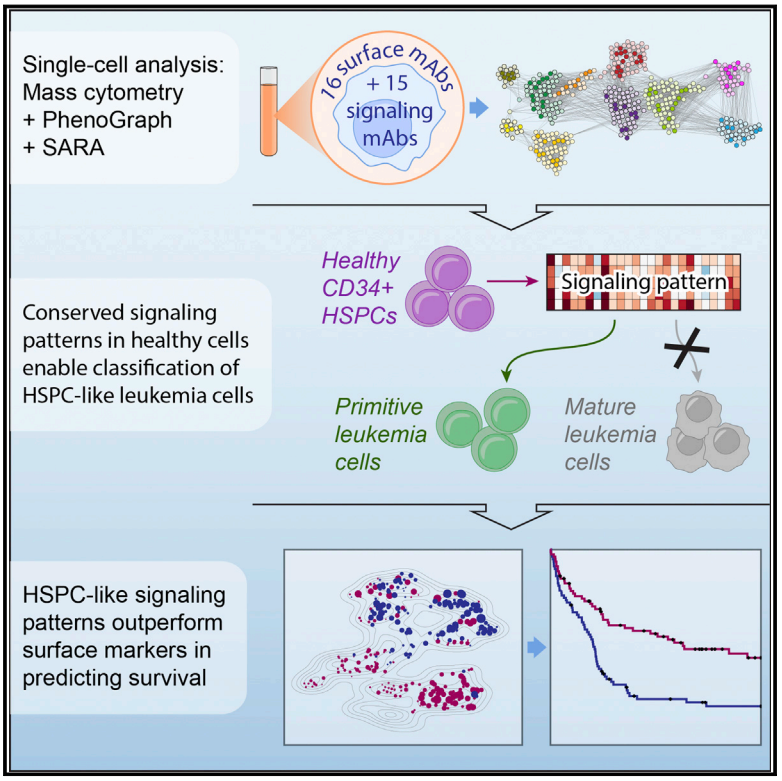
Resource

Cell

Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis

Graphical Abstract



Highlights

- PhenoGraph partitions high-dimensional single-cell data into subpopulations
- PhenoGraph plus mass cytometry elucidate intra- and intertumor heterogeneity in AML
- Surface phenotypes and regulatory intercellular signaling are decoupled in leukemia
- Signaling-based definition of primitive cells correlates with clinical outcome

Authors

Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, ..., James R. Downing, Dana Pe'er, Garry P. Nolan

Correspondence

dpeer@biology.columbia.edu (D.P.), gnolan@stanford.edu (G.P.N.)

In Brief

The PhenoGraph algorithm robustly partitions high-parameter single-cell data into phenotypically distinct subpopulations, aiding the study of complex tissues and disease cohorts. Applying PhenoGraph to a pediatric acute myeloid leukemia dataset revealed a recurrent population of leukemic cells with variable cell surface markers, but consistent signaling dynamics that mimicked normal hematopoietic progenitors.

Levine et al., 2015, Cell *162*, 184–197
July 2, 2015
http://dx.doi.org/10.1016/j.cell.2015.05.047

CellPress

Cell

Resource

Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis

Jacob H. Levine,[1,5] Erin F. Simonds,[2,5] Sean C. Bendall,[3,5] Kara L. Davis,[2] El-ad D. Amir,[1] Michelle D. Tadmor,[1] Oren Litvin,[1] Harris G. Fienberg,[2] Astraea Jager,[2] Eli R. Zunder,[2] Rachel Finck,[2] Amanda L. Gedman,[4] Ina Radtke,[4] James R. Downing,[4] Dana Pe'er,[1,6,*] and Garry P. Nolan[2,6,*]

[1]Departments of Biological Sciences and Systems Biology, Columbia University, New York, NY 10027, USA

[2]Baxter Laboratory in Stem Cell Biology, Department of Microbiology and Immunology, Stanford University, Stanford, CA 94305, USA

[3]Department of Pathology, Stanford University, Stanford, CA 94305, USA

[4]Department of Pathology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA

[5]Co-first author

[6]Co-senior author

*Correspondence: dpeer@biology.columbia.edu (D.P.), gnolan@stanford.edu (G.P.N.)

http://dx.doi.org/10.1016/j.cell.2015.05.047

SUMMARY

Acute myeloid leukemia (AML) manifests as phenotypically and functionally diverse cells, often within the same patient. Intratumor phenotypic and functional heterogeneity have been linked primarily by physical sorting experiments, which assume that functionally distinct subpopulations can be prospectively isolated by surface phenotypes. This assumption has proven problematic, and we therefore developed a data-driven approach. Using mass cytometry, we profiled surface and intracellular signaling proteins simultaneously in millions of healthy and leukemic cells. We developed PhenoGraph, which algorithmically defines phenotypes in high-dimensional single-cell data. PhenoGraph revealed that the surface phenotypes of leukemic blasts do not necessarily reflect their intracellular state. Using hematopoietic progenitors, we defined a signaling-based measure of cellular phenotype, which led to isolation of a gene expression signature that was predictive of survival in independent cohorts. This study presents new methods for large-scale analysis of single-cell heterogeneity and demonstrates their utility, yielding insights into AML pathophysiology.

INTRODUCTION

Intratumor heterogeneity is accepted to be functionally and clinically significant (Marusyk et al., 2012). Recent evidence implies that the pathobiology of cancer results from the actions and interactions of diverse subpopulations within the tumor. Thus, it is necessary to study tumors with methods that preserve single-cell resolution. Emerging technologies such as mass cytometry (Bendall et al., 2011) and single-cell RNA-seq (Patel et al., 2014) have attained dramatic increases in dimensionality and throughput, bringing unprecedented resolution to the diversity of cellular states detectable in a given tissue. Yet, to take advantage of these technological gains, computational methods are required to robustly identify high-dimensional phenotypes and compare them within and between individuals. Data-driven phenotypic dissection may then form the basis for downstream analyses in which subpopulations are isolated and compared, revealing the role of population structure in complex systems such as malignancies.

Intratumor heterogeneity is pervasive in acute myeloid leukemia (AML), an aggressive liquid tumor of the bone marrow characterized by an overwhelming abundance of poorly differentiated myeloid cells ("blasts"). Arising from the disruption of regulated myeloid differentiation (Tenen, 2003), AML results in a disordered developmental hierarchy wherein leukemic stem cells (LSCs) are capable of re-establishing the disease in immunodeficient mice (Bonnet and Dick, 1997). LSCs were first thought to be restricted to the same $CD34^+/CD38^-$ cellular compartment as normal hematopoietic stem cells (HSCs). Subsequent studies have demonstrated that both $CD38^+$ (Taussig et al., 2008) and $CD34^-$ (Taussig et al., 2010) AML blasts can have LSC capacity, indicating that AML does not follow the hierarchy of normal hematopoiesis. While AML exhibits a differentiated hierarchy, no uniform phenotypic identifier for LSCs has been found across patients (Eppert et al., 2011).

Recognizing a disconnect between functionally primitive (e.g., tumor-initiating) cells associated with cancer persistence and their surface phenotype, we simultaneously examined surface antigen expression and regulatory signaling in individual AML cells. We reasoned that intracellular signaling rather than surface antigen profile more accurately represents the functional state of a diseased cell. We used mass cytometry to measure protein expression and activation state in millions of cells from AML patients and healthy bone marrow donors in 31 simultaneous dimensions. By measuring cells after ex vivo perturbations, we further expanded the dimensionality of the data by revealing

functional responses to environmental cues, reflecting the broader cellular network beyond what can be inferred from the unperturbed state (Irish et al., 2004). To avoid the pitfalls of manual gating, we developed PhenoGraph, a robust computational method that partitions high-dimensional single-cell data into subpopulations. Building on these subpopulations, we developed additional methods to extract high-dimensional signaling phenotypes and infer differences in functional potential between subpopulations.

Our data-driven approach revealed two new perspectives on the pathobiology of AML. First, we found that pediatric AML draws from a surprisingly limited repertoire of surface phenotypes, indicating some memory of normal myelopoiesis. Despite genetic diversity, patterns of surface antigen expression followed trends in myeloid development, indicating limits in the ability of leukemic cells to phenotypically diverge from normal antigen profiles. Second, we found that the signaling pattern of undifferentiated hematopoietic progenitors defined a primitive signaling phenotype that was recapitulated in a majority of AML samples at varying frequencies. Functionally primitive leukemic cells—defined by signaling—were not linked to a consistent surface phenotype, including the standard HSC/LSC antigen profile (i.e., $CD34^{+}/CD38^{-}$), demonstrating that surface antigens are decoupled from regulatory networks in leukemia. The frequency of these functionally primitive cells enabled isolation of a gene expression signature that was enriched for stem cell annotations and formed a significant predictor of overall survival in independent AML clinical cohorts.

Taken together, we provide an alternative paradigm for identifying primitive cancer cells that complements the immunophenotypic definitions of cancer stem cells traditionally used in both AML and other systems. Moreover, this analysis framework is robust and broadly applicable to the characterization of subpopulation structure and function from single-cell data in a wide range of systems.

RESULTS

High-Dimensional Single-Cell Profiling of Pediatric AML by Mass Cytometry

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 (Supplemental Experimental Procedures). 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; Table S1) to elicit functionally relevant signaling responses (Bendall et al., 2011; Irish et al., 2004). The complete dataset contained over 15 million single cells from 21 individuals measured in 31 simultaneous protein epitope dimensions following exposure to one of 17 conditions (Figure 1A).

PhenoGraph Dissects Population Structure in High-Dimensional Single-Cell Data

Complex tissues such as bone marrow are composed of biologically meaningful subpopulations that are phenotypically coherent despite the intrinsic variability that makes each cell unique. A fundamental challenge is to establish the major phenotypes present, enabling an efficient and meaningful profile of the tissue. 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. Thus a data-driven, unsupervised approach is needed that takes single-cell measurements and returns a grouping of cells into distinct subpopulations (i.e., clusters).

Dimensionality reduction techniques such as *t*-distributed stochastic neighbor embedding (*t*-SNE) (Amir et al., 2013; Maaten and Hinton, 2008) help visualize the data but do not explicitly identify and partition cells into subpopulations. Moreover, not all subpopulations are visually distinct when rendering high-dimensional data in only two dimensions. We evaluated a number of leading methods for clustering fluorescence cytometry data and found that these did not perform well for mass cytometry data (Aghaeepour et al., 2013). Parametric methods (Pyne et al., 2009) require strong assumptions about the high-dimensional shape of cellular populations (e.g., ellipsoid, convex), which are violated in single-cell data (Amir et al., 2013). Therefore a non-parametric approach is needed, yet these currently use unstable heuristics or suffer from computational inefficiency and do not scale well to higher dimensions. We found that as the number of dimensions increased, available methods routinely failed to correctly identify known subsets, gave inconsistent results and were prohibitively slow (Supplemental Experimental Procedures).

To robustly discover subpopulations in high-dimensional single-cell data, we developed PhenoGraph. The parameters measured for each cell define a point in high-dimensional space wherein clustering is tantamount to finding dense regions. The difficulty is that density detection in high dimensions is both computationally hard and statistically unstable. Following our previous work (Bendall et al., 2014), we model this high-dimensional space using a nearest-neighbor graph. In this graph, each cell is represented by a node and connected by a set of edges to a neighborhood of its most similar cells. The graph distills the high-dimensional distribution of single cells into a compact, information-rich data structure that captures phenotypic relatedness and overcomes many pitfalls of standard geometries.

