 (2001).

19. Finck, R. *et al.* Normalization of mass cytometry data with bead standards. *Cytometry A* **83,** 483–494 (2013).

20. Bodenmiller, B. *et al.* Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nature biotechnology* **30,** 858–867 (2012).

21. Becker, A. J., McCulloch, E. A. & Till, J. E. Cytological Demonstration of the Clonal Nature of Spleen Colonies Derived from Transplanted Mouse Marrow Cells. *Nature* **197,** 452–454 (1963).

22. Suda, T., Suda, J. & Ogawa, M. Single-cell origin of mouse hemopoietic colonies expressing multiple lineages in variable combinations. *Proceedings of the National Academy of Sciences of the United States of America* **80,** 6689–6693 (1983).

23. Civin, C. I., Banquerigo, M. L., Strauss, L. C. & Loken, M. R. Antigenic analysis of hematopoiesis. VI. Flow cytometric characterization of My-10-positive progenitor cells in normal human bone marrow. *Experimental Hematology* **15,** 10–17 (1987).

24. Terstappen, L. W., Huang, S., Safford, M., Lansdorp, P. M. & Loken, M. R. Sequential generations of hematopoietic colonies derived from single nonlineage-committed CD34+CD38- progenitor cells. *Blood* **77,** 1218–1227 (1991).

25. Sachs, L. The Differentiation of Myeloid Leukaemia Cells: New Possibilities for Therapy. *British Journal of Haematology* **40,** 509–517 (1978).

26. Sachs, L. Control of normal cell differentiation and the phenotypic reversion of malignancy in myeloid leukaemia. *Nature* **274,** 535–539 (1978).

27. Coombs, C. C., Tavakkoli, M. & Tallman, M. S. Acute promyelocytic leukemia: where did we start, where are we now, and the future. *Blood Cancer J* **5,** e304 (2015).

28. Nowak, D., Stewart, D. & Koeffler, H. P. Differentiation therapy of leukemia: 3 decades of development. *Blood* **113,** 3655–65 (2009).

29. Tenen, D. G., Hromas, R., Licht, J. D. & Zhang, D.-E. Transcription Factors, Normal Myeloid Development, and Leukemia. *Blood* **90,** 489–519 (1997).

30. Zhang, P. *et al.* Enhancement of Hematopoietic Stem Cell Repopulating Capacity and Self-Renewal in the Absence of the Transcription Factor C/EBPα. *Immunity* **21,** 853–863 (2004).

31. Pabst, T. & Mueller, B. U. Transcriptional dysregulation during myeloid transformation in AML. *Oncogene* **26,** 6829–6837 (2007).

32. Duprez, E., Wagner, K., Koch, H. & Tenen, D. G. C/EBPβ: a major PML-RARA-responsive gene in retinoic acid-induced differentiation of APL cells. *The EMBO Journal* **22,** 5806–5816 (2003).

33. Wouters, R. & Lowenberg, B. On the maturation order of AML cells: a distinction on the basis of self-renewal properties and immunologic phenotypes. *Blood* **63,** 684–689 (1984).

34. Lowenberg, B. & Bauman, J. G. Further results in understanding the subpopulation structure of AML: clonogenic cells and their progeny identified by differentiation markers. *Blood* **66,** 1225–1232 (1985).

35. Pessano, S. *et al.* Subpopulation heterogeneity in human acute myeloid leukemia determined by monoclonal antibodies. *Blood* **64,** 275–281 (1984).

36. Bhatia, M., Wang, J. C. Y., Kapp, U., Bonnet, D. & Dick, J. E. Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 5320–5325 (1997).

37. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Medicine* **3,** 730–737 (1997).

38. Taussig, D. C. *et al.* Anti-CD38 antibody-mediated clearance of human repopulating cells masks the heterogeneity of leukemia-initiating cells. *Blood* **112,** 568–575 (2008).

39. Taussig, D. C. *et al.* Leukemia-initiating cells from some acute myeloid leukemia patients with mutated nucleophosmin reside in the CD34(-) fraction. *Blood* **115,** 1976–1984 (2010).

40. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501,** 328–337 (2013).

41. Rhenen, A. v. *et al.* High Stem Cell Frequency in Acute Myeloid Leukemia at Diagnosis Predicts High Minimal Residual Disease and Poor Survival. *Clinical Cancer Research* **11,** 6520–6527 (2005).

42. Pearce, D. J. *et al.* AML engraftment in the NOD/SCID assay reflects the outcome of AML: implications for our understanding of the heterogeneity of AML. *Blood* **107,** 1166–1173 (2006).

43. Eppert, K. *et al.* Stem cell gene expression programs influence clinical outcome in human leukemia. *Nature Medicine* **17,** 1086–1093 (2011).

44. Pattabiraman, D. R. & Weinberg, R. A. Tackling the cancer stem cells — what challenges do they pose? *Nature Reviews Drug Discovery* **13,** 497–512 (2014).

45. Okuno, Y. *et al.* Differential regulation of the human and murine CD34 genes in hematopoietic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 6246–6251 (2002).

46. Irish, J. M. *et al.* Single Cell Profiling of Potentiated Phospho-Protein Networks in Cancer Cells. *Cell* **118,** 217–228 (2004).

47. Irish, J. M., Kotecha, N. & Nolan, G. P. Mapping normal and cancer cell signalling networks: towards single-cell proteomics. *Nature Reviews Cancer* **6,** 146–155 (2006).

48. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).

49. Bock, T., Bausch-Fluck, D., Hofmann, A. & Wollscheid, B. CD proteome and beyond - technologies for targeting the immune cell surfaceome. *Frontiers in Bioscience (Landmark Edition)* **17,** 1599–1612 (2012).

50. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).

51. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402,** C47–C52 (1999).

52. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).

53. Bendall, S. C. *et al.* Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* **157,** 714–725 (2014).

54. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518,** 331–336 (2015).

55. Rives, A. W. & Galitski, T. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 1128–1133 (2003).

56. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* **34,** 166–176 (2003).

57. Elowitz, M. B. Stochastic Gene Expression in a Single Cell. *Science* **297,** 1183–1186 (2002).

58. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* **6,** 451–464 (2005).

59. Gupta, P. B. *et al.* Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. *Cell* **146,** 633–644 (2011).

60. MacQueen, J. *Some methods for classification and analysis of multivariate observations* in. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (The Regents of the University of California, 1967). <http://projecteuclid.org/euclid.bsmsp/1200512992> (visited on 09/28/2015).

61. Gray, R. & Karnin, E. Multiple local optima in vector quantizers (Corresp.) *IEEE Trans. Inform. Theory* **28,** 256–261 (1982).

62. Rose, K. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE* **86,** 2210–2239 (1998).

63. Mahajan, M., Nimbhorkar, P. & Varadarajan, K. in *WALCOM: Algorithms and Computation* 274–285 (Springer Verlag, 2009). doi:10.1007/978-3-642-00202-1_24. <http://dx.doi.org/10.1007/978-3-642-00202-1_24>.

64. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 14863–14868 (1998).

65. Lee, J. A. & Verleysen, M. *Nonlinear dimensionality reduction* (Springer, 2007).

66. Wood, B. in, 559–576 (Academic Press, 2004).

67. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

68. Amir, E.-a. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology* **31,** 545–552 (2013).

69. Waddington, C. H. *Organisers and Genes* (Cambridge University Press, 1940).

70. Huang, S. The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology? *Bioessays* **34,** 149–157 (2011).

71. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nature methods* **10,** 228–238 (2013).

72. Boedigheimer, M. J. & Ferbas, J. Mixture modeling approach to flow cytometry data. *Cytometry* **73A,** 421–429 (2008).

73. Pyne, S. *et al.* Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 8519–8524 (2009).

74. Roberts, S. J. Parametric and non-parametric unsupervised cluster analysis. *Pattern Recognition* **30,** 261–272 (1997).

