Additional supplementary methods

t-Distributed Stochastic Neighbor Embedding.

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

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

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

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

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

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

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

The joint similarity of x_i and x_i is defined as:

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

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

$$q_{ij} = \frac{(1 + \left|\left|y_i - y_j\right|\right|^2)^{-1}}{\Sigma_{k \neq l} (1 + \left|\left|y_k - y_l\right|\right|^2)^{-1}}$$ (Equation 5)

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

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

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

The gradient of the KL divergence between P and Q is derived in [1]:

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

The most challenging computational part of t-SNE is computation and storage of the pairwise similarities in high- and low- dimensional space. Given an input X with n points, each of these matrices will consist of n^2 values. For example, for n = 100,000, ten billion values are needed. For high values of n, this requirement is beyond the capability of modern computers. However, all of t-SNE's calculations are local: for a given point y_i in low-dimensional space, only the vectors p_i and q_i are needed to calculate the next iteration for y_i.

viSNE is a distributed implementation of t-SNE that relies on the locality of t-SNE's calculations. Given N cores, viSNE splits the input into partitions where each partition contains n/N points. Each core receives one partition. Instead of storing the entire matrix (n^2 values), each core only stores the submatrix consisting of its n/N points, times all other points (n^2/N values). Going back to the previous example, for n = 100,000 and N=64 cores, each core will need to store approximately 150 million values per matrix. For each iteration, each core locally updates Y_i for points in its partition and broadcasts these values to the other cores, guaranteeing that all cores have the updated similarity matrix.

The *cyt* visualization tool.

cyt is an interactive visualization tool designed for the analysis of viSNE maps and the high-dimensional mass or flow cytometry data from which these maps were projected. It plots viSNE maps as scatter and density plots,

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

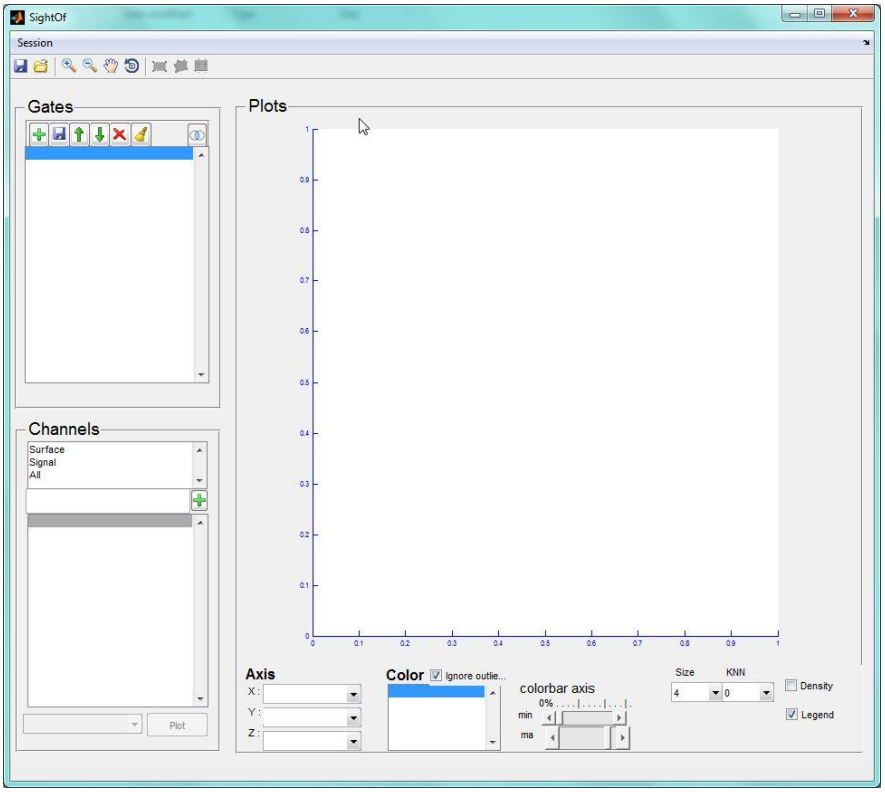
cyt contributes in correlating multiple viSNE maps of the same data. While viSNE consistently separates the various immune subtypes, their position on different maps could vary (most often due to rotations and reflections of the map). In healthy samples, we initially colored the immune subtypes using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

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

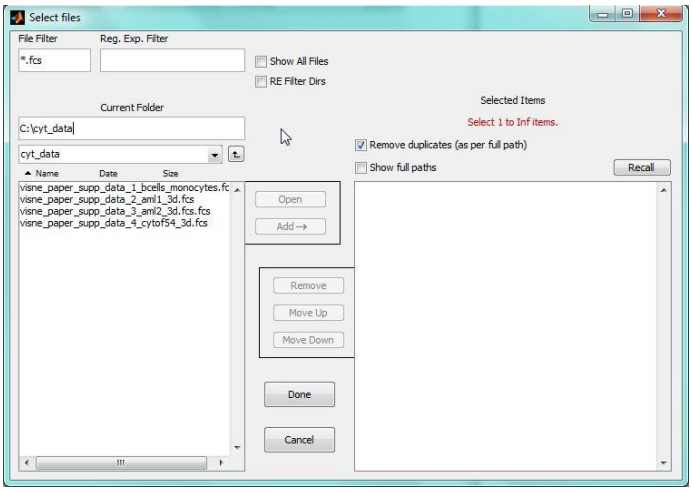
cyt is written in Matlab and is available along with viSNE at http://www.c2b2.columbia.edu/danapeerlab/html/index.html.

Using cyt to visualize 2D and 3D viSNE maps.

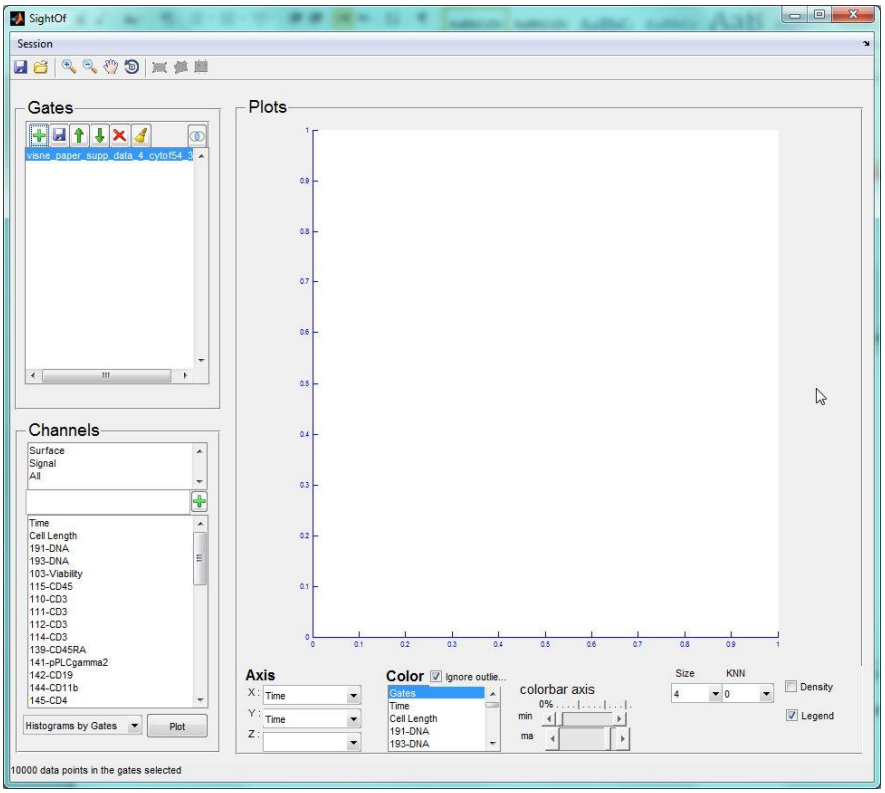
This paper's supplement includes four FCS files that can be visualized and explored using *cyt*. To launch *cyt*, unzip cyt.zip to a folder of your choice, launch Matlab, direct it to *cyt*'s folder, and run the script *run_cyt.m*. You might wish to copy the FCS files to that folder as well (alternatively, you can load them into *cyt* from a different folder). You will be greeted by the following:



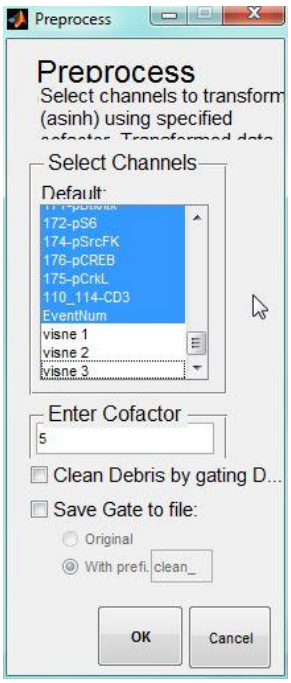
In the "Gates" panel, you can load and export FCS files, rearrange the order of FCS files and gates, remove gates, and transform the data using hyperbolic arcsin (the brush icon). Click the load button (the green plus icon) to load a FCS file:



Click on a file's name and click "Add" to add it to the selected items list. For this tutorial, add "visne_paper_supp_data_4_cytof54_3d.fcs". Click Done, and the file will be loaded:

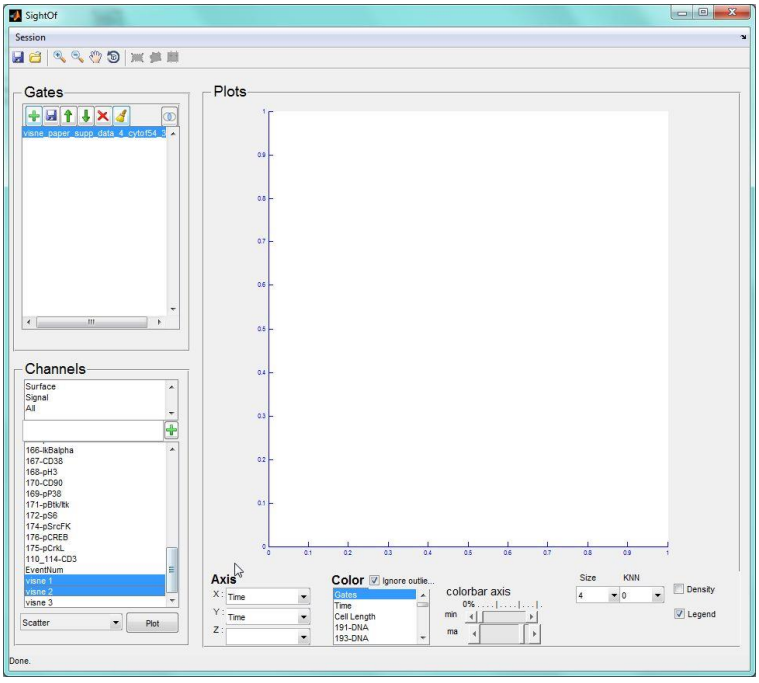


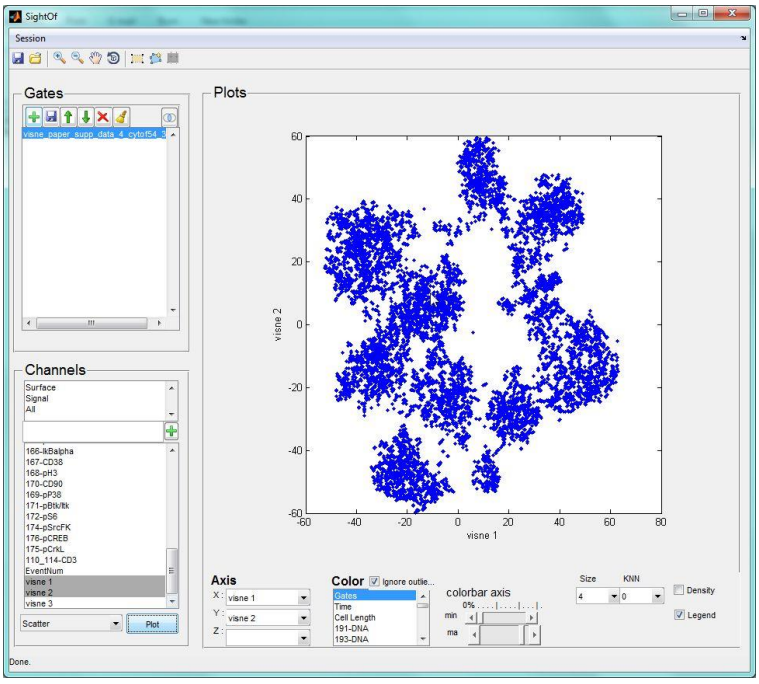
The first step when working with CyTOF data is to transform it using hyperbolic arcsin. Click the brush icon at the top left:



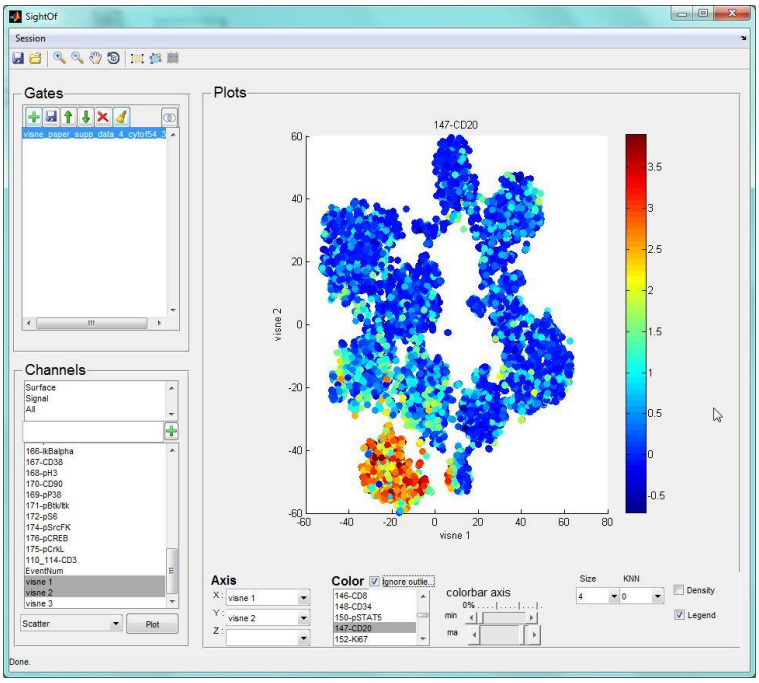
Uncheck the three viSNE channels by pressing Ctrl and clicking on the channel names. Next, press OK and the data will be transformed.