After the nearest-neighbor graph is constructed, the problem of density detection corresponds to the task of finding sets of highly interconnected nodes (Figure 1B). To this end, we borrow from the social network field, which has developed powerful algorithms to partition large social networks into communities (Girvan and Newman, 2002). 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. Community

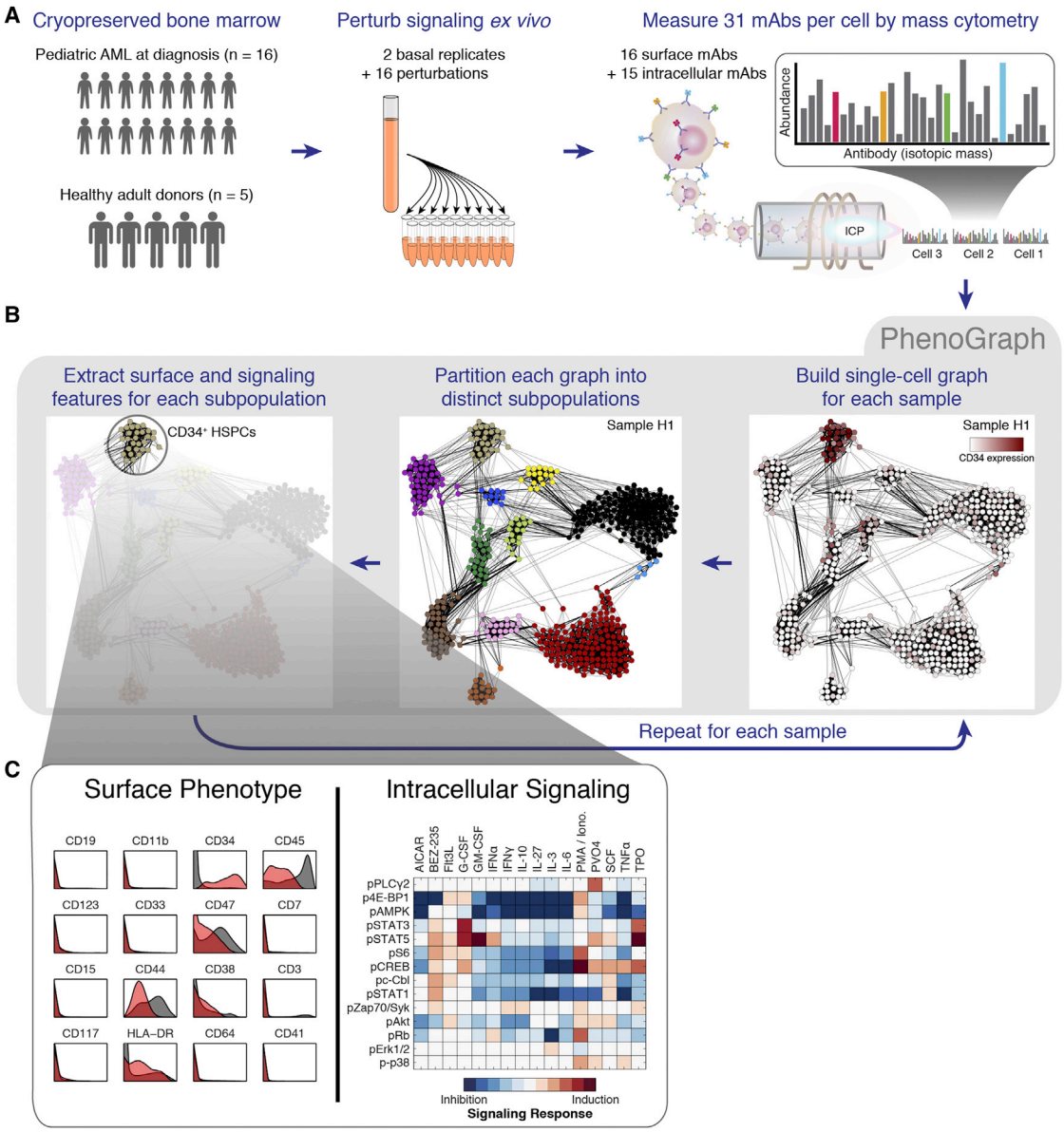


Figure 1. Mass Cytometry Analysis of Signaling Responses in Pediatric Acute Myeloid Leukemia

(A) Summary of experimental design.

(B) PhenoGraph method for clustering high-dimensional single-cell data. Each node in the neighbor graph represents one of 500 random cells from healthy donor H1 colored by CD34 expression. $CD34^+$ HSPCs form a dense subgraph and are automatically assigned to a single subpopulation. See Figure S1 and Experimental Procedures for more details on the PhenoGraph algorithm.

(C) HSPCs identified by PhenoGraph from donor H1. This subpopulation (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.

detection algorithms make no assumption about the size, number, or form of subpopulations (Fortunato, 2010). Importantly, communities need not be convex, symmetric, or ellipsoid—assumptions that are questionable for complex cellular populations. Efficient implementations can partition large graphs in minimal computation time (Blondel et al., 2008).

A key step in the PhenoGraph method is converting the single-cell data to a graph that faithfully represents the phenotypic relationships between cells. Without a carefully constructed graph, large populations can obscure rare ones (which may be outnumbered by orders of magnitude). This problem is further exacerbated by measurement noise that can spuriously link unrelated parts of the graph. We addressed both problems by constructing the graph in two iterations, using the Jaccard similarity coefficient in the second iteration. Thus, the similarity between cells is redefined by the number of shared neighbors

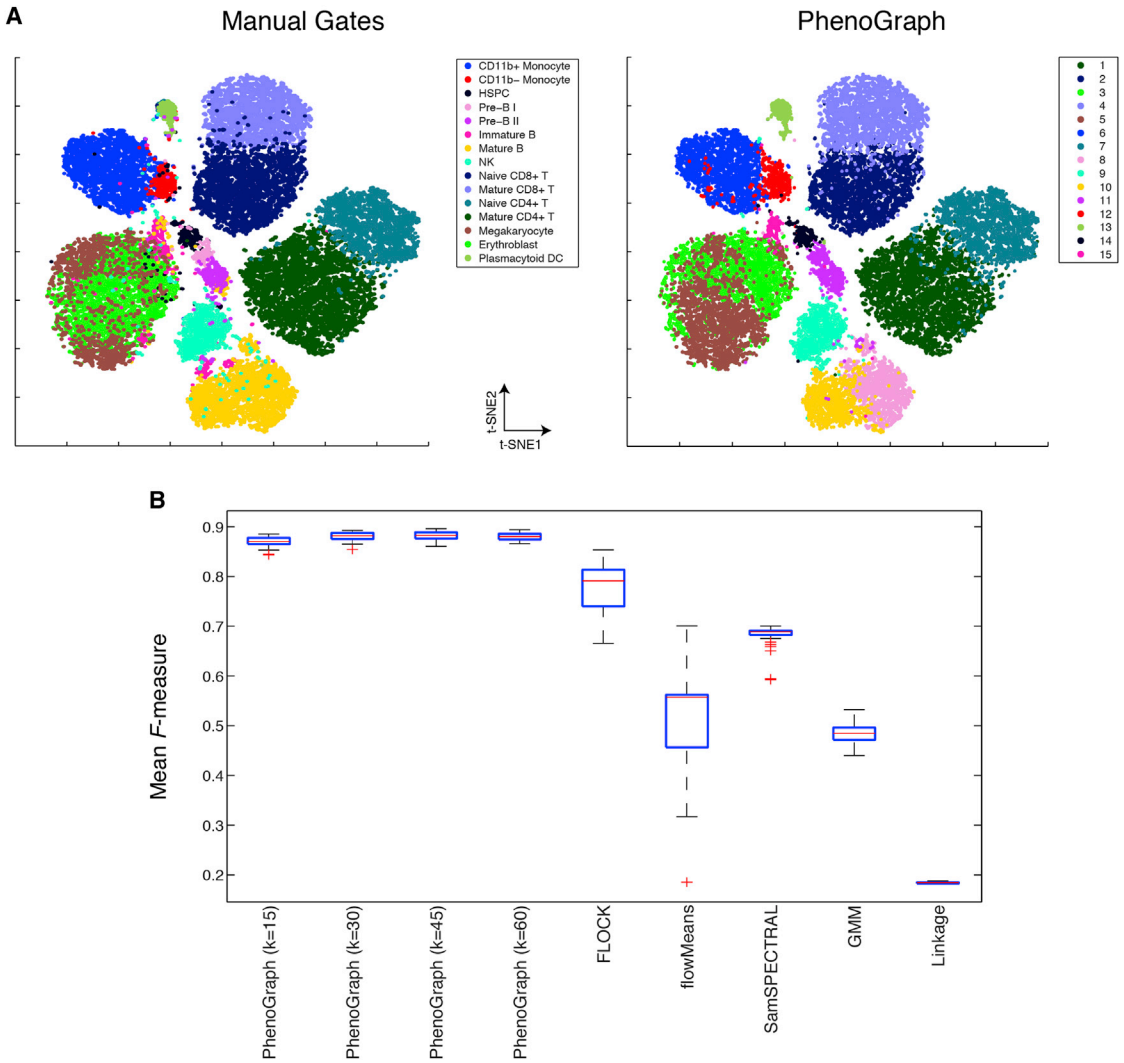


Figure 2. PhenoGraph Clustering Recapitulates Manual Assignments of Healthy Immune Cells

(A) viSNE (Amir et al., 2013) display of 30,000 cells from healthy BMMC benchmark data (Bendall et al., 2011). Cells are colored by cell-type assignments established by manual gating (left) or subpopulations detected by PhenoGraph (right).

(B) Comparison of PhenoGraph to other methods on the benchmark dataset, assessed for ability to recover the manual cell-type assignments quantified using the *F*-measure statistic (Aghaeepour et al., 2013) and normalized mutual information (Figure S2C). Box plots (generated by MATLAB's "boxplot" function) show the distributions of *F*-measure computed from 50 random samples of 20,000 cells from the full dataset. PhenoGraph was tested with four different settings of its single parameter *k*. Small interquartile ranges demonstrate that PhenoGraph accurately identifies the structure of the original population and is robust to random subsampling and to choice of parameter *k*. Comparison on additional benchmark datasets is provided in Data S1G–S1I.

following the first iteration (see Experimental Procedures and Figure S1). The Jaccard metric exploits the local density at each data point, removing spurious edges and strengthening well-supported ones. The co-occurrence of rare cells in the same phenotypic vicinity produces strongly interconnected modules that distinguish these rare cells from noise. Overall, the modular nature of the population is better revealed in the resulting graph.

Healthy human bone marrow, which is rich in distinct and well-characterized immunological cell types, presents a benchmark case for phenotypic dissection. We tested PhenoGraph on three different mass cytometry datasets of healthy human bone marrow (Bendall et al., 2011) and PhenoGraph correctly identified labeled immune cell types, displaying superior precision, recall, and robustness against leading methods (Aghaeepour et al., 2013) (Supplemental Experimental Procedures and Figures 2 and S2A–S2C, and Data S1). PhenoGraph runs efficiently on large datasets with substantially better scaling than other methods (Figure S2D) and can process millions of cells with modest computational resources. PhenoGraph is able to resolve subpopulations as rare as 1/2,000 cells and is robust to random subsampling and to the choice of the single user-defined parameter (Figures 2 and S2A–S2C, and Data S1).

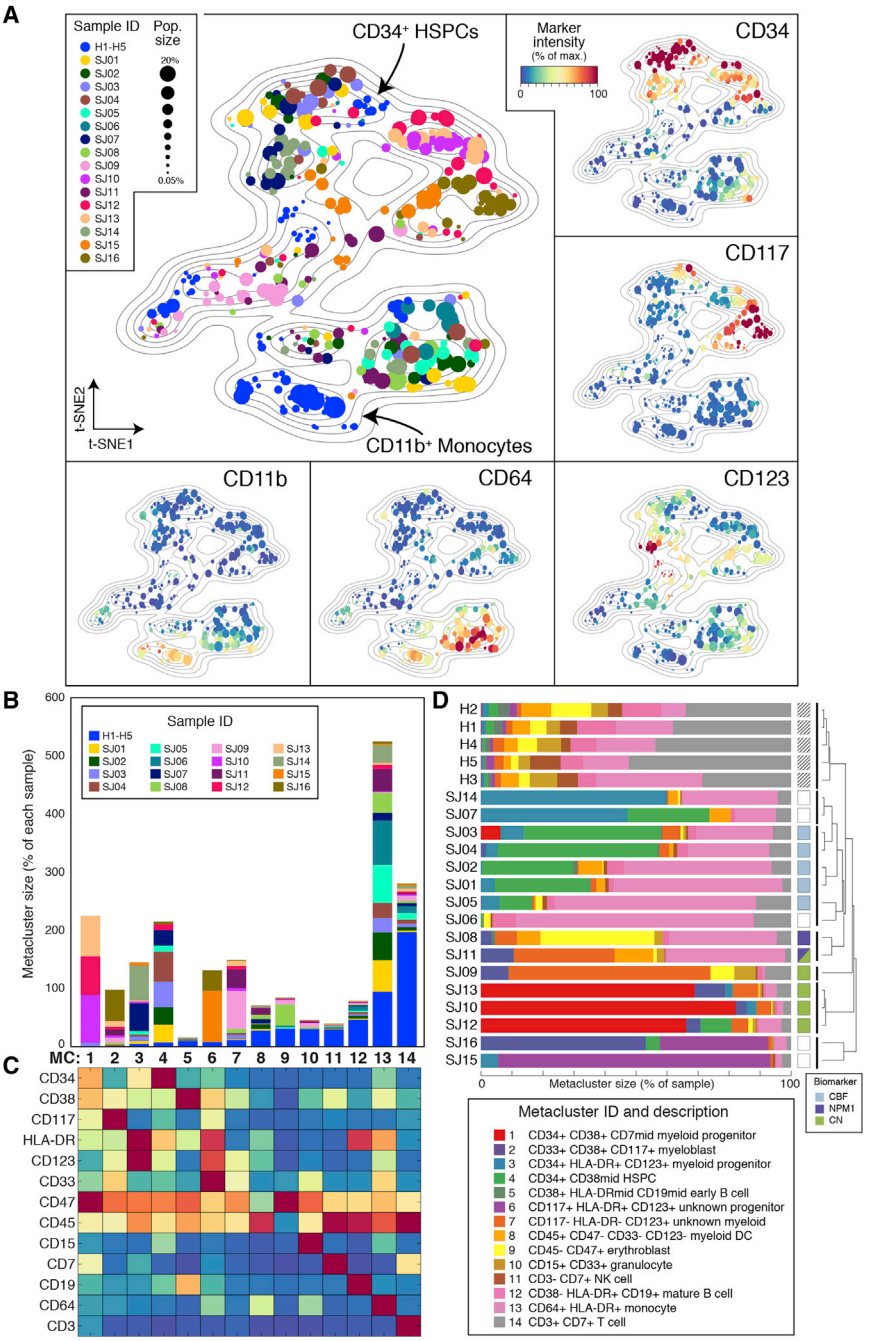


Figure 3. Intra- and Intertumor Heterogeneity Is Visible across the Phenotypic Landscape of Pediatric AML