75. Aghaeepour, N., Nikolic, R., Hoos, H. H. & Brinkman, R. R. Rapid cell population identification in flow cytometry data. *Cytometry Part A* **79A,** 6–13 (2011).

76. 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. Part B, Clinical Cytometry* **78 Suppl 1,** S69–82 (2010).

77. Zare, H., Shooshtari, P., Gupta, A. & Brinkman, R. R. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* **11,** 403 (2010).

78. Von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing* **17,** 395–416 (2007).

79. Jaccard, P. *Le coefficient générique et le coefficient de communauté dans la flore marocaine* <`https://books.google.com/books?id=u9K8IwAACAAJ`> (Impr. Commerciale, 1926).

80. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 7821–7826 (2002).

81. Newman, M. E. J. Analysis of weighted networks. *Physical Review E* **70.** doi:10.1103/physreve.70.056131. <http://dx.doi.org/10.1103/PhysRevE.70.056131> (2004).

82. Reichardt, J. *Structure in Complex Networks* (Springer Verlag, 2009).

83. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008 (2008).

84. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 184–197 (2015).

85. Doyle, P. G. & Snell, L. *Random Walks and Electric Networks* (Mathematical Association of America, 1984).

86. Grady, L. Random Walks for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **28,** 1768–1783 (2006).

87. Addison, E. G. *et al.* Ligation of CD8alpha on human natural killer cells prevents activation-induced apoptosis and enhances cytolytic activity. *Immunology* **116,** 354–361 (2005).

88. Bissell, M. J. & Radisky, D. Putting tumours in context. *Nature Reviews Cancer* **1,** 46–54 (2001).

89. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer* **12,** 323–334 (2012).

90. Tenen, D. G. Disruption of differentiation in human cancer: AML shows the way. *Nature Reviews Cancer* **3,** 89–101 (2003).

91. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer Verlag, 2009).

92. Clark, P., Normansell, D. E., Innes, D. J. & Hess, C. E. Lymphocyte subsets in normal bone marrow. *Blood* **67,** 1600–1606 (1986).

93. Craig, F. E. & Foon, K. A. Flow cytometric immunophenotyping for hematologic neoplasms. *Blood* **111,** 3941–3967 (2008).

94. Becskei, A., Séraphin, B. & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *The EMBO Journal* **20,** 2528–2535 (2001).

95. Ferrell, J. E. & Machleder, E. M. The biochemical basis of an all-or-none cell fate switch in Xenopus oocytes. *Science (New York, N.Y.)* **280,** 895–898 (1998).

96. Levina, E. & Bickel, P. *The Earth Mover's distance is the Mallows distance: some insights from statistics* in. **2** (2001), 251–256 vol.2. doi:10.1109/ICCV.2001.937632.

97. Kenneth D. Gibbs, J. *et al.* Single-cell phospho-specific flow cytometric analysis demonstrates biochemical and functional heterogeneity in human hematopoietic stem and progenitor compartments. *Blood* **117,** 4226–4233 (2011).

98. Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64,** 479–498 (2002).

99. Barber, D. *Bayesian reasoning and machine learning* (Cambridge University Press, 2011).

100. Gilliland, D. G. & Griffin, J. D. The roles of FLT3 in hematopoiesis and leukemia. *Blood* **100,** 1532–1542 (2002).

101. Wandzioch, E., Edling, C. E., Palmer, R. H., Carlsson, L. & Hallberg, B. Activation of the MAP kinase pathway by c-Kit is PI-3 kinase dependent in hematopietic progenitor/stem cell lines. *Blood* **104,** 51–57 (2004).

102. Kornblau, S. M. *et al.* Signaling changes in the stem cell factor-AKT-S6 pathway in diagnostic AML samples are associated with disease relapse. *Blood Cancer Journal* **1,** e3 (2011).

103. Martelli, A. M. *et al.* Phosphoinositide 3-kinase/Akt signaling pathway and its therapeutical implications for human acute myeloid leukemia. *Leukemia* **20,** 911–928 (2006).