You can now visualize channel intensities over the viSNE channels. To do so, first use the channel panel to pick the two viSNE channels; then, click Plot:

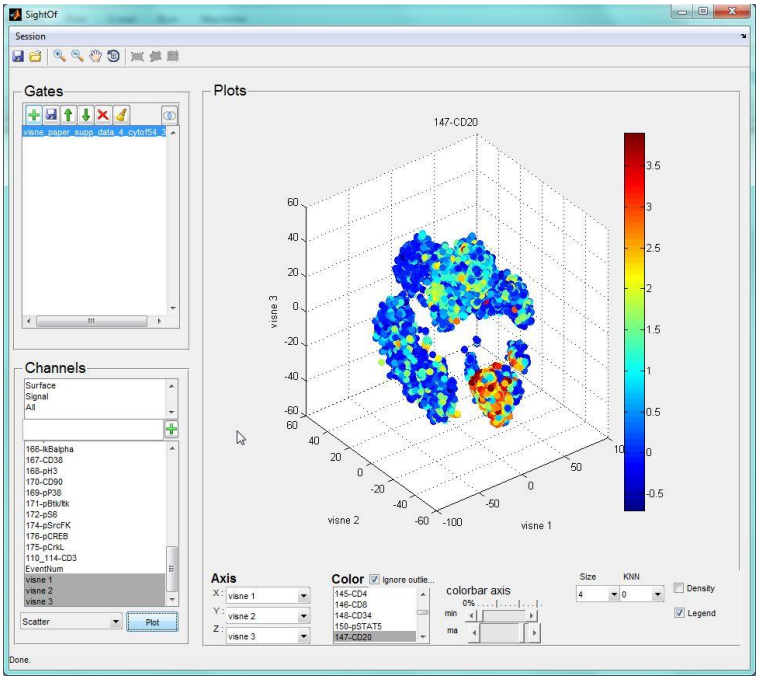




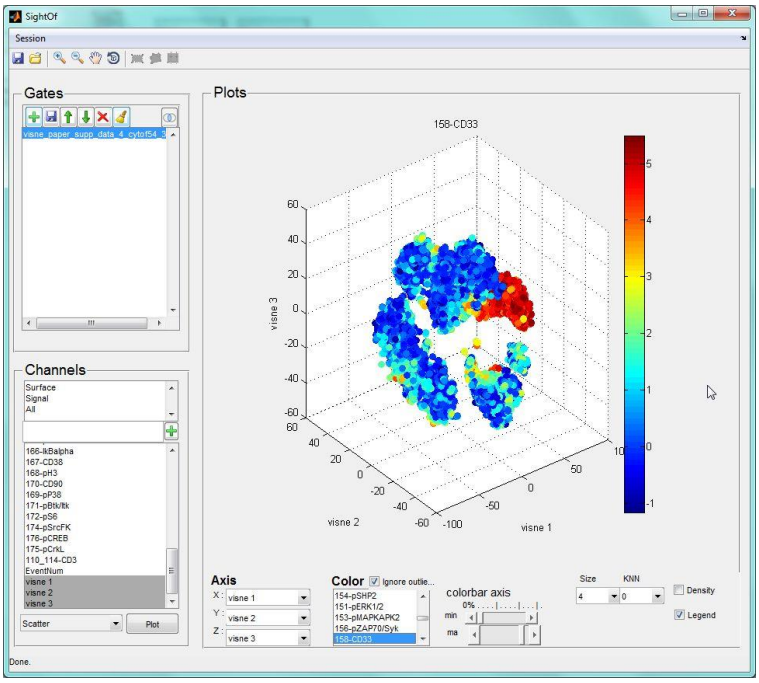
You can now color code cells by channel intensities by picking the channel under the Color panel at the bottom. For example, scroll down and pick CD20:



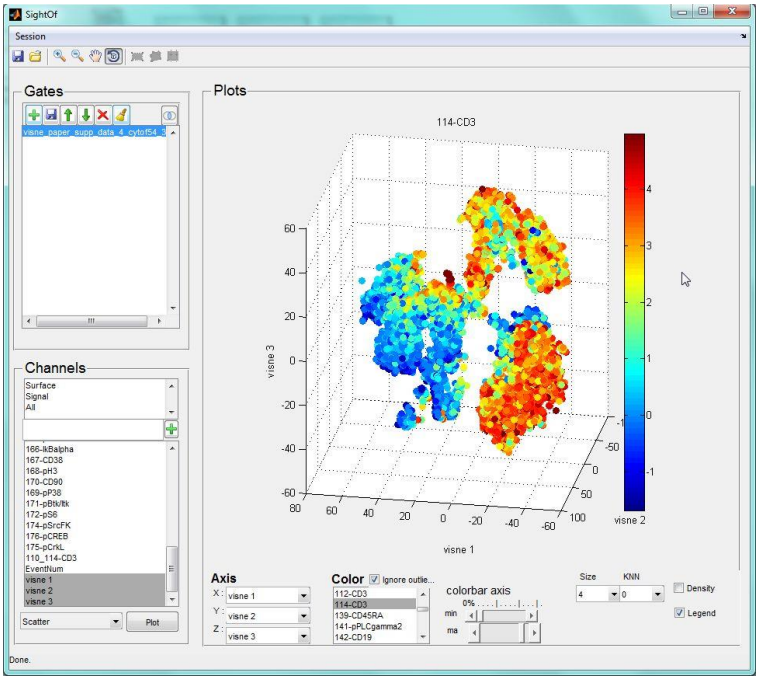
These files include a third viSNE channel, allowing visualization of the data in 3D. Pick all three viSNE channels and click Plot; *cyt* will retain the CD20 color coding:



You can visualize the other channels by picking another channel from the Color panel. For example, scroll and pick CD33 to see the separation between B cells and myeloids:



Finally, you can rotate the 3D view using the rotate tool from the top bar (the circular arrow). Here is an example with CD3 coding, slightly rotated to highlight CD4+ and CD8+ T cells:



The Jensen-Shannon divergence.

The Jensen-Shannon (JS) divergence is an information theory-based, symmetric measure of the similarity between two probability distributions. It is defined as:

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

Where M is:

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

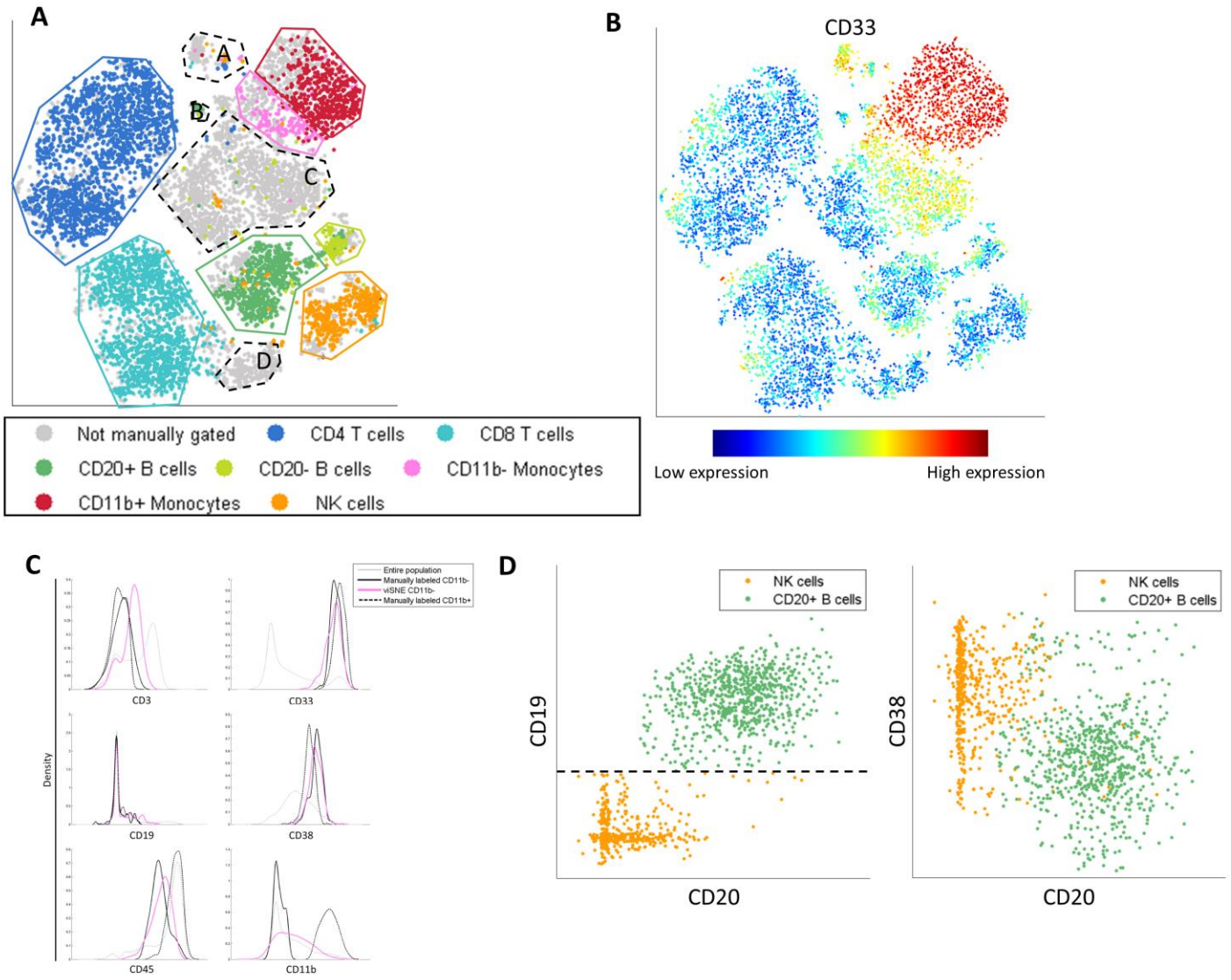
And KL is the Kullback-Leibler divergence:

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

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

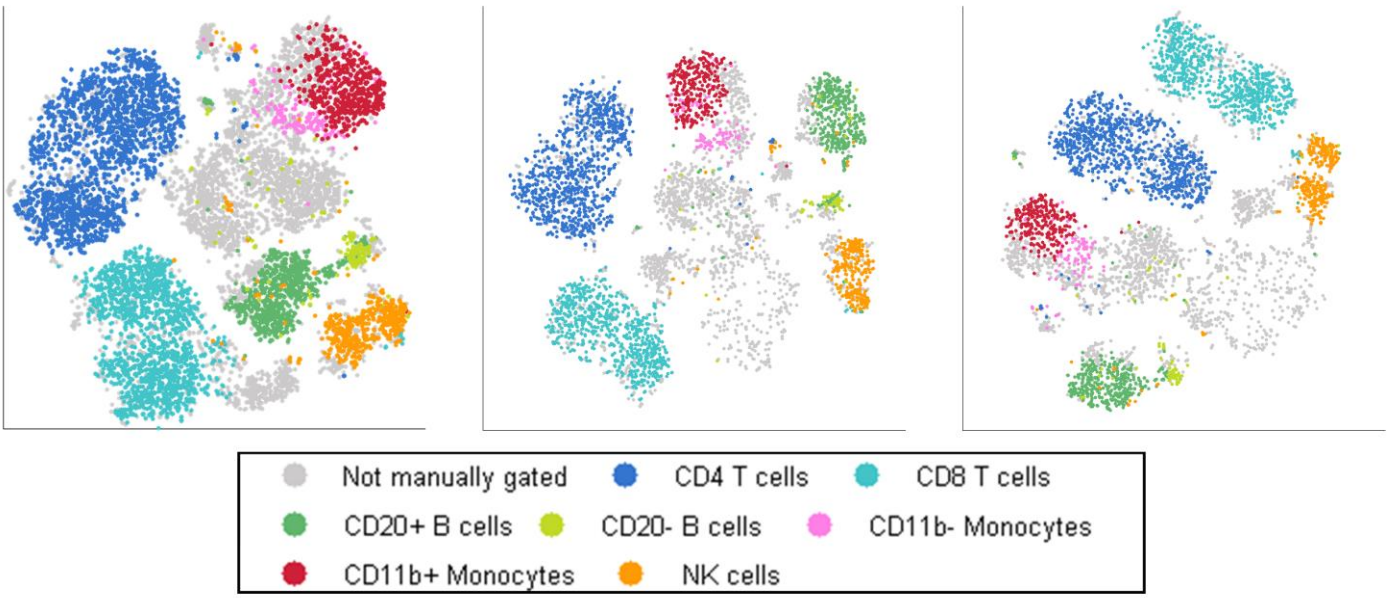
References

1. Bendall, S.C., et al., *Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum.* Science, 2011. **332**(6030): p. 687-96.
2. van Dongen, J.J., A. Orfao, and C. EuroFlow, *EuroFlow: Resetting leukemia and lymphoma immunophenotyping. Basis for companion diagnostics and personalized medicine.* Leukemia, 2012. **26**(9): p. 1899-907.
3. Jolliffe, I., *Principal component analysis*. 2005: Wiley Online Library.
4. Tenenbaum, J.B., V. de Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction.* Science, 2000. **290**(5500): p. 2319-23.
5. Roweis, S.T. and L.K. Saul, *Nonlinear dimensionality reduction by locally linear embedding.* Science, 2000. **290**(5500): p. 2323-6.
6. Shawe-Taylor, J. and N. Cristianini, *Kernel methods for pattern analysis*. 2004: Cambridge university press.
7. Van der Maaten, L., E. Postma, and H. Van Den Herik, *Dimensionality reduction: A comparative review.* Journal of Machine Learning Research, 2009. **10**: p. 1-41.
8. Qian, Y., et al., *Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data.* Cytometry B Clin Cytom, 2010. **78 Suppl 1**: p. S69-82.
9. Van der Maaten, L. and G. Hinton, *Visualizing data using t-SNE.* Journal of Machine Learning Research, 2008. **9**(2579-2605): p. 85.



