Cell

Supplemental Information

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

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

Supplemental Experimental Procedures

Table of Contents

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 (Radtke et al., 2009). 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 *Mass cytometry measurement*, below) did not pass quality control (QC) thresholds. The final pediatric AML cohort was 50% male with a mean age of 10.4 years (Table S1). 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 differentiation. 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.

Antibodies

Purified monoclonal antibodies were obtained from commercial vendors (Table S1) and labeled with stable metal isotopes using the MAXPAR™ X8 chelating polymer kit (Fluidigm Corporation, South San Francisco, CA) following the manufacturer's instructions. Anti-cleaved caspase-3 was measured but was omitted from downstream analysis because we instead relied on cisplatin viability staining to exclude dead cells (Fienberg et al., 2012). The final data set included a total of 30 antibodies against 14 intracellular targets and 16 cell-surface targets.

Selection of a Minimal Set of Surface Markers to Capture Diversity of AML Blasts

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 S1). 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. Data was gated to remove doublets (DNA^{hi}, $cell\_length^{hi}$) and non-nucleated cells or debris (CD45- DNA-).
2. PCA was performed on gated single-cell data from each patient separately.
3. A "non-redundancy score" (NRS) was calculated according to Equation 1 for each marker A in each patient p using c principal components, where coeff(k) is the coefficient of the marker in component k, and eigenvalue(k) is the eigenvalue of component k.

$$NRS(A_p) = \sum_{k=1:c} |coeff(k)| \times eigenvalue(k)$$

The average NRS was calculated for each of the 42 surface markers in the 36 bone marrow samples using the first 3 principal components (Table S1). 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.

Ex Vivo Stimulation of Bone Marrow Mononuclear Cells

Cryovials of frozen BMMCs were thawed by resuspending cells with 10 mL of 37°C thawing media (RPMI (GIBCO®, Life Technologies, Carlsbad, CA) supplemented with 10% FCS, 1X penicillin/streptomycin (GIBCO®, Life Technologies, Carlsbad, CA), 1X L-glutamine (GIBCO®, Life Technologies, Carlsbad, CA), 25 U/mL benzonase (Sigma-Aldrich, St. Louis, MO) and 20 U/mL sodium heparin (Sigma-Aldrich, St. Louis, MO)), and immediately centrifuging 5 min. at 250 rcf at 20°C. The supernatant was aspirated, and the cell pellet was disrupted by flicking the tube. Cells were resuspended in 2 mL serum-free RPMI containing cisplatin (VWR International, Radnor, PA) at a final concentration of 25 mM for 1 minute at room temperature while swirling vigorously before quenching with 8 mL growth media. Cells were washed immediately with 10 mL 37°C thawing media as described above. Cells were resuspended at 2

x 10^6 cells/mL in 37°C thawing media and distributed to 1 mL aliquots to accommodate different stimulation conditions. Cells were rested for 60 minutes in a 37°C humidified environment with 5% CO_2. Some samples were treated for the final 30 minutes with NVP-BEZ235 (BEZ235, Novartis, Basel, Switzerland) at 1 mM final concentration. Cytokines and chemical perturbations (Table S1) were added and incubated with cells for either the final 5 or 15 minutes, as indicated in the table.

Surface Antibody Staining

Immediately at the end of the incubation period, paraformaldehyde (16% stock, Electron Microscopy Sciences, Hatfield, PA) was added to a final concentration of 1.5%, and cells were fixed for 10 minutes at room temperature. Fixed cells were centrifuged at 600 rcf for 5 minutes at 4°C and aspirated using a 96-well aspirator (V&P Scientific, San Diego, CA). Cells washed twice by adding 1 mL 4°C cell staining media (CSM; PBS with 0.5% bovine serum albumin and 0.02% sodium azide) and centrifuging and aspirating as described above. Cells were resuspended in exactly 50 uL CSM containing metal-conjugated antibodies against the surface markers in the staining panel (Table S1). Cells were stained for 30 min. at room temperature on a rotating shaker at 500 RPM. Stained cells were washed twice with 1 mL CSM. Cells were permeabilized by adding 900 uL of 4°C 100% methanol (Fisher Scientific, Waltham, MA), then immediately transferred to -80°C for overnight storage prior to metal cell barcoding and intracellular antibody staining.

Metal Cell Barcoding

To achieve increased throughput, minimize antibody usage and improve comparability between conditions, metal cell barcoding was used (Bodenmiller et al., 2012; Zunder et al., 2015). All stimulation conditions for a given patient sample were mixed in a single tube prior to staining and acquisition to allow for direct comparison of basal and post-stimulation conditions and then deconvoluted *in silico* based on the barcode of each cell. Each stimulation condition was encoded with a unique error-correcting combination of exactly 3 isotope channels selected from 6 dedicated barcoding channels (Zunder et al., 2015). The metal isotopes used for barcoding were: ^{104}Pd, ^{106}Pd, ^{108}Pd, ^{113}In, ^{115}In and ^{139}La. Lanthanum chloride salt was obtained from Sigma-Aldrich (St. Louis, MO); all other metal chloride salts were enriched stable metal isotopes obtained from Trace Sciences International (Richmond Hill, ON, Canada). Metal barcoding reagents containing enriched palladium isotopes were prepared by combining 2 molar equivalents of isothiocyanobenzyl-EDTA (Dojindo Molecular Technologies, Rockville, MD) with 1 molar equivalent of metal chloride in ammonium acetate buffer (20 mM, pH 6.0). All other barcoding reagents were prepared by combining 2 molar equivalents of maleimido-mono-amide-DOTA (Macrocyclics, Dallas, TX) with 1 molar equivalent of metal chlorides in ammonium acetate buffer. Chelated metal solutions were immediately lyophilized and resulting solids were dissolved in DMSO at 10 mM final concentration for long-term storage at -20°C. Cells in methanol were equilibrated 5 min. at room temperature, centrifuged at 600 rcf for 5 min. at 4°C, carefully aspirated, then vortexed vigorously to disrupt pellets. Cells were washed twice with 1 mL 4°C PBS, then resuspended in 100 uL 4°C PBS. A unique combination of exactly 3 metal cell barcoding reagents, diluted in 900 uL 4°C PBS, was added to each aliquot of cells at final concentrations of 150 nM for palladium isotopes, 100 nM for indium isotopes, and 20 nM

for lanthanum. Barcoding reactions were pipetted to mix, then incubated 1 h at 4°C on a rotating shaker at 60 RPM. Cells were centrifuged and aspirated as above and washed 5 times with 1 mL 4°C CSM. Barcoded cells were then combined in a single tube and washed again with 1 mL 4°C CSM.

Intracellular Antibody Staining

Combined barcoded cells, previously stained with the surface markers, were resuspended in 400 uL CSM containing metal-conjugated antibodies against the intracellular markers in the staining panel (Table S1). Cells were stained for 1 h at room temperature on a rotating shaker at 500 RPM. Stained cells were then washed once with 1 mL 4°C CSM, and stained for 20 minutes in 1 mL of Cell-ID™ Intercalator-Ir solution (Fluidigm Corporation, South San Francisco, CA) at a final concentration of 50 nM. Cells were washed once with 1 mL CSM, washed again with 1 mL ultrapure water, resuspended in ultrapure water at approximately 1 x 10^6 cells/mL, then filtered through a 40 um nylon mesh (BD, Franklin Lakes, NJ) prior to measurement on the CyTOF™ mass cytometer (Fluidigm Corporation, South San Francisco, CA).