104. Finbloom, D. S. & Winestock, K. D. IL-10 induces the tyrosine phosphorylation of tyk2 and Jak1 and the differential assembly of STAT1 alpha and STAT3 complexes in human T cells and monocytes. *The Journal of Immunology* **155,** 1079–1090 (1995).

105. Radtke, I. *et al.* Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 12944–12949 (2009).

106. Lu, P., Nakorchevskiy, A. & Marcotte, E. M. Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 10370–10375 (2003).

107. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–15550 (2005).

108. Ross, M. E. Gene expression profiling of pediatric acute myelogenous leukemia. *Blood* **104,** 3679–3687 (2004).

109. Jaatinen, T. *et al.* Global Gene Expression Profile of Human Cord Blood-Derived CD133+ Cells. *Stem Cells* **24,** 631–641 (2006).

110. Gallacher, L. *et al.* Isolation and characterization of human CD34(neg)Lin(neg) and CD34(pos)Lin(neg) hematopoietic stem cells using cell surface markers AC133 and CD7. *Blood* **95,** 2813–2820 (2000).

111. Collins, A. T., Berry, P. A., Hyde, C., Stower, M. J. & Maitland, N. J. Prospective Identification of Tumorigenic Prostate Cancer Stem Cells. *Cancer Research* **65,** 10946–10951 (2005).

112. O'Brien, C. A., Pollett, A., Gallinger, S. & Dick, J. E. A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. *Nature* **445,** 106–110 (2007).

113. Metzeler, K. H. *et al.* An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* **112,** 4193–4201 (2008).

114. Toren, A. *et al.* CD133-positive hematopoietic stem cell "stemness" genes contain many genes mutated or abnormally expressed in leukemia. *Stem Cells (Dayton, Ohio)* **23,** 1142–1153 (2005).

115. Hosen, N. *et al.* CD96 is a leukemic stem cell-specific marker in human acute myeloid leukemia. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 11008–11013 (2007).

116. Lovász, L. Random walks on graphs: A survey. *Combinatorics* **2,** 1–46 (1993).

117. Fienberg, H. G., Simonds, E. F., Fantl, W. J., Nolan, G. P. & Bodenmiller, B. A platinum-based covalent viability reagent for single-cell mass cytometry. *Cytometry* **81A,** 467–475 (2012).

118. Akavia, U. D. *et al.* An Integrated Approach to Uncover Drivers of Cancer. *Cell* **143,** 1005–1017 (2010).

Appendix

A.1 Patient samples

Twenty (20) deidentified diagnosis bone marrow mononuclear cell (BMMC) specimens, selected to span a variety of different AML subtypes, were obtained from the tissue bank at St. Jude Children's Hospital (Memphis, TN). These samples had been analyzed for genome-wide gene expression and copy number alteration, as well as mutation status in a small set of putative oncogenes [105]. Sixteen (16) samples were included in the final pediatric AML cohort: Three samples were excluded due to insufficient cell number and one was excluded because signal from internal standard beads (see §A.3, below) did not pass quality control (QC) thresholds. The final pediatric AML cohort was 50% male with a mean age of 10.4 years. The final cohort reflected much of the interpatient heterogeneity observed in pediatric AML, and included samples with t(8;21) chromosomal translocations, inv(16) inversion, MLL rearrangements, as well as cytogenetically normal samples. Samples were classified by the French-American-British (FAB) system as M1, M2, M4, or M5, and thus reflected a broad range of histopathological categories. For all healthy adult controls, deidentified cryopreserved healthy BMMC samples were purchased from AllCells, Inc. (Emeryville, CA). The healthy adult cohort for signaling studies (n = 5, sample IDs: H1–H5) ranged in age from 19–28 years with a mean of 23.4 years. Nine deidentified adult AML BMMC specimens were obtained from Princess Margaret Hospital (Toronto, ON) and included in surface marker selection experiments only. All human samples were obtained with informed consent in compliance with IRB-approved protocols.