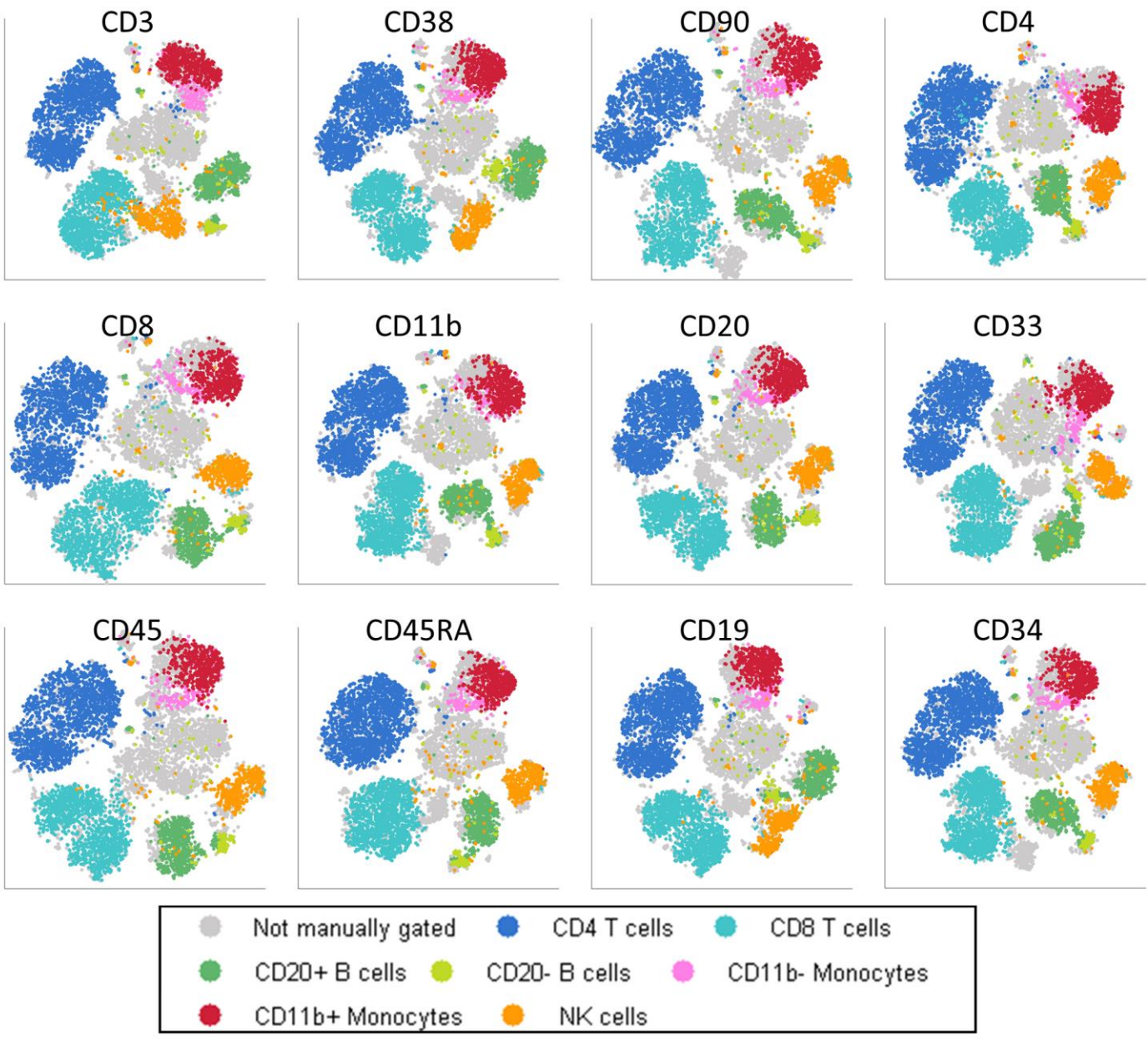
Supplementary Figure 1: A viSNE map can classify cells that were labeled incorrectly by manual gating. (A) viSNE map is identical to Figure 1B. Each of the cell subtypes is surrounded by a gate corresponding to that subtype's color as designated by manual gating. Grey points inside the viSNE map gate were not classified by manual gating, although further examination reveals that they belong to the relevant subtype. In addition, there are four regions that do not conform to known subtypes. Region A has cells positive for CD45 only, we suspect these are missing a marker needed to classify them. Region B includes doublets: pairs of cells that were read together by the machine as if a single cell, and are therefore positive for suspect marker combinations (for example, CD19+ CD11b+). The cells in Region C are negative for all channels and are probably debris. Region D has cells positive for CD45 and CD3 only, these might be NKT cells whose canonical marker is missing. (B)

The same map as in A, coded by CD33 (myeloid marker) expression. The monocyte cell population is clearly visible to the top right. (C) Marker expression level densities for the entire population (grey), the manually gated CD11b- monocyte cells (black) and the viSNE gated CD11b- monocytes cells (pink, as shown in panel A). The two CD11b- monocyte populations have almost identical marker distributions, except the pink population shows some CD3 staining due to reagent "stickiness", but this level of CD3 expression is significantly lower than that on bona fide CD3+ cells. Rather than exclude these cells based on a hard threshold on a single marker, viSNE groups the cells together based on their tight similarity in all other markers. Taken together these data support viSNE's identification of CD11b- monocytes. (D) Gating using the viSNE map is typically more accurate than manual gating, since 2D views and hard thresholds can be misleading. There are a number of cells labeled as CD20+ B-cells by viSNE (dark green outline in A) and labeled NK cells based on biaxial gating (orange dots within the B-cell region). These cells just miss the hard CD19 threshold in one of the gates (black dashed line) and therefore their CD20 level is never examined during the gating (the presented biaxial plot is not part of the gating scheme used). Their high CD20 levels and borderline CD19 levels support their labeling as B-cells, rather than NK-cells. Their CD38 levels further support their B-cell label.



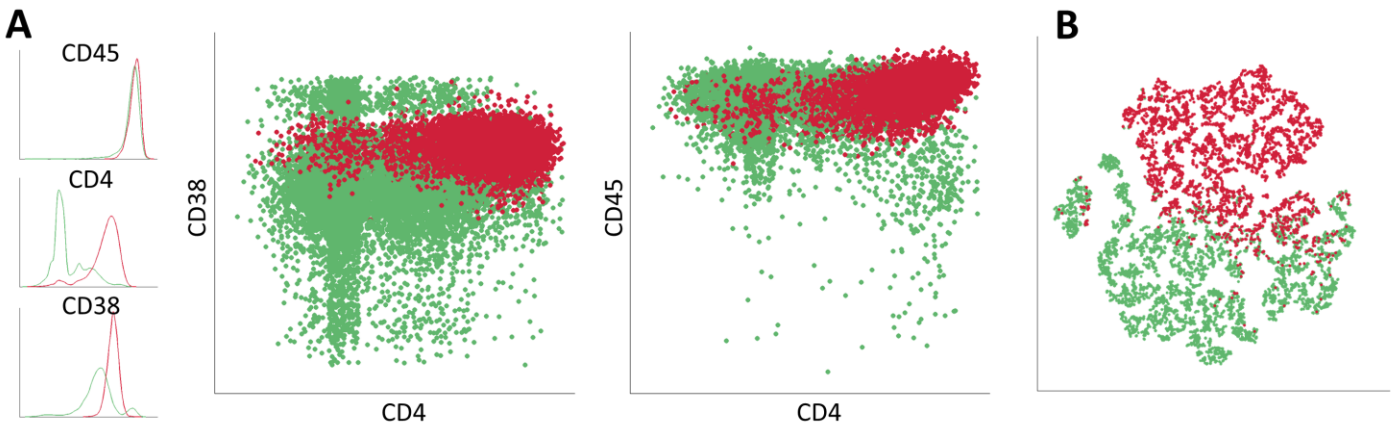
Supplementary Figure 2: viSNE is robust to subsampling. Each map includes a different subsample of 6,000 cells from Marrow1. Despite using different cells for the viSNE runs, we see that the healthy subtypes are

similarly separated, with the same subpopulations identified. The axes in viSNE are arbitrary and therefore maps often vary in rotation and reflection, as these are defined based on pairwise distances. The cyt tool helps match between equivalent subpopulations (see Using cyt to visualize 2D and 3D viSNE maps).