(A) *t*-SNE landscape of average surface marker expression of non-lymphoid PhenoGraph clusters from the AML cohort. Each cluster is represented by a single point scaled to represent its sample proportion and in the main panel colored by patient identity. Normal bone marrow cell types (H1-H5; blue) provide landmarks for interpreting the phenotypes of the leukemic bone marrow samples (SJ01-SJ16). In additional panels each subpopulation is colored by median expression of indicated surface markers.

(B) PhenoGraph applied to cluster centroids consolidated the 616 patient-level subpopulations into 14 cohort-level metaclusters (MCs). Stacked columns indicate the contribution made by each patient to each MC.

(C) Average surface marker expression in each MC, summarizing the major phenotypes observed across the cohort. Columns match those represented in (B).

(D) Intrapatient heterogeneity for each patient is represented graphically by a horizontal bar in which segment lengths represent the proportion of the patient assigned to each MC, colored according to the accompanying legend (bottom right). Hierarchical clustering of these patient descriptions revealed that some patterns of intrapatient heterogeneity were significantly correlated with genetic biomarkers. (CBF, core binding transcription factor translocation, p = 0.0014; NPM1, nucleophosmin mutation, p = 0.0083; CN, cytogenetically normal, p = 0.018).

Conformity of Phenotypes in the Landscape of AML

After validating PhenoGraph on healthy cells, we applied it to our pediatric AML cohort. We ran PhenoGraph on each sample individually, defining subpopulations based on the 16 measured surface markers. This yielded an average of 28 subpopulations per sample (ranging between 17 and 48), totaling 616 subpopulations across the entire cohort. Subpopulation size varied by orders of magnitude, from 7×10^2 to 2×10^5 cells (or 0.06% to 20% of a sample). For each sample, we pooled data from all conditions, enabling characterization of subpopulation-specific signaling patterns. 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 1C).

Each leukemia presented a diversity of surface phenotypes defined by distinct combinations of marker expression (Data S2A). We sought an overview of the similarities and differences between detected subpopulations across patients that could reveal larger trends and enable direct comparison of all subpopulations simultaneously. To do so, we began by representing each PhenoGraph subpopulation by its surface marker centroid. We then used *t*-SNE (Maaten and Hinton, 2008), to reduce the 16-dimensional data to 2 dimensions, following an approach previously taken with cytometry data (Amir et al., 2013). The resulting 2D landscape provided an intuitive and comprehensive overview of the major phenotypes present in the cohort and also demonstrated the extent of intra- and intertumoral heterogeneity or similarity (Figure 3A). 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 (Supplemental Experimental Procedures) 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 (Figure 3A and Data S2B and S2C). Healthy $CD34^{+}/CD38^{mid}$ hematopoietic stem and progenitor cells (HSPCs) provided the most primitive landmark, located at the top of the landscape plot. 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 (Data S2B and S2C). 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 (Data S2D).

Metaclusters Highlight Interpatient Similarity

Despite the widespread phenotypic diversity observed within patients (Data S2E), the cohort landscape revealed a surprising conformity when comparing AML subpopulations across different patients. Multiple patients occupied each phenotypic region in the landscape and no patient presented a substantially unique phenotype, suggesting that subpopulations could be matched across patients, cohort-wide. To examine these cohort-level phenotypes further, we pursued a metaclustering approach in which subpopulations from each patient were merged by a secondary clustering analysis (Pyne et al., 2009). We represented each AML subpopulation by its centroid and used PhenoGraph to group centroids into metaclusters (MCs; see Experimental Proceduresand Figure S3A), identifying 14 MCs that delineated the major cohort phenotypes (Figure 3B-C). Each MC had a mixed patient composition, containing subpopulations from at least 2 patients and a median of 11 patients.

To evaluate the robustness of these MCs, we performed cross-validation and observed high reproducibility (Figure S3B and Supplemental Experimental Procedures). Subsequently, we used the healthy samples (H1–H5) to interpret the MCs by systematically matching cells from healthy bone marrow with the MC surface marker profiles (Supplemental Experimental Procedures). Several MCs corresponded clearly to non-malignant cell types (constituting a small proportion of each leukemic sample), while the remaining MCs represented presumptive blast phenotypes. We determined that 7/14 MCs represented malignant expansions (MC 1–4, 6, 7, 13), based on the relative frequency of healthy cognates (Figure 3B) and surface marker profiles (Figure 3C). As expected from the histopathology of AML, the blast phenotypes resembled normal primitive and progenitor phenotypes with a myeloid bias. Each malignant phenotype was detected in multiple patients, but only MC13 was detected in all patients. The $CD64^{+}/HLA\text{-}DR^{+}$ expression profile of MC13 indicates an immature monocytic phenotype that was often drastically more abundant in AML than in healthy samples. 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 (Tenen, 2003).

Samples were evaluated quantitatively in terms of their proportional occupancies of the 14 MCs (Figure 3D). As expected, the five 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 (Figure 3D). 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 three 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.

Signaling Phenotypes Define Functionally Distinct Subpopulations

Surface markers have become standard tools for clinical diagnosis and monitoring of blood neoplasia (Craig and Foon, 2008). In normal bone marrow, cell surface markers identify stem and progenitor cell populations with distinct lineage potential and intracellular signaling behaviors (Bendall et al., 2011). However, in AML, no surface marker phenotype has been established that consistently distinguishes the more primitive blasts universally across patients (Eppert et al., 2011; Taussig et al., 2008; 2010).

We hypothesized that intracellular signaling might be a better surrogate of the underlying functional potential and therefore included molecular perturbations known to elicit signaling responses that are functionally relevant to normal and malignant hematopoiesis (Table S1). Intracellular signaling markers were selected to represent pathways known to be functionally and clinically relevant in AML, including JAK/STAT, PI3K/AKT, and MAPK. The response of each of 14 signaling proteins to each of 16 perturbations revealed a facet of the underlying network that controls cellular function, resulting in 224 signaling responses per subpopulation. We used these data to build a quantitative signaling phenotype representing the structure and function of the intracellular signaling network in each subpopulation.

To fully harness the single-cell nature of our data, we developed SARA (Statistical Analysis of Perturbation Response; see

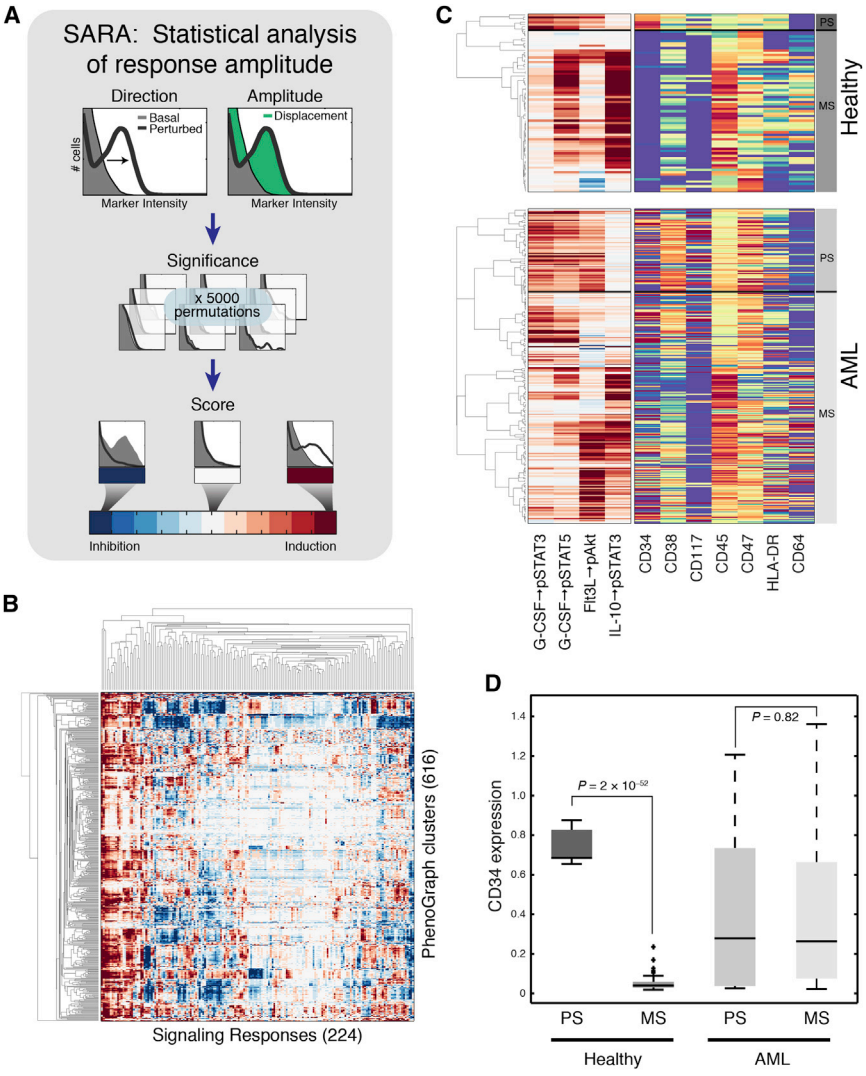


Figure 4. Analysis of Perturbation Response Generates Signaling Phenotypes

(A) An illustration depicting how SARA uses the single-cell distributions together with permutation testing to score signaling response.

(B) SARA, applied to every signaling molecule for every perturbation in every subpopulation, produced ~138,000 responses, which were compiled into 224-dimensional signaling phenotypes (columns) for each of the 616 subpopulations (rows). Rows and columns ordered by agglomerative linkage.

(C) Hierarchical clustering of four developmentally relevant signaling responses in the healthy samples (top) identified patterns of primitive signaling (PS) and mature signaling (MS) correlated with expression of CD34 and CD45, in the healthy samples. Hierarchical clustering of the same signaling responses in the AML samples (bottom) identified a cluster of subpopulations that recapitulated the primitive signaling pattern, but lacked a consistent surface phenotype. Color scales are as in Figures 3A and 4A.

(D) Box plots comparing CD34 expression between signaling clusters identified in (C). CD34 expression was significantly associated with primitive signaling only in the healthy samples. The box plots are generated by MATLAB's "boxplot" function, using default settings.

Experimental Procedures and Figure 4A). SARA examines the entire single-cell distribution of phosphoprotein intensities to detect meaningful changes between two conditions. SARA incorporates a measure of statistical significance through permutation testing, producing estimates that are sensitive to small responsive subsets yet robust to sampling error and noise. Together, PhenoGraph and SARA distilled high-dimensional data for 15 million cells into a single matrix of subpopulations and their signaling phenotypes (Figure 4B), revealing a rich variety of signaling potential across subpopulations.