A.2 Design of hybrid antibody panel

A.2.1 Antibodies

Purified monoclonal antibodies were obtained from commercial vendors and labeled with stable metal isotopes using the MAXPAR^{TM} X8 chelating polymer kit (Fluidigm Corporation, South San Francisco, CA) following the manufacturer's instructions (Table A.1). Anti-cleaved caspase-3 was measured but was omitted from downstream analysis because we instead relied on cisplatin viability staining to exclude dead cells [117]. The final data set included a total of 30 antibodies against 14 intracellular targets and 16 cell-surface targets.

A.2.2 A minimal set of surface markers to capture AML heterogeneity

We sought an efficient surface marker panel of approximately 16 markers to achieve segregation of the main subsets in the AML samples, thus freeing 15 analysis channels for simultaneous measurement of intracellular epitopes. To address this, we performed an initial surface-only phenotyping experiment on a panel of 32 AML samples (9 adult and 23 pediatric, including the 16 pediatric AML samples included in the final cohort). Each sample was stained with two overlapping surface marker panels of 31 antibodies each (42 unique antibodies in total, (Table A.2). We designed a simple feature-scoring algorithm based on principal component analysis (PCA) to identify the non-redundant markers in each patient while capturing the overall diversity across the patients. It consisted of the following steps:

1. Events were gated to remove doublets ($\text{DNA}^{\text{high}}/\text{event\_length}^{\text{high}}$) and non-nucleated cells or debris ($\text{CD45}^{-}/\text{DNA}^{\text{low}}$)

2. PCA was performed on single-cell data from each patient individually

3. A **non-redundancy score** (NRS) was calculated as follows. For each marker M in each patient p,

$$\text{NRS}(M_p) = \sum_{c=1}^{C} |\text{coeff}(c)| \times \lambda(c) \tag{A.4}$$

where coeff(c) is the coefficient of marker M in component c and $\lambda(c)$ is the eigenvalue

Label	Target	Clone (Vendor)	Staining conc.
141 Pr	PLCg2 (pY759)	K86-689.37 (BD)	1
142 Nd	CD19	HIB19 (BD)	2
143 Nd	4EBP1 (pT37/46)	236B4 (CST)	3
144 Nd	CD11b	ICRF44 (BL)	4
145 Nd	AMPK (pT172)	40H9 (CST)	3
146 Nd	STAT3 (pY705)	4 (BD)	2
147 Sm	S6 (pS235/pS236)	N7-548 (BD)	1
148 Nd	CD34	8G12 (BD)	2
149 Sm	CREB (pS133)	87G3 (CST)	1
150 Nd	STAT5 (pY694)	47 (BD)	2
151 Eu	CD123	9F5 (BD)	1.5
152 Sm	c-CBL (pY700)	47/c-Cbl (BD)	1
153 Eu	STAT1 (pY701)	4a (BD)	1
154 Sm	CD45	HI30 (BL)	2
156 Gd	ZAP70/SYK (pY319/pY352)	17a (BD)	0.75
158 Gd	CD33	WM53 (BL)	1
159 Tb	AKT (pS473)	D9E (CST)	2
160 Gd	CD47	B6H12 (BD)	0.75
162 Dy	CD7	M-T701 (BD)	0.5
164 Dy	CD15	W6D3 (BL)	0.5
165 Ho	RB (pS807/pS811)	J112-906 (BD)	0.4
166 Er	CD44	G44-26 (BD)	0.15
167 Er	CD38	HIT2 (BL)	1
168 Er	ERK1/2 (p44/42) (pT202/pY204)	20A (BD)	1
169 Tm	P38 (pT180/ pY182)	36/p38 (BD)	0.5
170 Er	CD3	UCHT1 (BL)	0.75
171 Yb	CD117	104D2 (BL)	0.5
172 Yb	Caspase-3 (active)	C92-605 (BD)	0.5
174 Yb	HLA-DR	L243 (BL)	1
175 Lu	CD41	HIP8 (BL)	0.2
176 Yb	CD64	10.1 (BL)	2

Table A.1: Mass cytometry staining panel used for main experiments. BL: Biolegend; BD: BD Bioscences; CST: Cell Signaling Technologies. Staining concentration is given in μg/mL.

of component c. The score considers the first C principal components, in descending order of $\lambda(c)$.