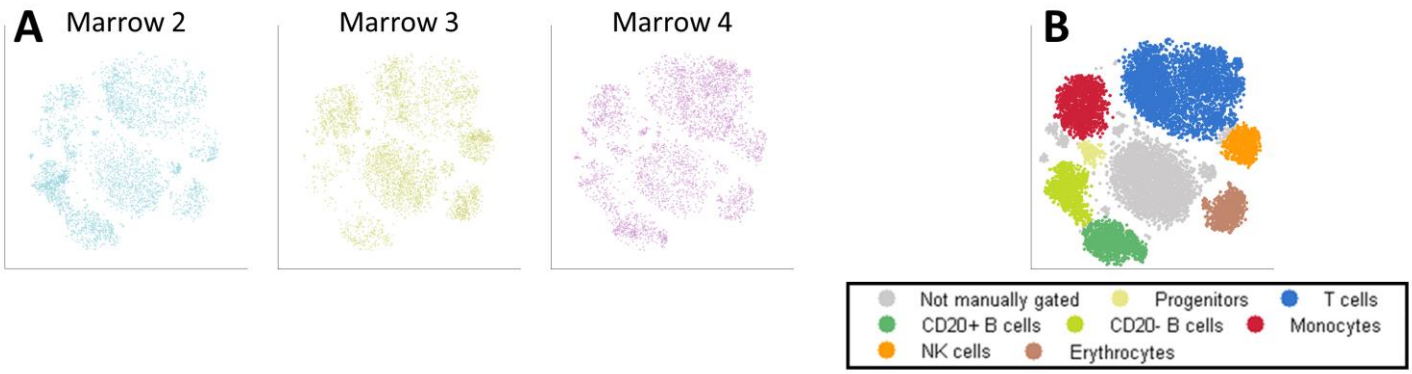


Supplementary Figure 3: viSNE is robust to marker panel selection. Leave-one-out viSNE maps of Marrow1 are shown. Each panel is a viSNE map with twelve markers; the marker listed above the panel was excluded in that run. The data itself is the same data as Figure 1B. Despite removing markers, subtypes are identified and separated correctly in each panel. The different maps are almost identical to each other. In this figure the same

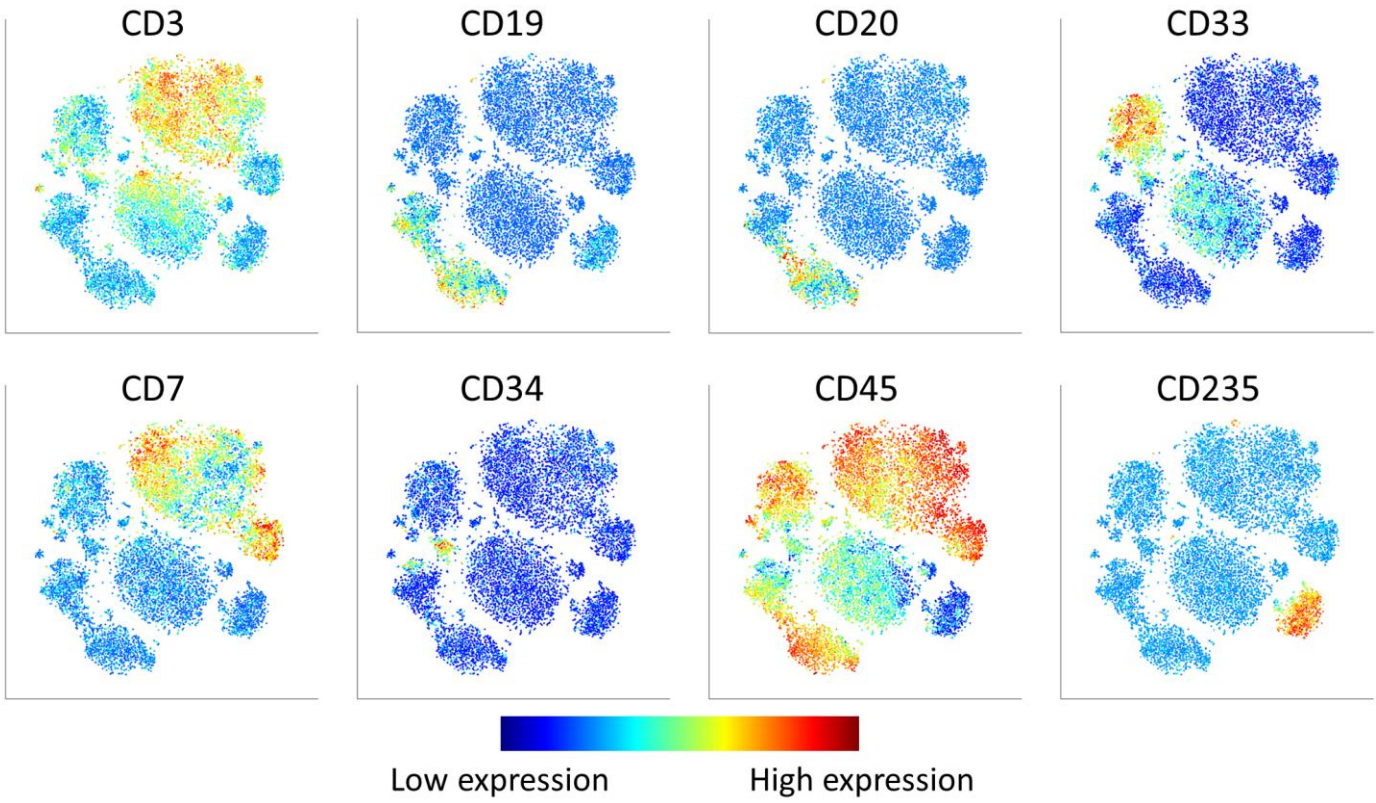
cells are used to make all maps, in contrast to Supplementary Figure 2, that used different cells from the same sample in each run. As such, here the maps look similar than they do in Supplementary Figure 2.