Within the healthy samples, surface and signaling phenotypes were tightly coupled, consistent with previous reports (Bendall et al., 2011; Gibbs et al., 2011). 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 (Figures 4C and 4D; $p = 2.0 \times 10^{-52}$, 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; Figure 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 (Eppert et al., 2011; Gibbs et al., 2011; Taussig et al., 2008; 2010). We therefore sought to use signaling phenotypes rather than surface phenotypes as alternative proxies for functional state.

Classification of Leukemic Maturity by Signaling Phenotype

PhenoGraph and SARA yielded two alternative representations for each subpopulation: a 16-dimensional surface phenotype and a 224-dimensional signaling phenotype (Figure 5A). 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 would more faithfully capture 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. PhenoGraph metaclustering of the five normal marrow samples identified 20 healthy cell types (Figure 5B and Data S3A). Using ANOVA, we examined

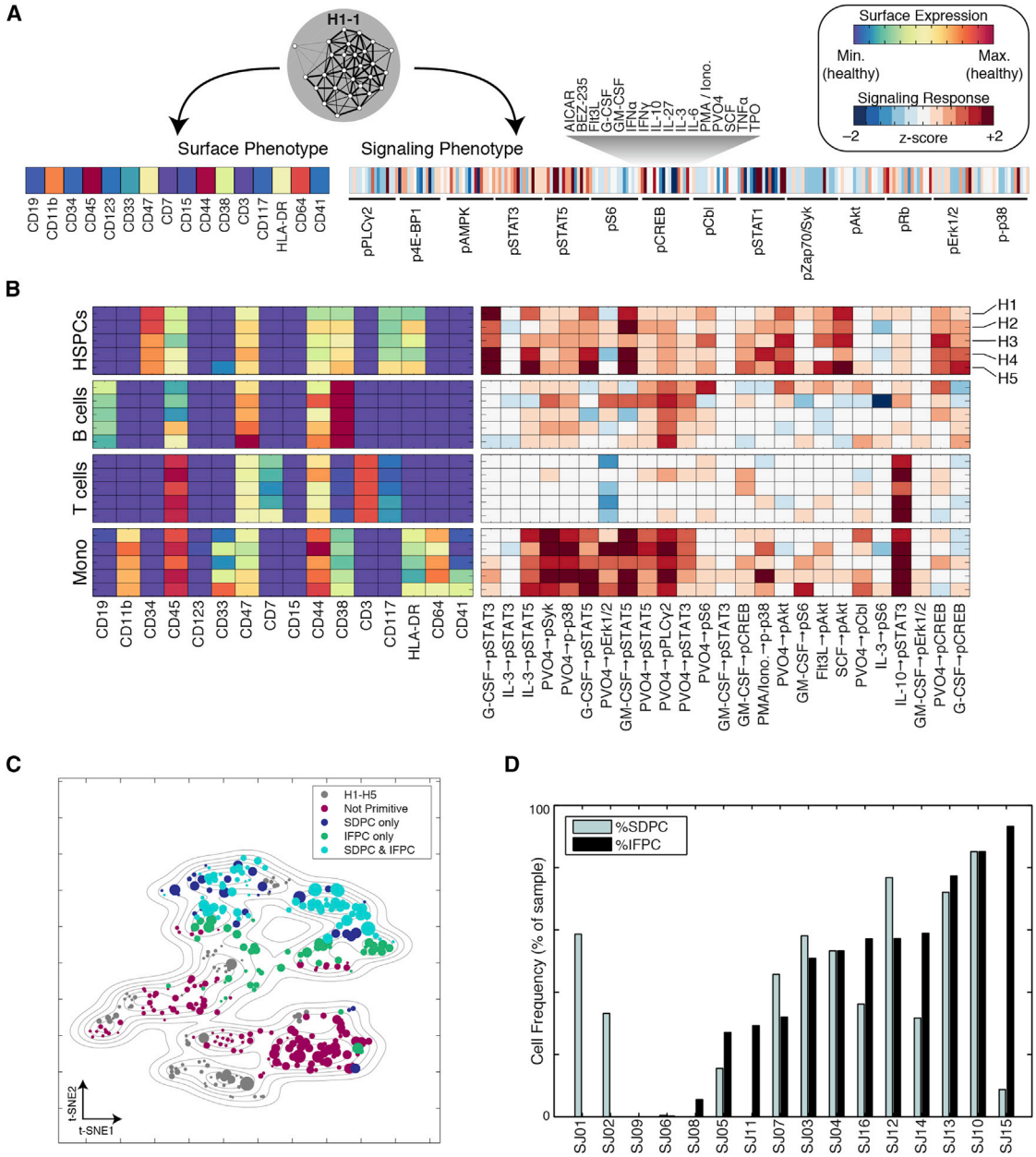


Figure 5. Data-Driven Scoring of Leukemic Maturity by Either Surface or Signaling Phenotype

(A) Each PhenoGraph subpopulation has two alternative phenotypes: surface and signaling.

(B) Normal cell types identified in healthy samples display characteristic surface and signaling phenotypes, represented by heat maps. Each row represents the indicated cell type (Mono = Monocyte). Surface markers (left) and signaling responses (right) are colored as in (A). Signaling responses are ordered from left to right by decreasing significance of association with cell type (Table S2).

(C) The same *t*-SNE map presented in Figure 3A, labeled by results of PhenoGraph classification. Colors depict whether a subpopulation was assigned to either, both, or neither primitive class as determined by signaling (IFPC) or surface (SDPC). (see Figures S4A and S4B).

(D) Frequencies of primitive cells: %IFPC or %SDPC for each patient sample.

their signaling profiles for responses that were consistently associated with particular cell types and found that a large number of signaling responses had significant associations with cell type (Table S2). Many of these were induction responses specific to undifferentiated cells, including G-CSF→pSTAT3 ($Q = 6.4 \times 10^{-42}$) and SCF→pAKT ($Q = 1.0 \times 10^{-9}$), as previously reported (Gibbs et al., 2011).

We then asked whether cell types could be distinguished entirely by their signaling phenotypes, rendering surface phenotypes dispensable for characterizing the subpopulations. To test

this, we developed a framework for classifying subpopulations based on either their surface or signaling phenotypes. We derived an extension of PhenoGraph that uses the same graph-based model but assigns observations to classes according to user-defined training examples ("PhenoGraph classification"; see Experimental Procedures and Figure S4A). First, we verified that PhenoGraph was capable of recovering "held out" healthy cell-type labels using a graph derived from surface phenotypes. Performance was evaluated using the cross-validated correct classification rate (CCR) and indeed, PhenoGraph correctly recovered 99.42% of the cell-type labels in this test. Next, we constructed a graph based only on similarity among signaling phenotypes, withholding all surface phenotype information. Using this graph, PhenoGraph's ability to recover the surface-defined labels was modestly diminished (CCR = 94%) due to errors distinguishing mature cell types for which characteristic signaling phenotypes had not been measured. Focusing on the task of distinguishing the most primitive cells (i.e., HSPCs) from the mature cell types, we found that signaling phenotypes performed equivalently to surface phenotypes (CCR = 99.85%; see Experimental Procedures).

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 examination of their signaling phenotypes. With the healthy subpopulations as training examples, we used PhenoGraph to classify the AML subpopulations, producing an estimate of functional state for each subpopulation (e.g., HSPC-like or monocyte-like). Because there were two alternative phenotypes for each subpopulation—surface and signaling—we performed two separate classifications (Figure S4B). The result was a data-driven assessment of each AML subpopulation, indicating which healthy cell type it resembled in its surface marker expression on one hand and in its high-dimensional signaling phenotype on the other.

Inferred Functional Maturity Diverges from Surface Phenotype 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 ~25% of subpopulations as primitive, but only 16% were identified as primitive by both classifiers simultaneously (Figure 5C). 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 Cells (IFPCs). For each patient, the sample proportion assigned to each of these labels produced two alternative measures of maturity (%SDPC or %IFPC; Figure 5D and Table S2). 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 (Craig and Foon, 2008). Indeed, we found that %SDPC was highly correlated with this standard manual gating procedure (Pearson's $r = 0.96$, $p = 4.4 \times 10^{-9}$; Figure S4C and Data S3B). 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 6, 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 (Figure 6; see Figure S5A for all patients). We used canonical variates analysis (Supplemental Experimental Procedures for details) to quantify this importance, finding that the majority of discriminative power could be attributed to 5 responses: G-CSF→pSTAT3, SCF→pAkt, G-CSF→pSTAT5, Flt3-L→pAkt, and IL-10→pSTAT3 (Figure S5B and Table S2). Primitive subpopulations displayed strong activation in the first four of these responses, which have all been previously implicated in the biology of HSPCs (Gibbs et al., 2011) and in the pathobiology of AML (Irish et al., 2004). Additionally, attenuation of the IL-10→pSTAT3 response—a response exhibited by mature immune cells—was also a distinctive feature of IFPCs. Other signaling responses were strongly associated with primitive subpopulations despite being less powerful for classification (Table S2).

Evaluating the ability of surface markers to identify IFPCs, it was clear that no surface phenotype could be applied universally across patients (Figure 6 and Figure S5A). CD34 was often an important label for IFPCs, but in a subset of cases. For example, CD34 marked both primitive and mature subpopulations in patient SJ03, where HLA-DR was a more specific marker of IFPCs ($p = 0.0007$ versus $p = 0.003$, Student's t test). In SJ05, where CD34 expression was tightly associated with IFPCs ($p = 7.4 \times 10^{-8}$), the multiparameter surface measurements revealed that CD123 was also an important marker ($p = 4.4 \times 10^{-6}$), whereas CD123 did not identify IFPCs in SJ03. Patient SJ11 lacked CD34 expression almost entirely, as expected for this nucleophosmin-mutated case (Taussig et al., 2010). 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-signaling cells ($p = 0.0027$) and IFPCs were marked instead by elevated expression of CD117 ($p = 0.0026$). Complete median surface marker profiles for IFPC and non-IFPC subpopulations are displayed in heat maps for each patient in Figures 6 and S5A.

Primitive Signaling Phenotype Identifies Clinically Prognostic Gene Expression Signature

Ultimately, the importance of intratumor heterogeneity depends on whether functionally distinct subpopulations influence clinical outcomes, especially patient survival (Pearce et al., 2006). While

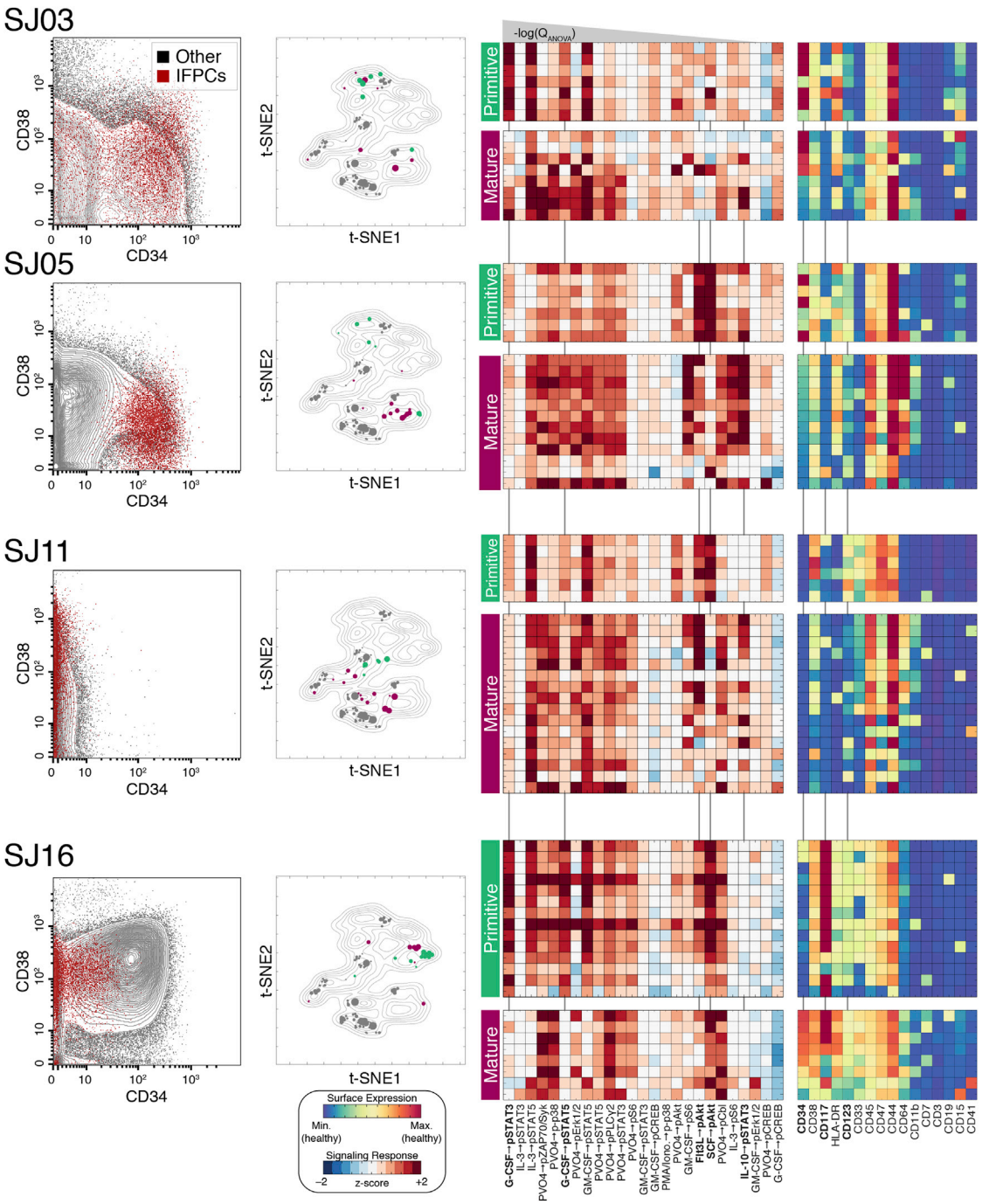


Figure 6. Leukemic Subpopulations with Primitive Signaling Exhibit Diverse Surface Phenotypes