The NRS was calculated for each of the 42 surface markers in the 36 bone marrow samples using the first 3 principal components (Table A.2). The 16 top-scoring markers were selected for carrying forward to future experiments, with the exception of CD2 and CD11c, which were manually excluded on the basis that they are expressed on mature lymphocytes and monocytes, respectively, and therefore not likely to be essential for discriminating AML blast subsets. Two surface markers were manually selected for inclusion in future experiments, despite the fact that they did not appear in the top 16 scores—CD117 and CD19. These were selected based on the following rationale: CD117 is a hematopoietic progenitor marker, and CD19 is a marker of B cells, which were otherwise difficult to identify using only the top-scoring markers. The final set of 16 markers carried forward for future experiments was: CD3, CD7, CD11b, CD15, CD19, CD33, CD34, CD38, CD41, CD44, CD45, CD64, CD47, CD117, CD123, and HLA-DR.

A.3 Mass cytometry data collection and preprocessing

The protocols for surface and intracellular antibody staining, *ex vivo* molecular stimulation, and cell barcoding are described elsewhere [20, 84].

Mass cytometry data were acquired on the CyTOF$^{\text{TM}}$ mass cytometer as previously described [17]. Raw mass cytometry data was extracted into listmode FCS files using CyTOF Instrument Control Software version 5.1.451 (DVS Sciences, Sunnyvale, CA) using default parameters except for the following: The instrument dual-count slopes were recalibrated weekly using solution-based standards and "Instrument" dual-count calibration was used for FCS file extraction; cell events with event_length values between 10 and 65 were extracted.

Machine sensitivity was monitored using polystyrene internal standard beads containing 5 embedded lanthanide elements (139La, 141Pr, 159Tm, 169Tb, 175Lu) (a gift from Scott Tanner, University of Toronto). Beads were spiked into the cell suspension immediately before measurement at approximately $2x10^4$ beads/mL. To facilitate quantitative com-

Rank	Target	NRS1	NRS2	Max. NRS	Included in hybrid panel?
1	HLA-DR	22.5	20.5	22.5	Yes (high score)
2	CD34	19.3	21.4	21.4	Yes (high score)
3	CD64	–	17.0	17.0	Yes (high score)
4	CD44	15.6	16.7	16.7	Yes (high score)
5	CD15	16.5	15.5	16.5	Yes (high score)
6	CD33	15.8	15.6	15.8	Yes (high score)
7	CD45	13.7	14.5	14.5	Yes (high score)
8	CD38	14.1	11.2	14.1	Yes (high score)
9	CD11b	13.2	–	13.2	Yes (high score)
10	CD3	13.1	12.9	13.1	Yes (high score)
11	CD7	12.9	12.4	12.9	Yes (high score)
12	CD41	12.3	–	12.3	Yes (high score)
13	CD2	–	12.2	12.2	No (T / NK marker)
14	CD123	11.8	10.5	11.8	Yes (high score)
15	CD11c	11.7	–	11.7	No (monocyte marker)
16	CD47	10.6	10.3	10.6	Yes (high score)
17	CD8	10.4	–	10.4	No (low score)
18	CD49d	–	10.1	10.1	No (low score)
19	CD117	9.7	9.7	9.7	Yes (HSPC marker)
20	CD14	9.6	7.9	9.6	No (low score)
21	CD5	–	9.1	9.1	No (low score)
22	CD45RA	8.3	6.6	8.3	No (low score)
23	CD4	8.1	–	8.1	No (low score)
24	CD16	8.0	5.7	8.0	No (low score)
25	CD13	6.9	–	6.9	No (low score)
26	CD61	6.6	–	6.6	No (low score)
27	CD184 (CXCR4)	6.6	5.7	6.6	No (low score)
28	CD133	–	6.1	6.1	No (low score)
29	CD235a/b	6.0	–	6.0	No (low score)
30	CD22	–	5.6	5.6	No (low score)
31	CD135 (Flt3)	–	4.8	4.8	No (low score)
32	CD20	4.5	3.4	4.5	No (low score)
33	CD19	3.1	4.5	4.5	Yes (B cell marker)
34	IgM	4.2	–	4.2	No (low score)
35	CD161	3.9	–	3.9	No (low score)
36	TIM3	–	3.9	3.9	No (low score)
37	CD10	3.9	–	3.9	No (low score)
38	IgD	–	3.6	3.6	No (low score)
39	CD90	3.5	2.0	3.5	No (low score)
40	CD114	–	3.4	3.4	No (low score)
41	CD56	2.5	2.4	2.5	No (low score)
42	CD79b	–	2.1	2.1	No (low score)