Mass Cytometry Measurement and Data Preprocessing

Data was acquired on the CyTOF™ mass cytometer as previously described (Bendall et al., 2011). 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 2 x 10^4 beads/mL. To facilitate quantitative comparisons between data acquired on different days, single-cell data was normalized as previously described (Finck et al., 2013). 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 (Bendall et al., 2011). To remove dead cells and debris, cells were gated based on event_length, DNA content, and cisplatin as described previously (Fienberg et al., 2012).

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 mapping solution. 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 over-expression is often of biological significance, we used the normal 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 x 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 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.

Microarray Data and Normalization

Matched gene expression profiles for our AML patients (Radtke et al., 2009) 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 (Metzeler et al., 2008) were downloaded from GEO (ID # GSE12417). This data set consisted of arrays from two different Affymetrix platforms (U133A and U133 Plus 2.0).

We processed and normalized all microarray data as described previously (Akavia et al., 2010). 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 75th percentile 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.

The PhenoGraph Algorithm for Clustering Single-Cell Data

PhenoGraph aims to partition a dataset of *N* individual cells into subpopulations representing the major phenotypes present, thus enabling an efficient and meaningful profile of the tissue. Single-cell measurements are represented as points in *D*-dimensional space, where each dimension records the expression of a particular molecular species (marker). PhenoGraph receives a *N* x *D* matrix (single-cell measurements) as input and partitions the rows of this matrix (cells) into clusters of cells that are mutually more similar to each other than to other cells. Clustering is often called ‘unsupervised learning’, because it both identifies the classes present (e.g., cell types and phenotypes) and assigns each cell to one of these classes directly from the data, without any *a priori* information. Clustering is a particularly powerful approach to explore less characterized tissues where the phenotypes and subtypes are largely unknown, such as cancer.

Our assumption is that these clusters represent cell populations with biologically meaningful phenotypes. Our premise is that cell populations accumulate in dense regions of *D*-dimensional space, defined by tight marker expression combinations. Therefore, our goal is to

discern these dense regions of cells in the *D*-dimensional space. However, we do not know the number, size, or high-dimensional shape (e.g., ellipsoid, convex) of clusters in the data. The single-cell domain is particularly challenging because cluster sizes can vary by orders of magnitude (e.g., hematopoietic stem cells versus T cells), and we wish to identify rare subsets (clusters) rather than discard them as outliers. Moreover, while most clustering algorithms assume ellipsoid clusters, we have demonstrated that many cell subsets have complex shapes and are not necessarily convex (Amir et al., 2013). Parametric approaches for density detection require strong assumptions about the shape of cellular populations (e.g., ellipsoid, convex), which are frequently violated in single-cell data. Existing non-parametric approaches that directly try to estimate density in high-dimensional space are unstable, sensitive to noise and suffer from computational burden that does not scale well to higher dimensions (see evaluation below, Fig. S2A-C & Data S1).

To overcome these hurdles, we construct a graph structure to represent the high-dimensional geometry of cell states. In this representation, each cell is represented as a node and is connected to its neighbors, the cells most similar to it, via an edge whose weight is set by the similarity between cells. Our approach is inspired by the manner in which cellular populations are generated *in vivo* through an articulated process of division and differentiation. Thus, if appropriate markers are measured, cells spanning multiple developmentally related populations can be represented as a graph structure encoding the paths taken by cells through phenotypic space as they develop. Dense regions of cells (phenotypes) will manifest as highly interconnected modules in this graph, characterized by a high density of edges within this module. Once constructed, this graph can be partitioned into these subsets of densely interconnected modules, called *communities*, which represent distinct phenotypic subpopulations. Community detection in these graphs provides a computationally efficient technique for identifying subpopulations (Girvan and Newman, 2002). Unlike parametric methods such as mixture models, this method makes no assumption about the size, distribution, or number of subpopulations.

Key to the success of our approach is constructing a graph structure that faithfully represents the geometry present in the *D*-dimensional space. However producing such a graph from a *N* x *D* matrix representing multi-dimensional single-cell measurements is a challenging task, especially given measurement noise typical in high-throughput data. What is a good metric of similarity between cells? How many neighbors should each cell have and how should this vary between cells (e.g., cells in dense regions versus cells in sparse regions)? Most importantly, how can we ensure that measurement noise does not obscure the underlying relationships in the data?

PhenoGraph builds a graph structure for single-cell data in two steps. In the first step, the *k* nearest neighbors are identified for each cell using the Euclidean distance, where *k* is the only parameter of the method. As we show in the next section, *k* has little effect on the final result when chosen within a broad range of reasonable values. However, if the final graph would be constructed based on neighbors defined by this process, *k* would have a large impact on the result. On one hand, if *k* is too large, smaller cell populations will be obscured by a large number of edges connecting the cells from true smaller populations to cells outside their population. On the other hand, a *k* that is too small will result in lack of connectivity within the cell populations

we seek to discover (see Fig. S1B). Moreover, as demonstrated in Fig. S1B-C, the problem is further exacerbated by the concave nature of many cell populations. In this example, there is no *k* that both disconnects the smaller (pink) cell population and keeps the larger concave (gray) population in a single connected component, when only a simple *k*-nearest-neighbor approach is used to construct the graph.

Therefore, in the second step, we refine the *k*-neighborhoods defined in the first step. We note that the output of the *k*-neighbor search for all cells defines a set of sets: Specifically, *N* sets of *k*-neighborhoods. We operate on these sets to build a weighted graph. In this graph, the weight between every pair of nodes (cells) is based on the number of neighbors they share. More formally, the weight between nodes i and j, is given by:

$$W_{ij} = \frac{|\nu(i) \cap \nu(j)|}{|\nu(i) \cup \nu(j)|}$$

where $\nu(i)$ is the *k*-neighborhood of point i. This weight—the Jaccard similarity coefficient between *k*-neighborhoods—handles many of the challenges mentioned above. In this metric, the similarity between two cells reaches a maximum when their *k*-neighborhoods are identical and decreases with the number of neighbors they share. Thus, the metric incorporates the structure of the data distribution into the weights, reinforcing edges in dense regions and penalizing edges that span sparse regions. Neighborhood size is automatically tuned such that nodes in dense regions have more edges than nodes in sparse regions. Rare populations are better resolved, as edges between multiple cells within this population are reinforced, whereas outliers tend to have *k*-neighborhoods unlike those of other points, causing them to be excluded from further analysis. For example, this metric naturally resolves a smaller population within the convex hull of a larger population, as demonstrated in Figure S1D.