Detailed surface and signaling phenotypes of IFPC subpopulations in four representative samples. Each row represents a particular patient using a number of visuals. Biaxial dot plots (*left*) show the CD34/CD38 phenotype of IFPCs (*red*) in each sample. IFPCs displayed the canonical primitive $CD34^{+}/CD38^{mid}$ phenotype in only a subset of samples. The IFPCs displayed using the *t*-SNE landscape of Figure 3A (center; IFPCs in green, non-IFPCs in maroon, healthy cells in gray). Heat maps (right) 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 5B. Signaling responses marked in bold with vertical lines were especially distinctive of IFPCs (Main Text and Supplemental Experimental Procedures). See Figure S5A for all patients not shown here.

our cohort was too small for survival analysis, genome-wide expression arrays for 15 of our 16 patients were available from a previous study (Radtke et al., 2009), 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 5D), we reasoned that this variance could be

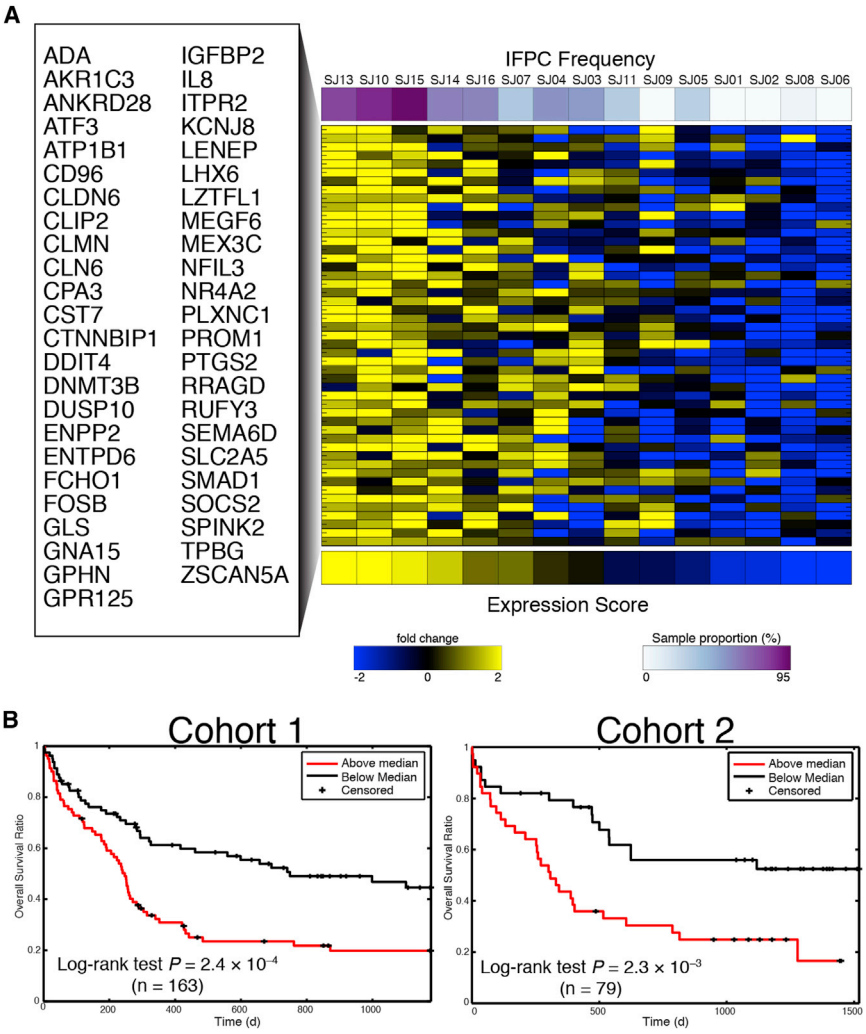


Figure 7. Frequency of IFPCs Identifies a Gene Expression Signature that Predicts Clinical Outcome

(A) IFPC gene signature identified by deconvolution of bulk expression data using IFPC frequency. The heat map displays expression of each gene in the 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 (Metzeler et al., 2008). Patients were assigned to groups for Kaplan-Meier analysis based on whether their IFPC expression score was below or above the cohort median. p values obtained from log-rank test.

exploited to identify genes whose expression covaried with these frequencies by in silico expression deconvolution (Lu et al., 2003). 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.

We developed a deconvolution method based on linear regression and cross-validation and used both %IFPC and % SDPC to produce two gene expression signatures, containing 42 and 49 genes, respectively (see Experimental Procedures, Figure 7A and Table S3). To characterize these signatures, we queried the Molecular Signatures Database (Subramanian et al., 2005) for significant annotations overlapping with each. The SDPC signature—which contained *CD34* among its top-ranked genes—was highly enriched for gene sets associated specifically with $CD34^+$ AML (Table S3). Alternatively, the most significant annotation for the IFPC signature was a set of genes upregulated in $CD133^+$ hematopoietic stem cells (Jaatinen et al., 2006) ($Q = 5.5 \times 10^{-8}$; Table S3). CD133 marks healthy stem cells that are possibly more primitive than $CD34^+$ HSCs (Gallacher et al., 2000) and has been linked to cancer stem cells in multiple cancer types (Collins et al., 2005; O'Brien et al., 2007). The mean expression of each signature was highly correlated with its corresponding subpopulation frequency (Figure S6A), indicating that the signature mean was an appropriate proxy for these frequencies in independent cohorts.

We tested our signatures in two independent cohorts of adult AML for which both gene expression and survival data were available (Metzeler et al., 2008). While the SDPC signature was associated with survival in one cohort, this was not replicated in the other (Figure S6B). Alternatively, the IFPC signature was predictive of poor survival in both cohorts (Figure 7B). Combining the data into a single, large cohort (n = 242), 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 examined the relationship between the IFPC signature and three signatures reported by (Eppert et al., 2011), which were also developed to capture primitive gene expression programs in AML. For each Eppert signature, we were able to reproduce the significant correlation with survival in the data from Metzeler et al. (2008). 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 (Supplemental Experimental Procedures). The IFPC signature proved to be a stronger predictor of survival than any of the Eppert signatures (Table S3). 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; Table S3).

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.

PhenoGraph revealed a phenotypic landscape of a pediatric AML cohort, providing a comprehensive overview of the major phenotypes and an explicit characterization of intra- and intertumor heterogeneity. The landscape resembled normal myeloid development, but with aberrations resulting from malignant accumulation of cells and neoplastic divergence from normal phenotypes. However, this AML landscape was surprisingly restricted to a limited repertoire of 14 MCs, each defined by distinct surface marker patterns. Importantly, these MCs were shared among a wide variety of AML genetic subtypes, yet genetics had a detectable influence on the phenotypic composition of each patient. Together 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.

We used mass cytometry in conjunction with molecular interrogation to construct signaling phenotypes that reflect differences in functional potential between subpopulations. Surface and signaling phenotypes displayed tight coregulation in healthy samples, whereas this coregulation was broken in AML. This 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.

Our demonstration that surface markers are unreliable reporters of signaling state in AML sheds light on the controversies surrounding the LSC model, which rely on manual gating and surface marker expression to define subpopulations. To avoid the assumption that surface markers indicate the functional state of leukemic cells, we used healthy HSPCs to define a primitive signaling phenotype, reflecting the functional state of undifferentiated hematopoietic cells. We found that the primitive signaling phenotype was present in most AML samples and could be used to identify intratumor functional heterogeneity. Leukemic cells displaying primitive signaling (Inferred Functionally Primitive Cells [IFPCs]), were thereby identified using data-driven techniques and without reference to surface phenotypes.

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 (Toren et al., 2005). 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 (Eppert et al., 2011) 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.

Our findings were enabled by computational dissection of intratumor heterogeneity. PhenoGraph creates a graph-based model of cellular phenotypes, similar to that used previously to identify developmental trajectories (Bendall et al., 2014) and in this case defining phenotypes as communities of densely interconnected nodes. PhenoGraph is general and highly 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. These methods are also applicable to healthy tissues, within which a large diversity of cell types remains uncharted.

Our 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. This study provides a framework for interrogating and discovering other features of cell biology that define network response states and their associated mechanistic or clinical outcomes.

EXPERIMENTAL PROCEDURES

Patient Samples

Sixteen (16) cryopreserved diagnostic bone marrow mononuclear cells (BMMC) of pediatric AML patients were supplied by St. Jude Children's

Hospital (Memphis, TN) (Table S1). For healthy adult controls, cryopreserved healthy BMMCs were purchased from AllCells (Emeryville, CA). All human samples were obtained with informed consent in compliance with IRB-approved protocols.

Mass Cytometry Analysis

Mass cytometry measurement and data pre-processing was performed as previously described (Bendall et al., 2011; Finck et al., 2013; Zunder et al., 2015). Surface marker expression was normalized based on the maximum intensity observed in healthy samples, determined as the 99.5th percentile of the ~3M healthy bone marrow cells. Data from all samples were divided by these maximum values, yielding expression values that can be interpreted as x-fold of the maximum observed in healthy. Mass cytometry data are publicly available at http://cytobank.org/nolanlab/reports. See Supplemental Experimental Procedures for full details.

Microarray Data and Normalization

Matched Affymetrix U133A gene expression arrays for 15 pediatric AML patients (Radtke et al., 2009) were downloaded from the Gene Expression Omnibus (GEO: GSE14471). Gene expression and survival data for 242 cytogenetically normal adult AML patients from two independent cohorts (Metzeler et al., 2008) were downloaded from the Gene Expression Omnibus (GEO: GSE12417). All microarray data were processed and normalized as described previously (Akavia et al., 2010).

The PhenoGraph Algorithm

PhenoGraph takes as input a matrix of N single-cell measurements and partitions them into subpopulations by clustering a graph that represents their phenotypic similarity. PhenoGraph builds this graph in two steps. First, it finds the k nearest neighbors for each cell (using Euclidean distance), resulting in N sets of k-neighborhoods. Second, it operates on these sets to build a weighted graph such that the weight between nodes scales with the number of neighbors they share. The Louvain community detection method (Blondel et al., 2008) is then used to find a partition of the graph that maximizes modularity. See Supplemental Experimental Procedures for full details on the method and an assessment of its accuracy, efficiency, and robustness compared to other methods. Source code for PhenoGraph is available online for MATLAB and Python (www.c2b2.columbia.edu/danapeerlab/html/software.html).

PhenoGraph Classification

Given a dataset of N d-dimensional vectors, M distinct classes and a vector providing the class labels for the first L samples, the PhenoGraph classifier assigns labels to the remaining $N - L$ unlabeled vectors. First, a graph is constructed as described above. The classification problem then corresponds to the probability that a random walk originating at unlabeled node x will first reach a labeled node from each of the M classes. This defines an M-dimensional probability distribution for each node x that records its affinity for each class. See Supplemental Experimental Procedures for full details on this method, as well as an evaluation of its performance on benchmark data.