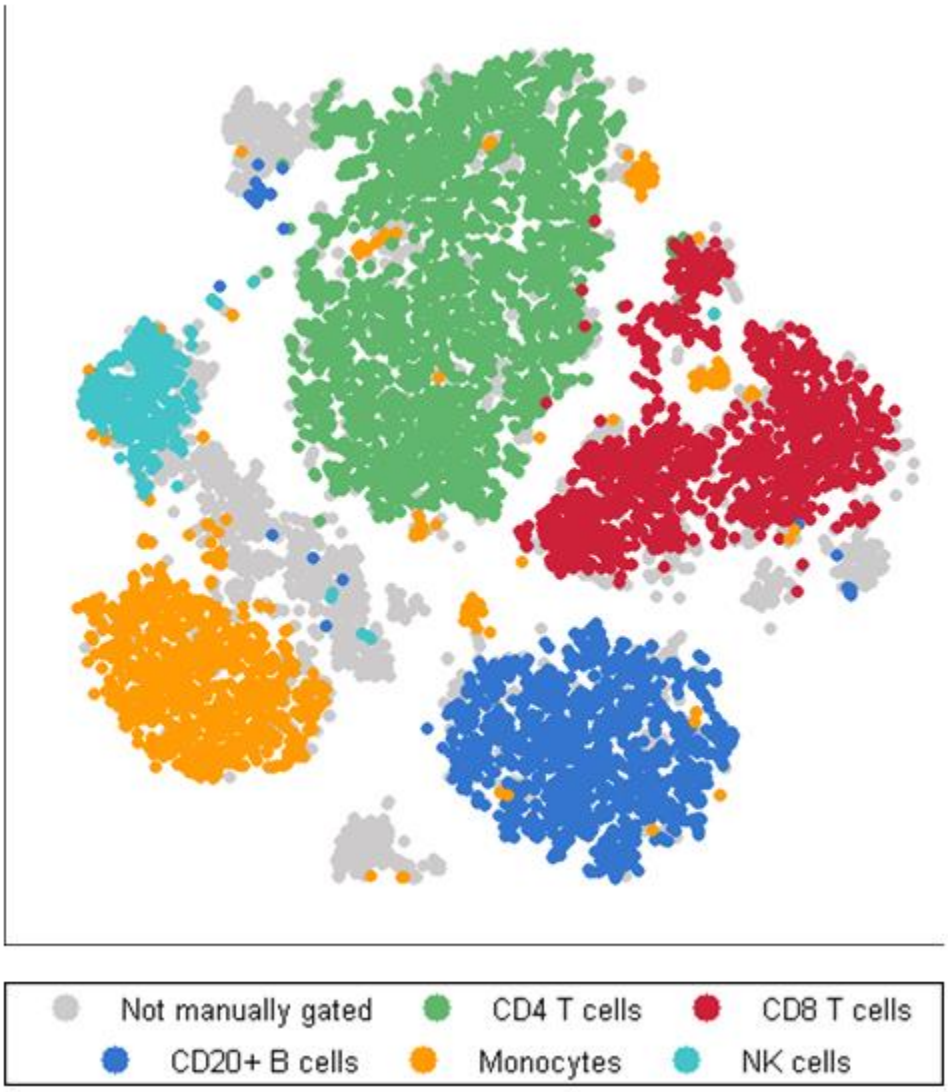
Supplementary Figure 4: We created a toy example from the Marrow1 data to better understand how non-canonical markers can be used to separate monocyte cells and B cells. (A) Demonstrates how three non-canonical markers can separate monocytes (red) and B cells (green), labeled as in Figure 1C. We deliberately chose three markers (CD45, CD4 and CD38) that cannot separate these cell types in either 1 or 2 dimensions: left, the one-dimensional distribution of each marker, x-axis represents marker intensity and y-axis represents density of cells at this intensity. Both populations overlap in all dimensions. Middle: biaxial scatter plots of CD4 (x-axis) and CD38 (y-axis), and CD4 (x-axis) and CD45 (y-axis), demonstrating overlap of the two populations in two dimensions. (B) A viSNE map of the same two populations, colored by cell type, monocytes (red) and B cells (green). While the two subsets are not separated in any linear projection to two dimensions, viSNE successfully separates these two populations. The two populations are disconnected in three-dimensional space, each residing in a different region: each subtype takes an irregular (non-ellipsoidal) shape (see Supplementary Data 1: Marrow 1 toy example) and is separated from the other by a space void of any cells. viSNE identifies these shapes and the distinction between them, thus separating the two subtypes. This example is available in Supplementary Data 1 and can be visualized in *cyt* using the instructions provided above (see Using cyt to visualize 2D and 3D viSNE maps).



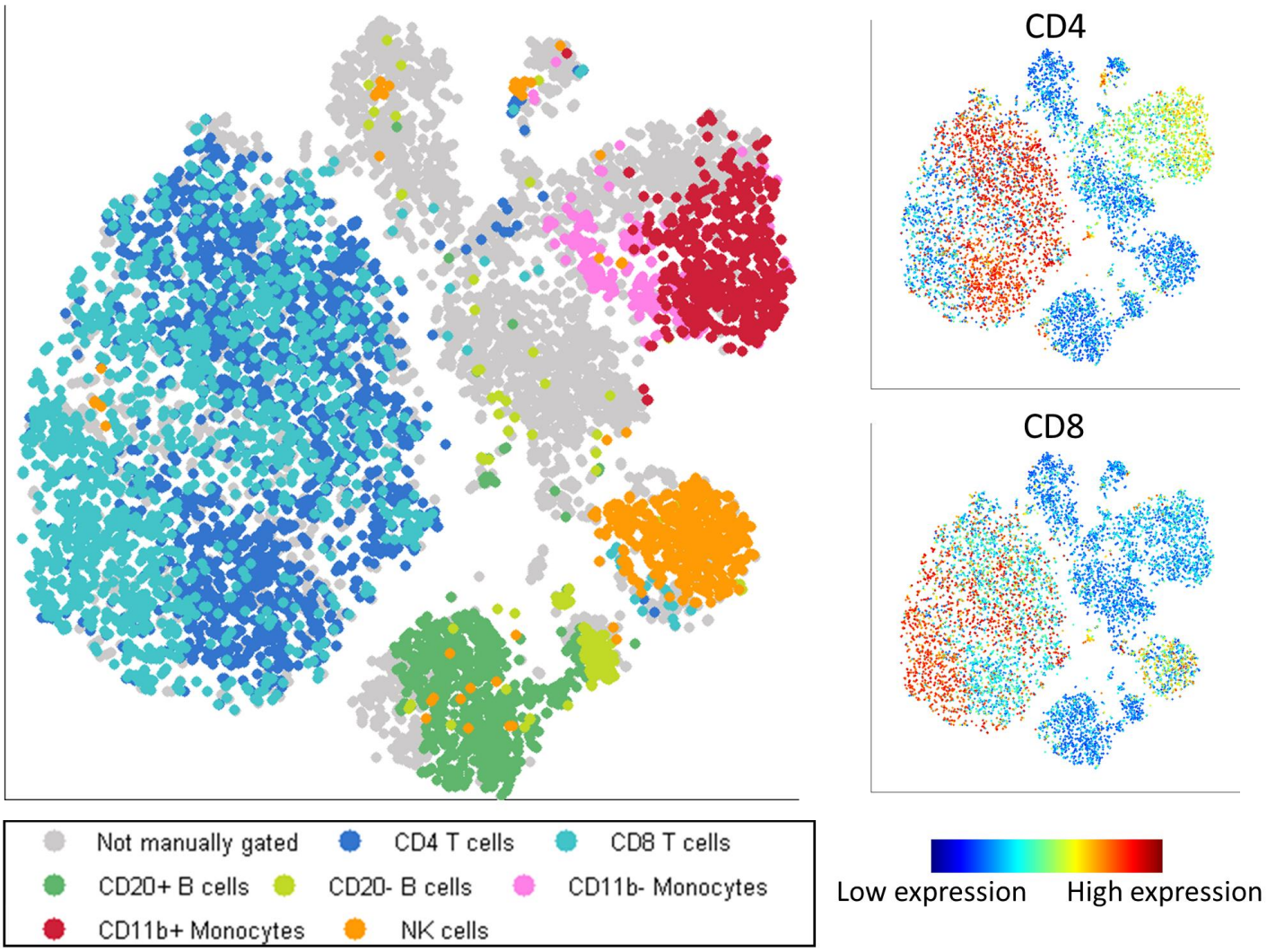
Supplementary Figure 5: The viSNE map has a conserved structure across healthy samples. (A) Bone marrow samples from three healthy donors (2-4) were mapped together using viSNE. This is the same data presented in Figure 2B. Each of the 3 plots represents a different individual. The viSNE plots across different individuals are very similar (Jenson-Shannon divergence of ~0.04), share the same cell types (see B) and shapes. Each individual shares the same cell populations, in different frequencies. (B) Same as Figure 2C, color coded by subtypes, as identified by marker expression levels (See Supplementary Fig. 6).