Graphs constructed from real data in this manner were characterized by an evident modular structure. For example, Figure 1B (*right panel*) depicts such a graph constructed from 500 cells, sampled randomly from healthy donor H1 and displayed by force-directed layout (Kobourov, 2012). *Modularity* refers to the presence of groups of nodes that are more densely connected with each other than with other nodes. This property of graphs has been studied extensively in social network research, where such groups of nodes are called *communities*. Community detection refers to finding a partition of the nodes into distinct communities that captures this modular structure. For a set of community assignments, $C = \{c_1, c_2, \ldots, c_K\}$, the modularity ($Q$) can be defined as (Newman & Girvan, 2004):

$$Q = \frac{1}{2m}\sum_{i,j}\left[W_{ij} - \frac{s_i s_j}{2m}\right]\delta(c_i, c_j)$$

where W_{ij} is the edge weight between nodes i and j, s_i is the sum of all edge weights involving node i, c_i is the community assignment for node i, the Kronecker delta function $\delta(u, v) = 1$ if $u = v$ and 0 otherwise, and $m = \frac{1}{2}\sum W_{ij}$ is a normalization constant. Figure S1E provides a schematic depiction of the assignment of nodes into distinct communities, and the quantities involved in computing the pairwise contribution of two nodes, i and j, to the overall score, Q.

We note that while much of the theory and algorithmic development regarding modularity have been pursued and applied in social network research, it is actually a more general quantity. The modularity (Q) 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 (Reichardt, 2008).

Modularity is bounded by [-1, 1] and can be computed for any graph with a corresponding set of community assignments. For example, the modularity of the community assignments depicted in Fig. 1B (*center panel*) is 0.88. As this score captures the quality of a graph partition, it can be used as an objective function maximized by community detection algorithms. Finding a set of community assignments that maximize Q is a combinatorial optimization problem for which exact solutions are computationally intractable (NP-complete)—yet, good heuristic approximations have been developed. One such approximation, called the Louvain Method (Blondel et al., 2008), has become popular due to its efficiency on large graphs containing hundreds of millions of nodes. This method is hierarchical and agglomerative. At the beginning of the first iteration, every node (cell) is placed into its own cluster. At each iteration, neighboring nodes are merged into clusters for all pairs whose mergers yield the largest increase in overall modularity (Q) of the graph. This process is repeated hierarchically (representing bottom-level clusters as nodes in the next iteration, etc.) until no further increase in Q is obtained.

Since a major advantage of mass cytometry is the ability to measure millions of cells—and detecting rare subpopulations will require large sample sizes—a design objective for PhenoGraph was the ability to model population sizes into the millions. Therefore, PhenoGraph uses the Louvain Method to maximize the modularity of its partitions. Specifically, PhenoGraph runs multiple random restarts of the Louvain Method, choosing a final partition in which the modularity reaches a maximum among all solutions. Due to the inherent modularity of the graph, 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 and reported for Fig. 1B above is extremely small (8.75×10^{-4}). For all results presented in this publication, PhenoGraph selected the best solution from 100 random restarts of the Louvain Method.

PhenoGraph Pseudocode

Input: data set of single-cell measurements $X = \{x_1, \ldots, x_N\}$, neighborhood size *k*
Output: subpopulation index $C = \{c_1, \ldots, c_M\}$ assigning each cell in X to one of *M* groups
Initialization:

for each cell x_i

find the indices $\nu(i)$ of the *k* nearest cells

Graph construction:

create a set of vertices $V = \{v_1, \ldots, v_N\}$ corresponding to each cell in X, and an empty set of edges E = { }

for each pair of cells x_i and x_j

compute $W_{ij} = \frac{|\nu(i) \cap \nu(j)|}{|\nu(i) \cup \nu(j)|}$

if $W_{ij} > 0$, add $e_{ij} = W_{ij}$ to E

return G = (V, E)

Community detection:

for t in {1, …, 100}

Decompose G into community assignments C_t that give a local modularity maximum Q_t using the Louvain Method

determine t such that $Q_t > Q_p \ \forall p \neq t$

return $C = C_t$

Two implementations of PhenoGraph are available: in MATLAB and in Python. Both use the Louvain Method for modularity optimization, implemented in C++ (sites.google.com/site/findcommunities/). The original C++ code was modified to increase file read/write speed.

Evaluation of PhenoGraph Performance on Healthy Data

Three healthy marrows taken from two independent data sets were used to evaluate PhenoGraph's performance and to test it against other available methods for clustering single-cell data.

Benchmark Data Set 1 was a publicly available mass cytometry data set (Bendall et al., 2011) (http://reports.cytobank.org/1/v1) of healthy adult BMMCs. It consisted of 167,044 cells collected from a healthy human donor ("Marrow 1"), 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 clustering in our testing. Manual gating assigned 49% of the cells to a known cell type while the remaining cells remained unlabeled by the gating strategy.

Benchmark Data Set 2 was a newly acquired mass cytometry data set measuring 32 surface marker antibodies (Table S1) on BMMCs from two healthy adult donors. These samples were taken from two of the same donors used in the signaling study (H1 and H2). 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. All 32 markers were used for clustering in our testing. (*Data are available for download at http://cytobank.org/nolanlab/reports*).

For both benchmark data sets, clustering was applied to all cells (including the unassigned cells) and evaluated based on the manually assigned cells. To quantify clustering quality, we used the mean *F*-measure, as was used in the recent FlowCAP competition and described in their recent publication (Aghaeepour et al., 2013). Briefly, for each subpopulation c_i in the benchmark data and a corresponding cluster k_j returned by the algorithm, Precision (Pr_{ij}) quantifies the proportion of k_j that correctly identifies c_i, Recall (Re_{ij}) quantifies the proportion of c_i that is correctly identified by k_j, and the *F*-measure is defined as the harmonic mean of these quantities:

$$F(c_i, k_j) = \frac{2 \times \mathrm{Pr}_{ij} \times \mathrm{Re}_{ij}}{\mathrm{Pr}_{ij} + \mathrm{Re}_{ij}}$$

The *F*-measure quantifies the accuracy of binary classifications, but can be extended to multiple classes $C = \{c_1, c_2, \ldots, c_M\}$ by taking the weighted average:

$$F_{\mathrm{mean}}(C, K) = \sum_{c_i \in C} \frac{|c_i|}{N} \max_{k_j \in K} F(c_i, k_j)$$

where $\frac{|c_i|}{N}$ denotes the proportion of the benchmark data in class (subpopulation) c_i. The mean *F*-measure is bounded by the interval [0, 1] with 1 representing a clustering result that perfectly matches the manually-defined subpopulations.

While the mean *F*-measure was used by the FlowCAP consortium (Aghaeepour et al., 2013), we added a second metric to avoid any bias that may be inherent in the *F*-measure statistic. For the second metric, we used the normalized mutual information (NMI) (Strehl and Ghosh, 2003), another popular score for cluster quality. 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:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y)\, log(\frac{p(x,y)}{p(x)p(y)})$$

which is non-negative but otherwise unbounded. The normalized variant is bounded such that perfect accuracy maximizes the score at 1.0,

$$NMI = \frac{I(X;Y)}{\sqrt{H(X)\, H(Y)}}$$

where $H(X)$ is the Shannon entropy of X.

PhenoGraph Correctly Identifies Most Healthy Subpopulations

First we used Benchmark Data Set 1 to assemble a curated dataset including only labeled cells (i.e., from manual gating) and PhenoGraph was run on the normalized surface phenotype matrices for this data, with the parameter *k* = 15, resulting in 15 clusters. We note that it is only by chance that number of nearest neighbors (*k* = 15), results in 15 clusters and the neighborhood size does not typically match the number of clusters. Figure 2A demonstrates 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.