Applying PhenoGraph and SARA to AML Cohort

We ran PhenoGraph on each sample individually, defining subpopulations based on expression of the 16 surface markers. For each sample, all ex vivo conditions were pooled, as we previously demonstrated that surface marker distributions are not altered by these short-term perturbations (Bendall et al., 2011). PhenoGraph was run on the normalized surface phenotype matrices for each sample, with the parameter $k = 50$.

Subpopulation signaling phenotypes were computed for each cluster using SARA, followed by Z score standardization. See Supplemental Experimental Procedures for full details.

Defining AML Metaclusters

Each AML subpopulation was represented by its centroid, resulting in a 425 × 16 matrix. PhenoGraph was run on 425 subpopulation centroids with the parameter $k = 15$, resulting in 14 metaclusters (MCs) delineating the major cohort phenotypes. These MCs are a robust feature of the data and remained consistent when the metaclustering was performed on subsets of patients (Supplemental Experimental Procedures and Figure S3B). To characterize these MCs, we systematically matched cells from healthy bone marrow (H1–H5) with the MC surface marker profiles using linear discriminant analysis. See Supplemental Experimental Procedures for full details.

PhenoGraph Classification of Leukemic Subpopulations

We used the PhenoGraph classifier to classify leukemic subpopulations based on training examples provided by the healthy subpopulations. Using similarities derived either from surface or signaling phenotypes, k-neighbor graphs ($k = 15$) were constructed over 616 subpopulations (healthy and leukemic). Specifically, we used a weighted Euclidean distance in which each phenotypic feature was weighted according to its statistical association with known cell types in the healthy samples. Each AML subpopulation was classified based on its phenotypic proximity to the healthy training examples. Classification was performed using surface and signaling classifiers separately, resulting in two alternative classifications per AML subpopulation (Figures 6 and S4B). See Supplemental Experimental Procedures for full details.

Gene Expression Signatures and Survival Analysis

For each score, %SDPC or %IFPC, a set of associated genes was defined based on correlation with the expression patterns across patients, using linear regression. This in silico gene expression deconvolution assumes that changes in bulk expression of certain genes will track with changes in subpopulation frequency. We used leave-two-out cross-validation across 15 patients to select genes that placed in the top one percentile and had a SD across subsets < 5%.

We used gene expression and survival data for 242 cytogenetically normal adult AML patients from two independent cohorts (Metzeler et al., 2008). For each patient, the frequency of a cell type (%IFPC or %SDPC) was estimated as the mean expression intensity of the associated gene signature. For Kaplan-Meier analysis, patients were stratified into two groups based on the median expression value of the signature of interest. See Supplemental Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, three tables, and three datasets and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.05.047.

AUTHOR CONTRIBUTIONS

J.H.L., E.F.S., S.C.B., D.P., G.P.N. conceived the study. J.H.L., E.F.S., S.C.B., E.D.A., D.P., G.P.N. designed experiments. J.H.L., E.D.A., D.P. designed and developed PhenoGraph. A.L.G., I.R., J.R.D. provided clinical samples. E.F.S., S.C.B., K.L.D., H.G.F., A.J. performed all data acquisition experiments. E.R.Z., R.F. provided barcoding methods. J.H.L., D.P. developed new analysis algorithms. J.H.L., E.D.A., M.D.T., O.L. implemented analysis tools. J.H.L., E.F.S., D.P. analyzed and interpreted the data. J.H.L., E.F.S., S.C.B., D.P., G.P.N. wrote the manuscript.

ACKNOWLEDGMENTS

We thank G. Behbehani, W. Fantl, B.J. Chen, and L. Zelnik for helpful discussion. E.F.S. and S.C.B. are supported by DRCRF Fellowships (DRG 2190-14 & DRG-2017-09) and NIH 1R00-GM104148 to S.C.B. Grants from NIH (DP1-HD084071, DP2-OD002414, R01-CA164729 U54-CA121852), Stand Up To Cancer Phillip A. Sharp Award SU2C-AACR-PS04 and Packard Fellowship for Science and Engineering supported D.P. Grants from NIH (1R01CA130826, 5U54CA143907, HHSN272200700038C, N01-HV-00242, P01 CA034233, U19 AI057229 and U54CA149145), CIRM (DR1-01477 and RB2-01592), EC (HEALTH.2010.1.2-1), US FDA (HHSF223201210194C),US DOD (W81XWH-12-1-0591), the Entertainment Industry Foundation, and the Rachford and Carlota Harris Endowed Professorship supported G.P.N. G.P.N., S.C.B, H.G.F., and E.F.S. have a personal financial interest in the

company Fluidigm, the manufacturer of the mass cytometer used in this manuscript.

Received: January 7, 2015
Revised: March 16, 2015
Accepted: May 4, 2015
Published: June 18, 2015

REFERENCES

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T.R., Brinkman, R., Gottardo, R., and Scheuermann, R.H.; FlowCAP Consortium; DREAM Consortium (2013). Critical assessment of automated flow cytometry data analysis techniques. Nat. Methods *10*, 228–238.

Akavia, U.-D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A., and Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. Cell *143*, 1005–1017.

Amir, E.-A.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol *31*, 545–552.

Bendall, S.C., Simonds, E.F., Qiu, P., Amir, A.D., Krutzik, P.O., Finck, R., Bruggner, R.V., Melamed, R., Trejo, A., Ornatsky, O.I., et al. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science *332*, 687–696.

Bendall, S.C., Davis, K.L., Amir, A.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell *157*, 714–725.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *10*, P10008.

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

Collins, A.T., Berry, P.A., Hyde, C., Stower, M.J., and Maitland, N.J. (2005). Prospective identification of tumorigenic prostate cancer stem cells. Cancer Res. *65*, 10946–10951.

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

Eppert, K., Takenaka, K., Lechman, E.R., Waldron, L., Nilsson, B., van Galen, P., Metzeler, K.H., Poeppl, A., Ling, V., Beyene, J., et al. (2011). Stem cell gene expression programs influence clinical outcome in human leukemia. Nat. Med. *17*, 1086–1093.

Finck, R., Simonds, E.F., Jager, A., Krishnaswamy, S., Sachs, K., Fantl, W., Pe'er, D., Nolan, G.P., and Bendall, S.C. (2013). Normalization of mass cytometry data with bead standards. Cytometry A *83*, 483–494.

Fortunato, S. (2010). Community detection in graphs. Phys. Rep. *486*, 75–174.

Gallacher, L., Murdoch, B., Wu, D.M., Karanu, F.N., Keeney, M., and Bhatia, M. (2000). Isolation and characterization of human CD34(-)Lin(-) and CD34(+)Lin(-) hematopoietic stem cells using cell surface markers AC133 and CD7. Blood *95*, 2813–2820.

Gibbs, K.D., Jr., Gilbert, P.M., Sachs, K., Zhao, F., Blau, H.M., Weissman, I.L., Nolan, G.P., and Majeti, R. (2011). Single-cell phospho-specific flow cytometric analysis demonstrates biochemical and functional heterogeneity in human hematopoietic stem and progenitor compartments. Blood *117*, 4226–4233.

Girvan, M., and Newman, M.E.J. (2002). Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA *99*, 7821–7826.

Irish, J.M., Hovland, R., Krutzik, P.O., Perez, O.D., Bruserud, Ø., Gjertsen, B.T., and Nolan, G.P. (2004). Single cell profiling of potentiated phospho-protein networks in cancer cells. Cell *118*, 217–228.

Jaatinen, T., Hemmoranta, H., Hautaniemi, S., Niemi, J., Nicorici, D., Laine, J., Yli-Harja, O., and Partanen, J. (2006). Global gene expression profile of human cord blood-derived CD133+ cells. Stem Cells *24*, 631–641.

Lu, P., Nakorchevskiy, A., and Marcotte, E.M. (2003). Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. Proc. Natl. Acad. Sci. USA *100*, 10370–10375.

Maaten, L.V.D., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

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

Metzeler, K.H., Hummel, M., Bloomfield, C.D., Spiekermann, K., Braess, J., Sauerland, M.-C., Heinecke, A., Radmacher, M., Marcucci, G., Whitman, S.P., et al.; Cancer and Leukemia Group B; German AML Cooperative Group (2008). An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. Blood *112*, 4193–4201.

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

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Pearce, D.J., Taussig, D., Zibara, K., Smith, L.-L., Ridler, C.M., Preudhomme, C., Young, B.D., Rohatiner, A.Z., Lister, T.A., and Bonnet, D. (2006). 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.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L.M., Baecher-Allan, C., McLachlan, G.J., Tamayo, P., Hafler, D.A., et al. (2009). Automated high-dimensional flow cytometric data analysis. Proc. Natl. Acad. Sci. USA *106*, 8519–8524.

Radtke, I., Mullighan, C.G., Ishii, M., Su, X., Cheng, J., Ma, J., Ganti, R., Cai, Z., Goorha, S., Pounds, S.B., et al. (2009). Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia. Proc. Natl. Acad. Sci. USA *106*, 12944–12949.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Taussig, D.C., Miraki-Moud, F., Anjos-Afonso, F., Pearce, D.J., Allen, K., Ridler, C., Lillington, D., Oakervee, H., Cavenagh, J., Agrawal, S.G., et al. (2008). Anti-CD38 antibody-mediated clearance of human repopulating cells masks the heterogeneity of leukemia-initiating cells. Blood *112*, 568–575.

Taussig, D.C., Vargaftig, J., Miraki-Moud, F., Griessinger, E., Sharrock, K., Luke, T., Lillington, D., Oakervee, H., Cavenagh, J., Agrawal, S.G., et al. (2010). Leukemia-initiating cells from some acute myeloid leukemia patients with mutated nucleophosmin reside in the CD34(-) fraction. Blood *115*, 1976–1984.

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

Toren, A., Bielorai, B., Jacob-Hirsch, J., Fisher, T., Kreiser, D., Moran, O., Zeligson, S., Givol, D., Yitzhaky, A., Itskovitz-Eldor, J., et al. (2005). CD133-positive hematopoietic stem cell "stemness" genes contain many genes mutated or abnormally expressed in leukemia. Stem Cells *23*, 1142–1153.

Zunder, E.R., Finck, R., Behbehani, G.K., Amir, A.D., Krishnaswamy, S., Gonzalez, V.D., Lorang, C.G., Bjornson, Z., Spitzer, M.H., Bodenmiller, B., et al. (2015). Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. Nat. Protoc. *10*, 316–333.

Cell

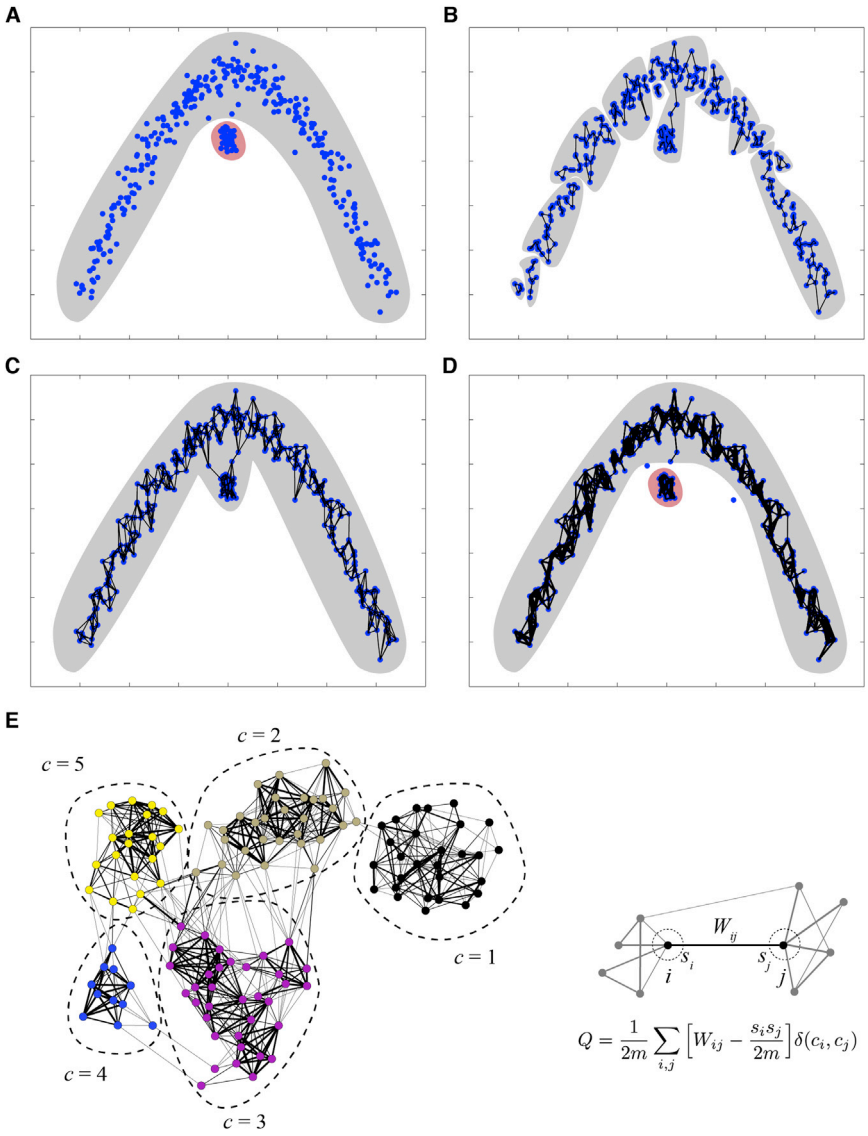


Figure S1. Overview of PhenoGraph Clustering, Related to Figure 1

(A) Simulated data containing two clusters. A small, convex cluster (red) is wrapped by a large concave cluster (gray). (B) k-nearest neighbor graph (k = 5) has disconnected components but fails to represent the separation between the clusters. (C) k-nearest neighbor graph (k = 10) has a single connected component that fails to represent the separation between the clusters. (D) Jaccard graph (k = 10) has two components that represent the two clusters. (E) Example of graph with community assignments that maximize modularity, and a schematic depiction of contribution of two nodes to modularity calculation.