Table A.2: Antibodies included in two pilot studies to determine a minimal set of informative markers for use in the hybrid panel. The NRS (Eq. 3) was calculated for each marker and averaged over the 36 bone marrow samples used for this pilot study. NRS1 and NRS2 refer to the two overlapping panels.

parisons between data acquired on different days, single-cell data was normalized as previously described [19]. Bead-normalized data are publicly available for download at `http://cytobank.org/nolanlab/reports`.

Bead-normalized single-cell measurement intensities were transformed using the hyperbolic inverse sine with cofactor 5, as previously described [17]. To remove dead cells and debris, cells were gated based on event_length, nucleic acid staining, and cisplatin as described previously [117].

For analysis of the surface markers we performed further normalization to: (1) Facilitate comparison between patients, as these were each collected in a different tube; (2) Better equalize the contribution of each surface protein to the clustering and dimension reduction solutions. Different antibodies have varied dynamic ranges that do not necessarily reflect the physical dynamic range or the marker's importance and markers with larger dynamic range can have a disproportionate influence on the clustering solution. Therefore we chose to rescale marker intensities across the surface panel.

As proteins can be highly overexpressed in cancer, and this overexpression is often of biological significance, we used the healthy bone marrow samples in our data as the standard for normalization. For each surface marker, the maximum intensity observed in healthy samples was determined as the 99.5th percentile of the 3×10^6 healthy bone marrow cells from the 5 donors. The top half percentile was excluded from this determination because mass cytometry data can have high-intensity outliers. Data from all samples (healthy and AML) were divided by these psuedo-maximum values, yielding expression values that can be interpreted as x-fold of the maximum expression observed in healthy. As a result, intensity values for different antibodies were placed in more commensurate dynamic ranges (largely falling between 0 and 1), and expression in AML samples exceeding 1 can be considered as fold-change above the maximum expression observed in normal bone marrow. Because only surface marker intensities were directly compared across samples, only these channels were normalized in this manner.

A.4 Microarray data and normalization

Matched gene expression profiles for our AML patients [105] were downloaded from the Gene Expression Omnibus (GEO; ID # GSE14471). This data consisted of gene expression measured with Affymetrix U133A arrays. Gene expression and survival data for 242 cytogenetically normal adult AML patients from two independent cohorts [113] were downloaded from GEO (ID # GSE12417). This data set consisted of arrays from two different Affymetrix platforms (U133A and U133 Plus 2.0).

All microarray data were processed and normalized as described previously [118]. Of the 19291 probe sets on these arrays, 7604 were removed for low intensity (defined as being below 7 on $\log_2$ scale in at least 12 of the 16 arrays). Probe sets targeting the same gene whose measurements were well correlated ($r > .75$) were averaged to produce consensus expression values for 8196 unique genes. Additionally, 286 genes from the X and Y chromosomes were excluded. Each array was normalized by dividing the $\log_2$ intensities by the third quartile of the array. Principal component analysis of the arrays before filtering or normalization identified the array for patient SJ12 as an outlier; this patient was excluded from all gene expression analyses.