PhenoGraph Is Robust to Parameter Choice

To evaluate the robustness of output, we produced 50 random subsamples of 20,000 cells from Benchmark Data Set 1 (which included both labeled and unlabeled cells) and ran PhenoGraph on each sample. The resulting mean *F*-measure was high (0.87) with a tight distribution (Fig. 2B). This demonstrates both the concordance of PhenoGraph with the manual labels, as well as the robustness of this solution.

PhenoGraph has only a single parameter, *k*, denoting the number of neighbors identified in the first iteration of the algorithm (see previous section for details on the algorithm). A key concern with such a user-defined parameter is these might have a large impact on the resulting output. To test PhenoGraph's sensitivity to the choice of *k*, PhenoGraph's performance was evaluated using 4 different choices spanning a four-fold range of reasonable values, *k*=(15, 30, 45, 60). The *F*-measure was similarly high (0.88 for *k*=30, 45, 60) and tightly distributed over this 4-fold range (Fig. 2B). Moreover, comparing PhenoGraph outputs on the curated data set, we observe nearly identical labeling of the cells across all values of *k* (Figs. 2A, S2B, and Data S1B-C).

Thus, we have demonstrated that PhenoGraph is robust to random resampling of the data and a wide range of values for the single parameter *k*.

PhenoGraph Outcompetes Leading Clustering Methods

Finally, we compared PhenoGraph to additional state-of-the-art methods for clustering single-cell data. Recently, the FlowCAP consortium (http://flowcap.flowsite.org/) published a critical assessment of available computational methods for clustering single-cell data (Aghaeepour et al., 2013). We chose the top 3 open-source methods from this assessment for testing on mass cytometry data: FLOCK (Qian et al., 2010), flowMeans (Aghaeepour, 2011), and SamSPECTRAL (Zare et al., 2010). Additionally, we chose 2 standard clustering methods: Gaussian mixture modeling (GMM) and hierarchical linkage clustering.

To evaluate the robustness of each method's output, we produced 50 random subsamples of 20,000 cells from Benchmark Data Set 1 and ran each method on each subsample, resulting in a distribution of mean *F*-measures for each method. Ideal performance is reflected in this distribution by a high center (solution quality) and small variance (solution robustness). Figure 2B shows the mean *F*-measure distributions for all tested methods. PhenoGraph (for each tested value of *k*) clearly outperformed all methods both in terms of quality and robustness. The same results were recapitulated with the normalized mutual information, demonstrating superior performance of PhenoGraph independent of quality metric (Fig. S2C). Figure S2A-B and Data S1A-F provide detailed visualization of the outputs of the various methods for the curated data presented in Figure 2A and discussed above. Figures 2B and S2B–D show PhenoGraph results for *k*=(15, 30, 45, 60), respectively. In all cases, PhenoGraph correctly distinguishes naive and mature CD4+ and CD8+ T cells; NK cells; Pre-B II, immature and mature B cells; monocytes; pDCs. Larger values of *k* tend to encourage coarse-graining of the results, for example, merging CD11b+ and CD11b- monocytes. All other methods failed to correctly distinguish naive and mature CD4+ or CD8+ T cells (Data S1D-F), a major distinctive feature of this data. These methods either over-simplified the structure of the

population (SamSPECTRAL, Data S1D; flowMeans, Fig. S2C), or identified spurious structure in homogeneous subpopulations (FLOCK, Data S1E).

We repeated our tests on Benchmark Data Set 2. This data set consists of 32-parameter single-cell measurements, and this higher dimensionality presents a greater challenge to the methods developed for lower-dimensional fluorescence cytometry. For example, FLOCK was unable to run at all because it assumes that the number of observations is vastly greater than the number of dimensions, such that a 40,000-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 computation time (Data S1G-I). 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.

Finally, we observed that competing methods don’t scale well with number of cells and are therefore unsuitable for large data such as the AML data we present here. Figure S2D shows run times for flowMeans, SamSPECTRAL, and PhenoGraph on random subsamples of increasing size from the 32-parameter Benchmark Data Set 2. All runs were performed on a 2.6 GHz Intel Core i7 with 16GB of RAM. Five random data sets were generated at each sample size and the figure represents the mean runtime and standard error over these five samples at each size increment. At 80,000 cells, average run time was 105 minutes for SamSPECTRAL, 254 minutes for flowMeans, and 5 minutes for PhenoGraph. On a laptop computer, PhenoGraph can process 1,000,000 cells in 229 minutes: less time than it takes flowMeans to process 80,000 cells.

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 (σ = 100, sf = 1). For hierarchical linkage clustering and Gaussian mixture models, MATLAB R2013b implementations were used.

SARA: Statistical Analysis of Perturbation Response

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 (Becskei et al., 2001; Ferrell and Machleder, 1998; Novick and Weiner, 1957). 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 Perturbation Response). SARA examines the entire single-cell distribution of phosphoprotein intensities to detect meaningful changes between two conditions.

SARA proceeds as follows. We treat each phospho-marker ϕ as a random variable whose distribution depends on two observed variables: cluster membership, C, and condition, Z. We are interested in comparing two distributions:

$$P_b(\,\phi \mid Z = b, C = c\,)$$
$$P_s(\,\phi \mid Z = s, C = c\,)$$

where b denotes measurement under the unstimulated ('basal') condition and s denotes measurement under the stimulated condition. In all cases, we compare basal and stimulated distributions within the same cluster. The quantity of interest for comparing these distributions is the cost of converting one distribution into the other. This quantity is known equivalently as the Earth Mover's or Mallows distance (Levina & Bickel, 2001) and, for the one-dimensional case is the L_1 norm between the empirical cumulative distribution functions, F_b and F_s:

$$\mathrm{EMD} \;=\; \sum_{\Phi} |F_b(\phi) \;-\; F_s(\phi)|$$

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. We therefore introduce a measure of statistical significance for the EMD. We build a null distribution by computing EMD over 5000 random permutations of Z, and calculate a corresponding p-value. We integrate the p-value into the final score as an inverse weight that dampens the magnitude of less significant responses.

Finally, while EMD is nonnegative, stimulation response typically has a sign—positive for induction or negative for inhibition—with respect to the basal distribution. The direction is incorporated by comparing the centers of P_b and P_s. Random noise may cause spurious changes in direction when the stimulation has an insignificant effect, in which case the penalized magnitude will cause these values to be distributed near 0. We note there are cases for which incorporation of direction is not appropriate (for example, a significant, evenly diverging response), but we did not observe any such case in our data. The final score given by SARA is (see Figure 4A):

$$\mathrm{score} \;=\; \mathrm{EMD} \cdot \mathrm{sgn}(\mathbb{E}[\phi_s] - \mathbb{E}[\phi_b]) \cdot (1 - p)$$

where $\mathbb{E}[\cdot]$ represents the expected value of a random variable and p denotes the permutation p-value.

PhenoGraph Classification

Central to PhenoGraph is its underlying graph structure, constructed by representing data points as nodes in a graph, which are connected via edges that capture their proximities. As we demonstrated, the resulting graph structure can be partitioned into communities of

interconnected nodes, effectively clustering the data into subsets in an unsupervised manner. Here, we use this same graph structure for a *classification* task, in which unlabeled samples ("test data") are assigned to classes, based on learning from a set of labeled samples ("training data"). In other words, given a partially labeled data set, classification uses the known examples to extend labels to the entire dataset. Our approach uses PhenoGraph's graph structure to infer "cell types" for uncharacterized cells or subpopulations, using a reference data set as a guide. In particular, we wanted to use this approach to generate semi-supervised characterizations of the leukemic subpopulations, using the healthy subpopulations as reference examples.