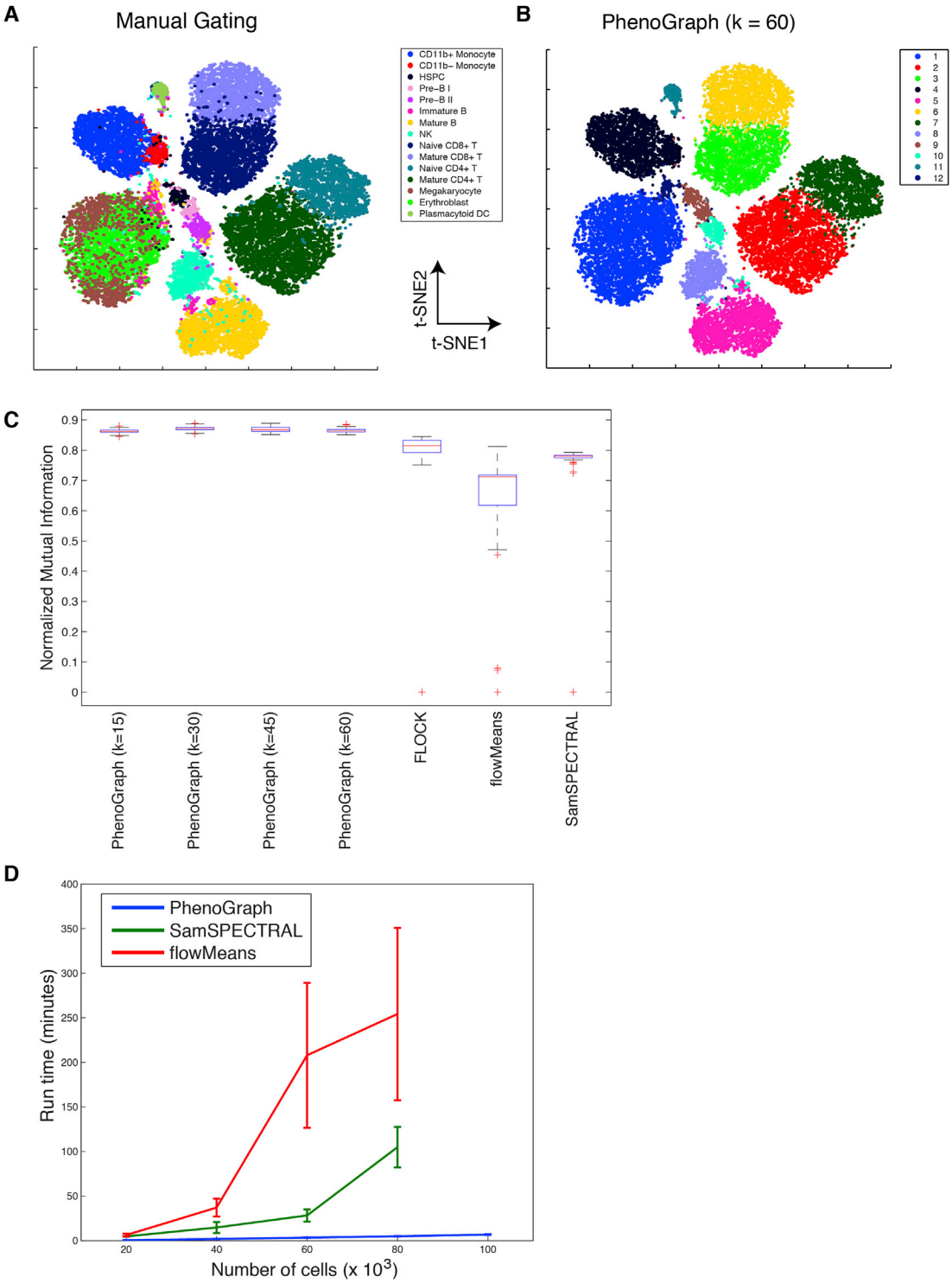


Figure S2. PhenoGraph Is Robust to Parameter Choice and Outperforms Other Methods, Related to Figure 2

(A) Curated benchmark dataset from Bendall et al. (Bendall et al., 2011). Only cells that were assigned to prominent cell types by manual gating were retained. 30,000 cells were sampled randomly from these gates. Colors display manual gating assignment. Data are plotted using t-SNE dimensionality reduction of the 13-dimensional input data. (B) PhenoGraph results for k: 60. For additional values of k, see Data S1B-C and Figure 2A. Colors represent cluster assignment and are ordered by cluster size. (C) Performance of various methods on Benchmark Dataset 1, evaluated by normalized mutual information. (D) 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 SE.

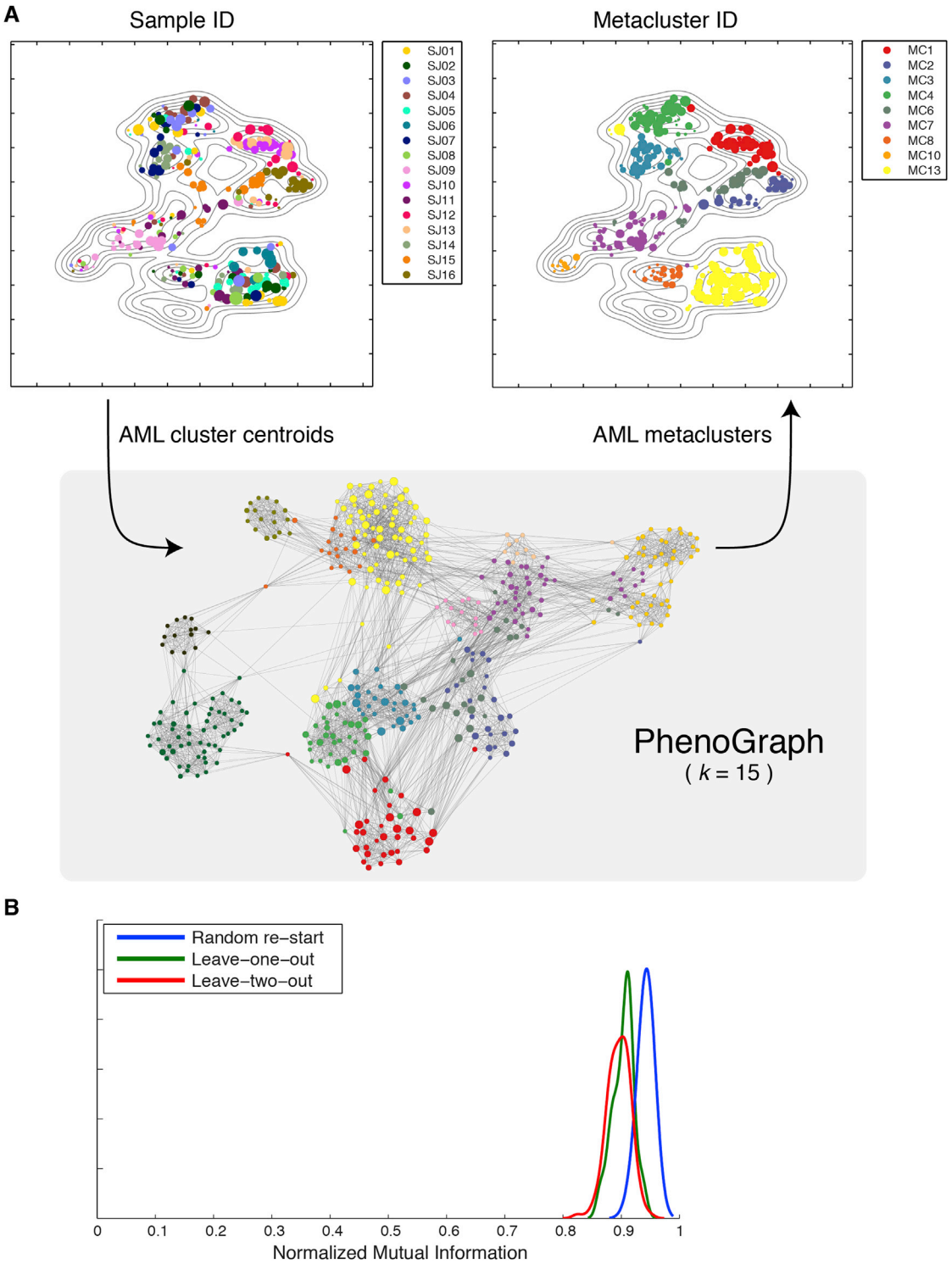
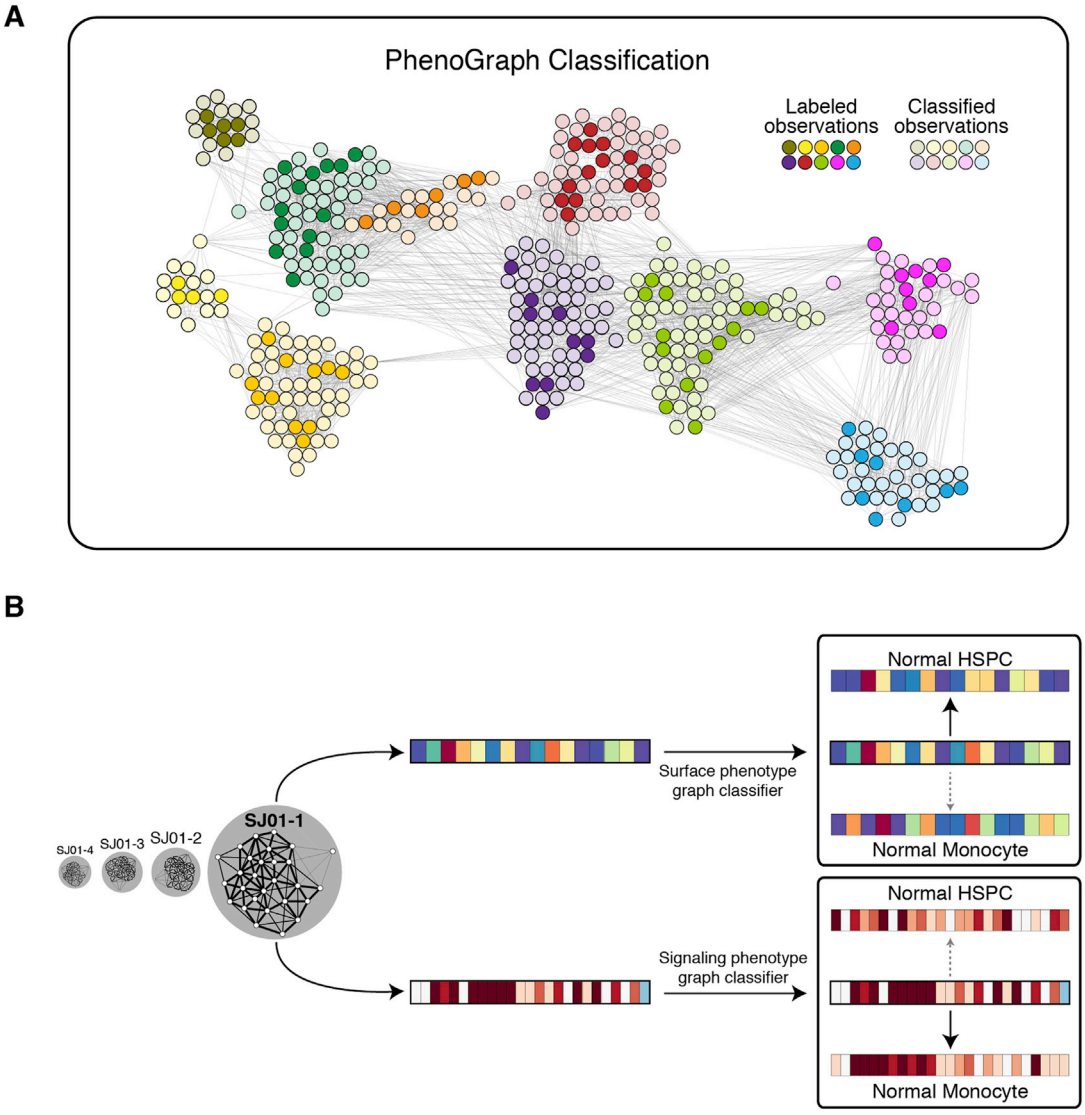