Given a dataset of N *D*-dimensional vectors $X = \{x_1, \ldots, x_N\}$, *M* distinct classes and a vector $Y = \{y_1,\ldots, y_L\}$ providing the class labels for the first L samples, the PhenoGraph classifier assigns labels $Y = \{y_{L+1},\ldots, y_N\}$ to the remaining unlabeled samples in the data. First, a graph is constructed as before, for all N samples, using the Jaccard similarity coefficient between *k*-neighborhoods, without reference to the labels. Classification methods are often based on "guilt by association" and the graph's neighborhood structure effectively encodes for such associations (see Figure S4A). Consider an unlabeled sample x: If a majority of x's labeled neighbors share a distinct class m, based on a "neighborhood vote" it would be natural to label x with m. However, the basic "guilt by association" approach is easily confounded when x's immediate neighborhood is only a small part of the available training data and might not even have any labeled samples. Therefore, we generalize this idea in a way that takes the entire graph structure into account. Intuitively, we allow x to be influenced by more distant labeled nodes beyond its immediate neighborhood, connected to x through multiple edges in the graph structure. The "vote" of each labeled node is then weighted by the number of edges between x and the labeled node and their weights. However, there are often multiple routes between two nodes and these should all be taken into account. An effective way to capture all possible paths is through the process of a random walk: At each step the walk continues to a new neighbor, selected based on the relative edge weights. Distance between two nodes is thus reformulated as the probability of the random walk first reaching one node, starting from the other.

The classification problem then becomes 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 a *M*-dimensional probability distribution for each node x that records its affinity for each class m. If a number of nearby nodes are all labeled with the class m, the probability that x would reach m first is high. Thus, the graph is used in a series of random walk simulations that propagate the class information from the labeled nodes onto the unlabeled nodes. A key advantage to the PhenoGraph structure is that the Jaccard metric enriches for edges that likely share a class label. Additionally, the modular nature of the graph structure further concentrates the probability distribution to labeled nodes within the same module.

Exploiting the connection between random walks on graphs and discrete potential theory (Hersh & Griego, 1969), the probability of cell x 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 nodes of each labeled class (Grady, 2006). Each unlabeled node is then assigned to the class that it reaches first with highest probability.

Given a partially labeled data set, X, we compute the Jaccard graph G = (V, E) as described above. From G, we define two alternative $\mathrm{N} \times \mathrm{N}$ sparse matrices:

$$A_{ij} = \begin{cases} W_{ij} & \text{if } e_{ij} \in \mathrm{E} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{ii} = \begin{cases} \sum_j A_{ij} & \\ 0 & \text{if } i \neq j \end{cases}$$

These matrices define the graph Laplacian:

$$\mathcal{L} = D - A$$

Note that each vertex of G is either labeled or unlabeled and therefore must be in one of two sets, V_L (labeled vertices) and V_U (unlabeled vertices), such that $V_L \cup V_U = \mathrm{V}$ and $V_L \cap V_U = \emptyset$. Thus, $\mathcal{L}$ can be decomposed into submatrices corresponding to V_L and V_U:

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

The graph Laplacian is used to compute the probability that random walkers originating at vertices in V_U first arrive at particular vertices in V_L. These probabilities can be calculated through the solution of a system of sparse linear equations (see Grady [2006] for mathematical derivation). Specifically,

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

where Q is a $L \times M$ binary matrix representing the label 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.

PhenoGraph Classifier Pseudocode

Input: data set of measurements X = {x_1, …, x_N}, M distinct classes, a vector of labels Y= {y_1,…, y_L} so that y_i denotes the class label for x_i , neighborhood size *k*

Output: vector of labels Y' = {y_{L+1}, …, y_N} so that y_i denotes the inferred class label for x_i

Graph construction:

initialize and construct graph G as specified in the PhenoGraph Pseudocode above

Classification:

simulate random walks as described in Grady (2006).

return P = {p_{L+1}, …, p_N}, a set of M-dimensional probability vectors defining the random walk proximity of each unlabeled node to each class.

define $y_i = j$ s.t. $p_{ij} = \max(p_i)$

return Y' = {y_{L+1}, …, y_N}

Applying PhenoGraph to Pediatric Cohort and Healthy Donors

We ran PhenoGraph on each sample (AML patient or healthy donor) individually, defining subpopulations based on expression of the 16 measured surface markers. This approach differs from many published methods for analyzing cytometry data, including SPADE (Qiu et al., 2011), which generally require pooling of the data from multiple samples into a single matrix prior to any clustering or dimensionality reduction steps. While pooling multiple samples may facilitate immediate comparison across samples by placing cells from different samples into the same cluster, this approach is vulnerable to bias caused by sample size and may produce clusters that are difficult to interpret. The decision to analyze each sample individually was a

conscious design choice, driven by both biological and technical considerations. We wanted to create a profile of each sample and ensure that each reported subpopulation was indeed present, rather than an outlier spuriously associated with a subpopulation present in a different sample. The 1.3–16 x 10^5 cells collected per sample were sufficient to detect rare subpopulations (as low as 0.06% of the sample). Additionally, since there is tube-to-tube variation in antibody staining, PhenoGraph was run on directly comparable cells stained within the same barcoded tube.

For each healthy volunteer or patient sample, single-cell measurements from all *ex vivo* conditions (including the two basal control conditions), which were all measured together in a single barcoded tube, were pooled and compiled into a single data matrix. This pooling both increased the total number of cells, facilitating the identification of rare populations and later enabled characterization of subpopulation-specific signaling dynamics. PhenoGraph was run using only the 16 surface markers to define phenotypic similarity. We assumed that the surface markers do not change within 15 minutes following stimulation (as previously demonstrated in (Bendall et al., 2011)), thus making these directly comparable across conditions. PhenoGraph was run on the normalized surface marker matrices for each sample, with the parameter k=50. We reasoned that 50 neighbors was a sufficiently large number to estimate the geometry of the local neighborhood while being small enough to avoid inappropriately large neighborhoods. Because the effective neighborhood size is tuned to the local density during the second step in the graph construction, results are robust to variations in reasonable choices for this parameter (see previous section).

This yielded an average of 28 subpopulations per sample (ranging between 17 and 48), totaling 641 subpopulations across the entire cohort of healthy and leukemic samples. Subpopulation size varied by orders of magnitude, from 7x10^2 to 2x10^5 cells (or .06% to 20% of a sample).

Defining Subpopulation Phenotypes

Subpopulation surface and signaling phenotypes were computed for each cluster returned by PhenoGraph (Figure 5A). Subpopulation surface phenotypes were defined simply as the median value of each surface marker among all cells in that cluster. Subpopulation signaling phenotypes were computed by SARA, followed by z-score standardization. SARA was computed for each phospho-marker X stimulation pair by comparing cells collected in the basal and post-stimulation condition, as described above. A vast majority of subpopulations contained sufficient cells from all conditions to compute signaling phenotypes. Subpopulations that contained less than 20 cells from all conditions were excluded from further analysis (25/641 original clusters), resulting in 616 (191 healthy + 425 AML) cohort subpopulations.

To facilitate comparability across samples and signaling phenotypes, SARA scores were converted into 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, we noted subtle sample-specific biases in these dynamic ranges, likely due to inevitable differences in handling of primary human samples from day to day. Therefore, within each condition and sample, we pooled SARA values from all subpopulations and all phospho-markers and standardized the

SARA scores by re-expressing them as z-scores. Thus each value in the signaling phenotype represents the relative magnitude of the response, within the contexts of condition and sample. Moreover, 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 also enhance interpretability by bringing insignificant fluctuations down to zero, highlighting the most significant responses (Figure 4A).

Defining AML Metaclusters

We performed metaclustering to define cohort-level phenotypes and enable a more rigorous matching of subpopulations across patients. Effectively, this entailed a second application of PhenoGraph to the previously defined PhenoGraph clusters from all patients in the cohort. To identify AML-centric phenotypes and avoid bias toward healthy phenotypes, we chose exclude the healthy samples from this analysis and thus perform metaclustering only on the 425 subpopulations derived from the 16 AML patients. Each subpopulation was represented by its centroid, a 16-dimensional vector computed by taking the median of each surface marker across all cells in that subpopulation, resulting in a 425 x 16 matrix (subpopulations x surface markers). PhenoGraph was run on this matrix with the parameter k=15, taking into account the smaller number of data points (only 425 clusters, relative to ~1 million cells in the previous application of PhenoGraph). We note that the resulting graph had a clear modular structure (see Figure S3A, which visualizes the neighbor graph for all subpopulations). PhenoGraph partitioned the 425 subpopulations into 14 metaclusters (MCs) delineating the major cohort phenotypes. Each MC had a mixed patient composition, containing subpopulations from at least 2 patients and a median of 11 patients (Fig. 3B).

AML Metaclusters Are Robust to Subsampling

With 16 patients the cohort is relatively small, therefore we conducted robustness analysis to determine if the MCs were biased by the inclusion of any particular patient(s). By subsampling the patients and recomputing the metacluster assignments, we assessed the degree to which we obtain similar MCs in the absence of certain patients. We produced 16 leave-one-out and 120 leave-two-out data sets in which all subpopulations from patient i (and j in the case of leave-two-out) were removed. For each of these subsamples, we computed metacluster assignments using PhenoGraph as described in the previous paragraph (k = 15). We then compared these assignments to the full-data MC assignments of the same populations, using normalized mutual information to quantify the similarity of the partitions. Using this cross-validation scheme, we found that the MCs were a robust description of the AML phenotypes across the entire cohort, changing very little as different patients were added or removed. The NMI scores are displayed in Fig. S3B. For comparison, we calculated NMI between our MC assignments and 16 random restarts of PhenoGraph on the full 425-subpopulation data, yielding an average score of NMI = 0.94. NMI scores for the leave-one-out and leave-two-out subsamples were only modestly diminished, both averaging at NMI = 0.9, and tightly distributed (standard deviation = 0.02).

Identification of Healthy Cell Types

Healthy cell types were defined in a manner analogous to AML metaclusters, described above. For the 191 healthy subpopulations, the surface phenotypes were used to build a healthy PhenoGraph model (*k*=15), producing 20 metaclusters (Fig. S4 and Data S3). The metaclusters generally displayed recognizable surface phenotypes corresponding to known cell types, such as monocytes (healthy metacluster 1) and HSPCs (healthy metacluster 9).

Healthy Cognates of AML Metaclusters

Some of the surface phenotypes of the AML metaclusters (MCs) resembled healthy cell types, for example the CD19+/HLA-DR+ phenotype of MC12 matched the phenotype of mature B-cells. To evaluate this similarity more formally, we systematically matched cells from healthy bone marrow (H1–H5) with the MC surface marker profiles using linear discriminant analysis (Hastie et al., 2009). We modeled each MC with a Gaussian density in 16-dimensional space and a shared covariance matrix, using maximum likelihood to estimate parameters directly from the AML data. By representing each MC as a probability distribution, we can formally evaluate the posterior probability that each healthy cell and each MC are generated by the same source. Formally, we seek the quantity

$$P(MC = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}$$

$$f_k(x) = N(x \mid \mu_k, \Sigma_k)$$

for each healthy cell, x, where $\mu_k$is the mean surface marker expression for the AML subpopulations in MCk, $\Sigma_k = \Sigma\ \forall k$ is the shared covariance matrix, and $\pi_k = \frac{1}{14}\forall k$is a uniform prior over the 14 MCs. This quantity was computed for each cell in the healthy samples (approx. 3 million cells) using the classify function implemented in the MATLAB R2013b Statistics Toolbox. A high posterior probability for MCk indicates that cell x falls within the phenotypic boundaries estimated from the observed expression of subpopulations in MC_k. Cells that were an extremely good fit to one of the MCs (with $P(MC = k \mid X = x) > 0.99$) were considered *healthy cognates* to that MC.

Using this technique, 71–81% of each healthy sample was identified as a healthy cognate to one MC. Approximately 40% of each healthy sample was assigned to MC14, consistent with the phenotype and frequency of T cells in normal bone marrow aspirates (Clark et al., 1986). We used the proportion of each sample (healthy and AML) assigned to each MC 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 is associated with metacluster k, the score

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

quantifies whether cells associated with metacluster k were more frequently identified in healthy samples than in leukemic samples. Metaclusters for which $s_k > 1$ were overrepresented 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 cell (MC5), myeloid dendritic cell (MC8), erythroblast (MC9), granulocyte (MC10), NK cell (MC11), mature B cell (MC12), T cell (MC14). While large numbers of healthy cells were assigned to MC13 (~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), rare healthy cognates were identified independently in each normal marrow, suggesting that leukemic phenotypes do not depart radically from cells that occur in normal hematopoiesis.

Mapping a Cohort Landscape

A picture is worth a thousand words and therefore we sought to visualize the similarities and differences between detected subpopulations across patients. Each subpopulation was represented by its centroid, a 16-dimensional vector computed taking the median of each surface marker across all cells in that subpopulation. The phenotypic relationships between subpopulations were visualized in 2D using the nonlinear dimensionality reduction algorithm *t*-SNE (*t*-distributed stochastic neighbor embedding; van der Maaten, 2008). *t*-SNE projects high-dimensional data into a lower dimension, so that pairwise distances and global geometric features of the high-dimensional data are best preserved. *t*-SNE was previously shown to be effective for displaying the phenotypic landscape of individual cells, visually organizing these into subpopulations (Amir et al., 2013). We reasoned that it would also be effective at visualizing higher order relationships between the subpopulations themselves.

We chose to exclude normal lymphoid subpopulations to focus on the myeloid, leukemic, and progenitor populations and to get finer resolution on the relationships between them. We used the previously defined MCs to determine which populations were lymphoid cell types. Specifically, AML metaclusters MC 5, 9, 11, 12, and 14 were excluded as normal lymphoid or erythroid cells in leukemic samples (Fig. 3B–C). Similarly, healthy metaclusters HMC 2–8, 11, 13, 16, and 19 were excluded as lymphoid (Data S3A). The remaining 360 subpopulations (76 healthy and 284 AML) were used to construct the cohort landscape shown in Fig. 3A. We chose to simultaneously map subpopulations from both healthy and leukemic samples, so the healthy cell types could act as “landmarks” to aid interpretation of the leukemic subpopulations. All *t*-SNE mappings were computed using the Barnes-Hut C++ implementation (van der Maaten, 2014; http://lvdmaaten.github.io/tsne/) with *perplexity* = 30 and *theta* = 0.