Figure S3. PhenoGraph Metaclustering of Consistent AML Phenotypes, Related to Figure 3

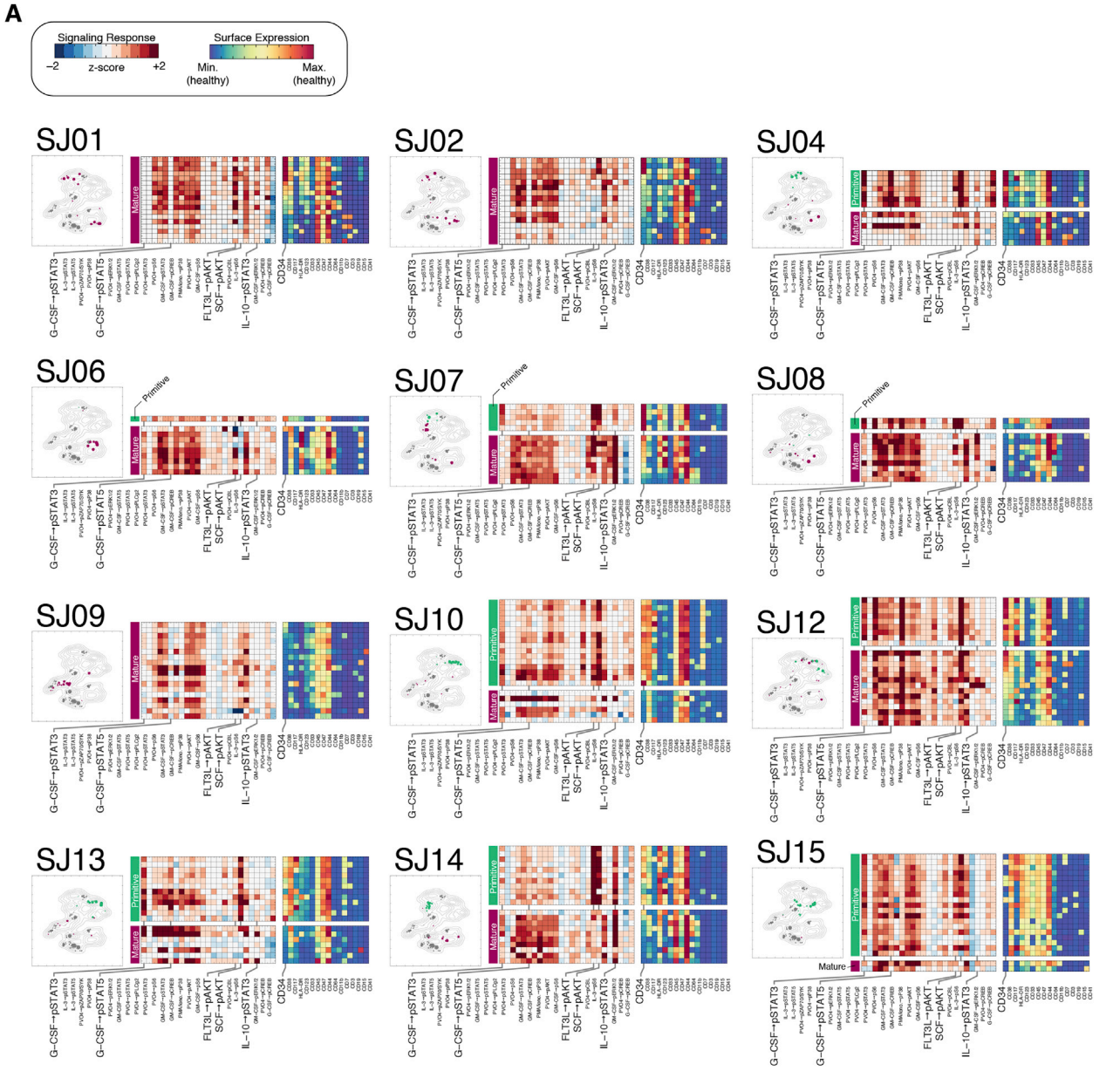
(A) PhenoGraph metaclusters split the AML landscape into major phenotypes, each containing subpopulations from multiple patients. (B) Reproducibility of PhenoGraph metaclusters as assessed by different cross-validation approaches. Metacluster assignments were computed by running PhenoGraph (k = 15) on 16 leave-one-out and 120 leave-two-out datasets. For comparison, 16 random restarts of PhenoGraph (k = 15) were performed on the full dataset. Normalized mutual information was used to quantify the similarity of the partitions, using the assignments from the full dataset as the standard.

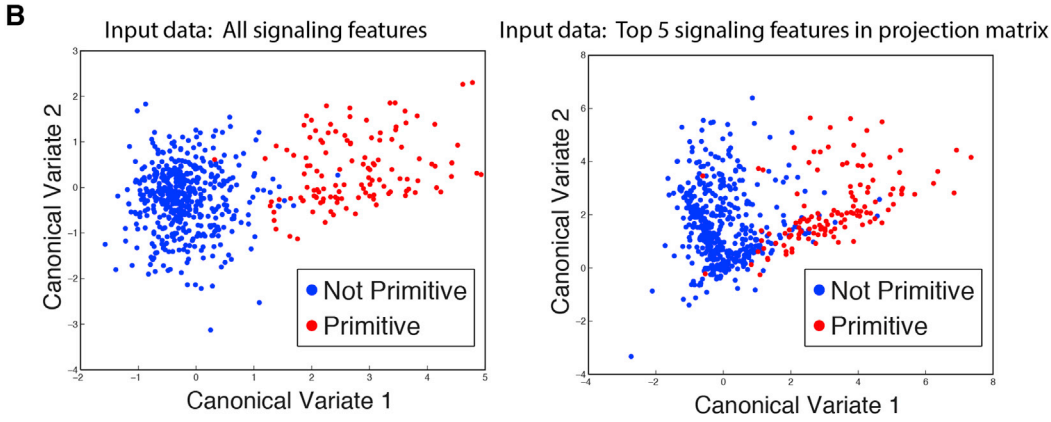


(legend on next page)

Figure S4. PhenoGraph Classification of Inferred Functionally Primitive Cells in AML Samples, Related to Figure 5

(A) Schematic of PhenoGraph classification, which uses proximity of labeled and unlabeled observations on a common graph to classify the unlabeled observations (Supplemental Experimental Procedures for details). (B) Schematic of PhenoGraph classification applied to AML. Each AML subpopulation was classified using normal cell types from the healthy samples as training examples. Of particular interest was whether AML subpopulations were found to be similar to the HSPC (primitive) class. Each AML subpopulation was assigned two classifications, one for its surface phenotype and one for its signaling phenotype. (C) Estimated frequency of cells with primitive surface phenotype as determined by PhenoGraph classification is highly correlated with traditional manual gating (shown in Data S3B).





(legend on next page)

Figure S5. Identifying Primitive AML Subpopulations by Surface and Signaling Phenotypes, Related to Figure 6

(A) Detailed features of subpopulations displaying primitive and mature signaling phenotypes for each sample. (B) Canonical variates analysis demonstrating the linear separability of IFPCs from non-IFPCs across the entire dataset on the basis of 224 signaling features (left) or on the basis of only the 5 most important features (right) as determined by this analysis.

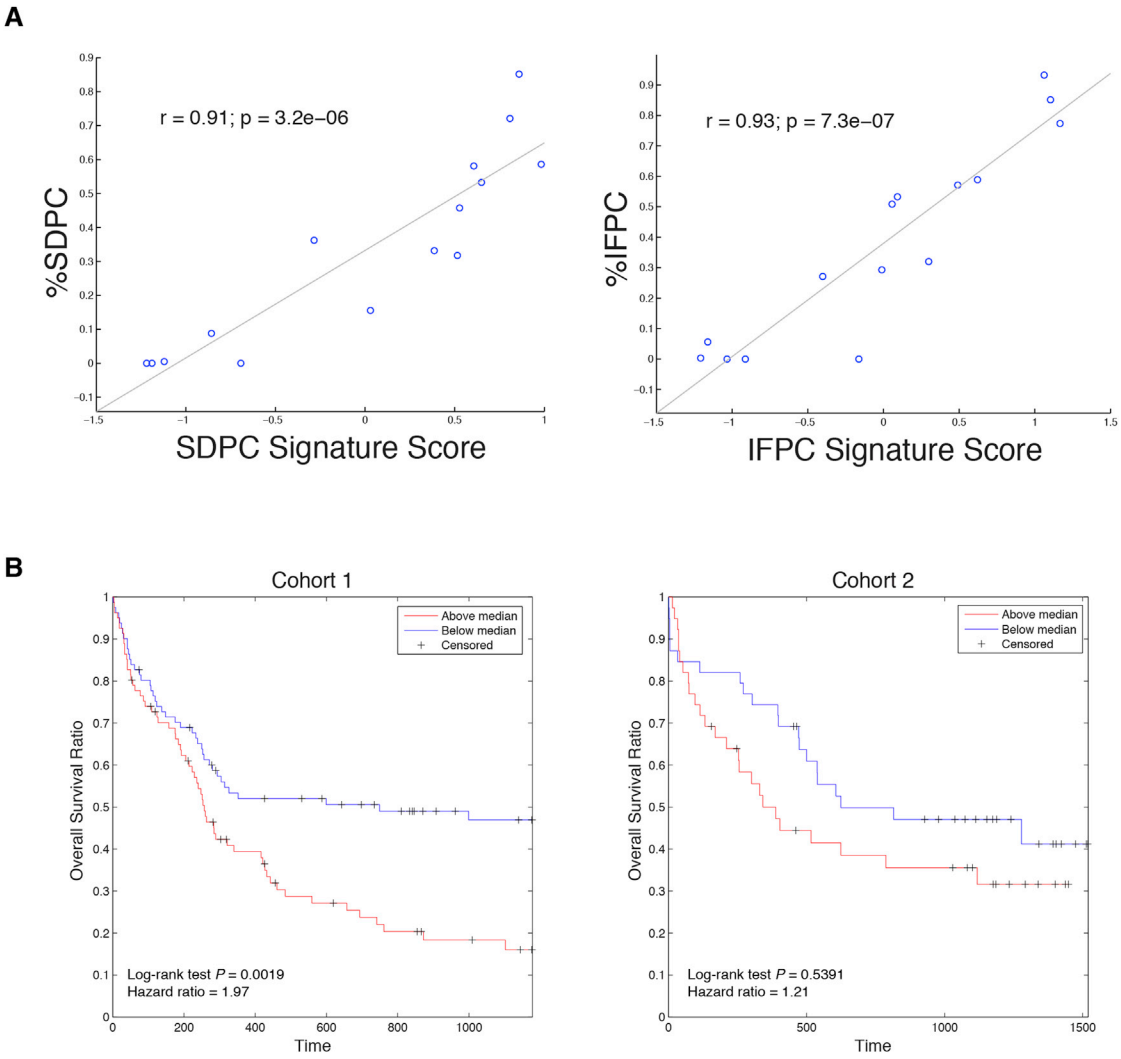


Figure S6. Frequency-Derived Gene Expression Scores, Related to Figure 7

(A) Mean expression of frequency-derived gene signatures are highly correlated with their corresponding frequencies in the 15 patient cohort (SJ12 was excluded from all expression analysis). (B) The SDPC signature was significantly associated with poor survival only in Cohort 1 of Metzeler et al. (Metzeler et al., 2008).