Identification of Signaling Responses that Are Predictive of Healthy Cell Types

Metaclustering of the 5 normal marrows (described above) identified 20 healthy cell types (Data S3A), each of which was associated with a surface phenotype and a signaling phenotype of 224 signaling responses (i.e. response of phospho-marker X to perturbation Y). While surface markers are typically used to distinguish healthy immune cell types, we tested whether signaling responses could inform cell type in general, as well as highlighting which signaling responses were most informative in this regard.

To identify signaling responses significantly associated with the healthy cell type assignments, the 20 cell types were used as a categorical variable for the healthy subpopulations in analysis of variance (ANOVA) of each signaling response. Due to small

numbers of cells collected under three conditions (BEZ235, tumor necrosis factor α, and thrombopoeitin) it was common for clusters to have insufficient numbers of cells (less than 20) to calculate reliable SARA scores in these conditions. Thus, for many analyses we excluded these conditions, resulting in signaling phenotypes comprising 224 – (3x14) = 182 signaling responses. A large number of signaling responses had significant associations with cell type (e.g., 68/182 responses at FDR < 0.05), reported in Table S2. Four significant responses (G-CSF→pSTAT3, G-CSF→pSTAT5, FLT3L→pAKT and IL-10→pSTAT3) were selected for clustering the subpopulations in Figure 4C because these pathways had been previously observed in primitive (e.g., G-CSF, FLT3L) or differentiated (e.g., IL-10) hematopoietic cell types. To evaluate which surface markers were most informative in distinguishing between the cell types, ANOVA was applied similarly to the 16 surface markers.

Evaluation of PhenoGraph Classification

Our ultimate aim was to use PhenoGraph's modeling capabilities to infer "cell type" classifications for uncharacterized cells or subpopulations. In particular, we wanted to use this approach to generate data-driven characterizations of the leukemic subpopulations, using the healthy subpopulations as reference examples. To have confidence in this approach, we first tested the quality of PhenoGraph's classifications on our healthy samples, based on the ability to recover "held out" healthy cell type labels.

We pooled subpopulations from all 5 healthy samples and constructed a graph containing 191 nodes, each corresponding to a healthy subpopulation. We evaluated the performance of PhenoGraph classification under two scenarios, each based on a different distance metric.

- Using 16-dimensional surface phenotype to measure similarity between subpopulations. Each subpopulation was represented by its centroid, a 16-dimensional vector computed taking the median of each surface marker across all cells in that subpopulation.
- Using the 224-dimensional signaling phenotype to measure similarity, withholding all surface phenotype information. Each subpopulation was represented by the SARA derived z-scores (as described in the previous section). The z-score normalization (as described above) facilitated comparison of signaling responses across samples.

To further hone the graph toward accurate classification we reweighted the contribution of each dimension, increasing the weight of dimensions that are more informative towards distinguishing between classes. Specifically, each vector was reweighted using the negative logarithm of the ANOVA p-values as computed in the previous section. For subpopulations x and y with *D*-dimensional phenotypes, the distance was computed as:

$$\text{dist} = \sqrt{\sum_{d=1}^{D} -\log(p_d)(x_d - y_d)^2}$$

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. This was particularly important for the signaling phenotype, with so many dimensions (224), many of which were more informative for cell type than others (see Table S2). To maximize comparability between both surface and signaling metrics, we applied the same reweighting scheme to both surface and signaling based classification (i.e., weights were calculated as the negative logarithm of ANOVA p-values representing the association of each feature with healthy cell types).

In each case, the first neighbor graph was constructed with *k*=15 using the weighted Euclidean metric described above. Following construction of the initial nearest neighbor graph, the Jaccard metric was used to refine the neighbor graph exactly as before. This graph was then used for classification. We used a cross-validation scheme in which the subpopulations from 4/5 healthy samples were used as training data to classify subpopulations from the remaining healthy sample, repeated 5 times, each time with a different "held-out" healthy sample. The cell type metacluster assignments (see previous section) were used as class labels. In each iteration, all subpopulations in the training set were labeled with their class label and PhenoGraph classification was used to label the held-out samples. Performance was evaluated using the cross-validated correct classification rate (CCR), calculated as the percentage of cells in the held-out sample whose cell type was correctly recovered, averaged over the 5 repeats.

PhenoGraph Classification of Leukemic Subpopulations

We used the PhenoGraph classifier to assign leukemic subpopulations to phenotypic categories based on training examples provided by the healthy subpopulations. We treated surface and signaling phenotypes separately, building one classifier for each (Figure S4C). To do so, all 616 subpopulations (healthy and leukemic) were considered together and *k*-neighbor graphs (k = 15) were constructed using similarities derived either from surface phenotypes or signaling phenotypes. Specifically, we used a weighted Euclidean distance in which each phenotypic feature (e.g., CD34 among surface or G-CSF→pSTAT3 among signaling) was weighted according to its statistical association with known cell types in the healthy samples, as defined by ANOVA (described above). In each classifier, 191 healthy subpopulations (from the 5 healthy marrow samples) defined 20 cell type categories (Fig. S4A), and each AML subpopulation was assigned to one category based on its phenotypic proximity to these healthy training examples. Classification was performed once using a graph built from (weighted) surface phenotypes and once using a graph built from (weighted) signaling phenotypes to define the random walk probabilities of the PhenoGraph classification procedure. Each graph produced two alternative classifications per AML subpopulation.

Once all AML subpopulations were classified, we could characterize the population structure of each patient. We focused our analysis on the primitive phenotype—in this case, subpopulations assigned to the class defined by healthy metacluster 9 (HMC9; Fig. S4A). Each patient was scored for the proportion of cells falling into the HMC9 class on the basis of surface marker expression (%SDPC, Surface-Defined Primitive Cells) or signaling response pattern (%IFPC, Inferred Functionally Primitive Cells). This resulted in two alternative characterizations of each patient (Table S2, Fig. 5D), both based on similarity to HMC9, but one based on surface

marker similarity and the other based on functional similarity, inferred from signaling response pattern.

Informative Signaling Phenotypes for Distinguishing the Primitive Subpopulations

To identify the signaling features that were most informative for distinguishing primitive (e.g., IFPC) from mature (e.g., non-IFPC) leukemic subpopulations, we used canonical variates analysis (CVA), a multi-class, multi-dimensional generalization of Fisher's linear discriminant (Barber, 2011). CVA is similar to principal component analysis in that it seeks a linear projection of the high-dimensional data into a lower-dimensional space, but it uses information about the class membership of each observation to find a projection that maximizes the separability of these classes in the target space. With this method, one can visually examine the linear separability of the classes in low dimensional space, and identify the features that are important for obtaining that separability by examining the projection matrix.

In our case, the input data were the 224-dimensional signaling phenotypes for each subpopulation. We split the subpopulations into two classes based on whether they were assigned to the primitive or mature class by the PhenoGraph classifier using the signaling-based graph. CVA revealed that the linear separability of these classes could be maximally preserved in a single dimension (Fig. S5B). The signaling responses most important for class separability could be identified as the entries of the projection matrix with the largest (absolute) magnitudes (Table S2). In this way, CVA performs feature selection implicitly. Figure S5B (*bottom panel*) shows the analysis repeated, using only the top 4 positive-magnitude features and the largest negative-magnitude feature (i.e., selecting 5/224 signaling features for the analysis). The projection displays the extent to which these 5 features alone (G-CSF→pSTAT3, SCF→pAKT, G-CSF→pSTAT5, FLT3L→pAKT, and IL-10→pSTAT3) preserve the separability obtained with all 224 dimensions. Other signaling responses were strongly associated with primitive subpopulations despite being less powerful for classification. These were typically downstream effects of the signaling cascades triggered by SCF, FLT3L and G-CSF, affecting targets such as pCREB, pS6 and p4EBP1 (Table S2). Primitive subpopulations also displayed significantly increased activation of pSTAT1 by IFNα, a mechanism implicated in the control of HSC quiescence (Essers et al., 2009; Sato et al., 2009).

Gene Expression Signatures

For each score, %SDPC or %IFPC, a set of associated genes was defined based on correlation with the expression patterns across the patients. This method, known as in silico gene expression deconvolution (Lu et al., 2003; Stuart et al., 2004), assumes that a subpopulation of interest will express certain genes at constant rates; therefore, changes in bulk expression will track with changes in subpopulation size. This can be formalized as a linear regression problem:

$$Y = X\beta + \epsilon$$

in which Y is a N x G matrix of G mean-centered gene expression values, X is a N x 1 vector of the patient feature (e.g., %IFPC), and β is a 1 x G vector of regression coefficients. β can be

obtained from the least squares solution and represents, for each gene, the strength of its association with the patient feature.

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 one percentile more often than any other bin and had a standard deviation across data subsets of less than 5 percent.

This resulted in two gene expression signatures based on the %SDPC and %IFPC frequencies, containing 42 and 49 genes, respectively (Fig. 7A and Table S3). The mean expression of genes in each signature ("signature score") was significantly correlated with the frequency measurement (e.g., %SDPC or %IFPC) used to generate that signature (Fig. S6A). We therefore assume it could serve as a proxy to estimate %SDPC or %IFPC when single-cell data is not available. Thus, the expression score for each signature was calculated for patients in larger, independent cohorts for which gene expression and survival data, but not single-cell data, were available.

Survival Analysis

We used gene expression and survival data for 242 cytogenetically normal adult AML patients from two independent cohorts (Metzeler et al., 2008), consisting of arrays from two different Affymetrix platforms (U133A, U133 Plus 2.0). Cytogenetically normal patients are an important cohort for survival analysis, as they comprise ~45% of all AML cases and present an intermediate risk group whose survival is not dominated by karyotypic factors (Schlenk et al., 2008). Microarray data were preprocessed as described above. Each gene was centered such that its mean expression across arrays was 0. Metadata, also downloaded from GEO (ID # GSE12417), provided the following clinical annotations for each patient (array): karyotype, FAB, age, overall survival, right-censor. For each patient, the frequency of a cell type (%IFPC or %SDPC) was estimated as the mean expression intensity of the associated gene signature because single-cell data was not available. For Kaplan-Meier analysis (Fig. 7B and Fig. S6B), patients were stratified into two groups based on the median expression value of the signature of interest. For survival analysis, the overall survival and right-censor fields of the metadata were used to estimate the clinical significance of each gene signature.

For comparison between the IFPC signature and three published gene signatures (Eppert et al., 2011), signature scores were calculated as the mean signature expression, as described above. We noted that all three Eppert signatures contained one or two genes in common with the IFPC signature. In all cases, we removed these genes from the IFPC signature and allowed them to be retained by the Eppert signature. Cox proportional hazards regression was performed with the *coxphfit* function in MATLAB R2013b. Details and statistics are given in Table S3.

References

Aghaeepour, N. (2011). flowType: Phenotyping Flow Cytometry Assays. Bioconductor Repository.

Aghaeepour, N., Finak, G., FlowCAP Consortium, DREAM Consortium, Hoos, H., Mosmann, T.R., Brinkman, R., Gottardo, R., and Scheuermann, R.H. (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.

Becskei, A., Séraphin, B., and Serrano, L. (2001). Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. Embo J *20*, 2528–2535.

Bendall, S.C., Simonds, E.F., Qiu, P., Amir, E.-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 (New York, NY) *332*, 687–696.

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

Bodenmiller, B., Zunder, E.R., Finck, R., Chen, T.J., Savig, E.S., Bruggner, R.V., Simonds, E.F., Bendall, S.C., Sachs, K., Krutzik, P.O., et al. (2012). Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nat Biotechnol.

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

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.

Essers, M.A.G., Offner, S., Blanco-Bose, W.E., Waibler, Z., Kalinke, U., Duchosal, M.A., and Trumpp, A. (2009). IFNalpha activates dormant haematopoietic stem cells in vivo. Nature *458*, 904–908.

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

Fienberg, H.G., Simonds, E.F., Fantl, W.J., Nolan, G.P., and Bodenmiller, B. (2012). A platinum-based covalent viability reagent for single-cell mass cytometry. Cytometry n/a–n/a.

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.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning (New York, NY: Springer New York).

Kobourov, S.G. (2012). Spring Embedders and Force Directed Graph Drawing Algorithms. arXiv:1201.3011 [Cs].

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.

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

Novick, A., and Weiner, M. (1957). Enzyme Induction as an All-or-None Phenomenon. Proc Natl Acad Sci USA *43*, 553–566.

Qian, Y., Wei, C., Eun-Hyung Lee, F., Campbell, J., Halliley, J., Lee, J.A., Cai, J., Kong, Y.M., Sadat, E., Thomson, E., et al. (2010). 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 *78 Suppl 1*, S59–S82.

Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs, K.D., Jr., Bruggner, R.V., Linderman, M.D., Sachs, K., Nolan, G.P., and Plevritis, S.K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol *29*, 886–891.

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. Proceedings of the National Academy of Sciences *106*, 12944–12949.

Reichardt, J. (2008). Structure in Complex Networks (Berlin, Heidelberg: Springer Science & Business Media).

Sato, T., Onai, N., Yoshihara, H., Arai, F., Suda, T., and Ohteki, T. (2009). Interferon regulatory factor-2 protects quiescent hematopoietic ste... - PubMed - NCBI. Nat Med *15*, 696–700.

Schlenk, R.F., Döhner, K., Krauter, J., Fröhling, S., Corbacioglu, A., Bullinger, L., Habdank, M., Späth, D., Morgan, M., Benner, A., et al. (2008). Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. N Engl J Med *358*, 1909–1918.

Strehl, A., and Ghosh, J. (2003). Cluster ensembles --- a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research *3*, 583–617.

Stuart, R.O., Wachsman, W., Berry, C.C., Wang-Rodriguez, J., Wasserman, L., Klacansky, I., Masys, D., Arden, K., Goodison, S., McClelland, M., et al. (2004). In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. Proc Natl Acad Sci USA *101*,

615–620.

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

Zunder, E.R., Finck, R., Behbehani, G.K., and El-ad, D.A. (2015). Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. Nature.

Supplemental data legends