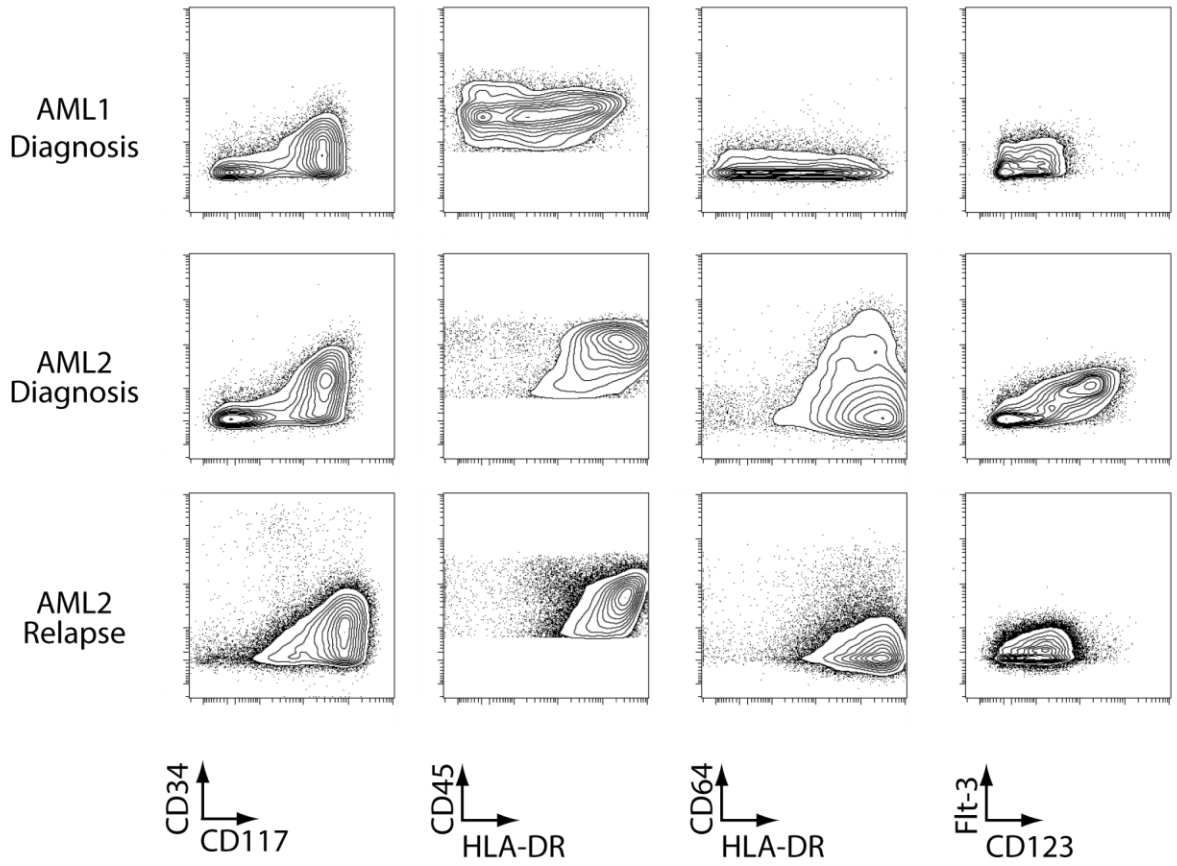
Supplementary Figure 6: Support of cell subset labeling in Figure 3B. The viSNE map of Figure 2B colored by expression level of the markers used to identify the different immune subsets. Gating logic follows the strategy in [2]. For example, T cells are CD3+ CD19- CD20- CD33- CD34- CD45+.



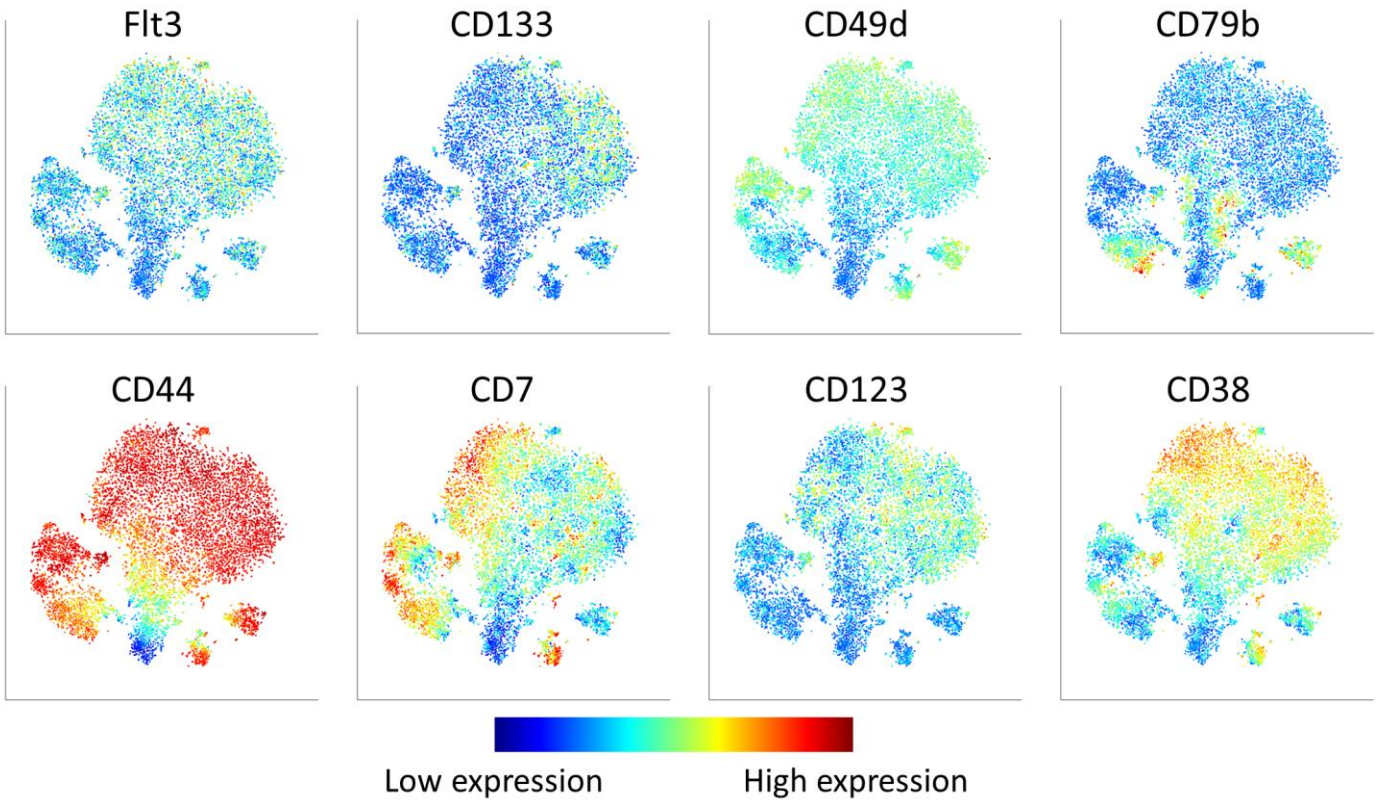
Supplementary Figure 7: viSNE can be used for the analysis of flow cytometry data. viSNE map of flow cytometry data of healthy bone marrow. The flow cytometry panel includes eight markers (CD45, CD45RA, CD3, CD4, CD8, CD33, CD20, CD7) and therefore identifies fewer subsets.



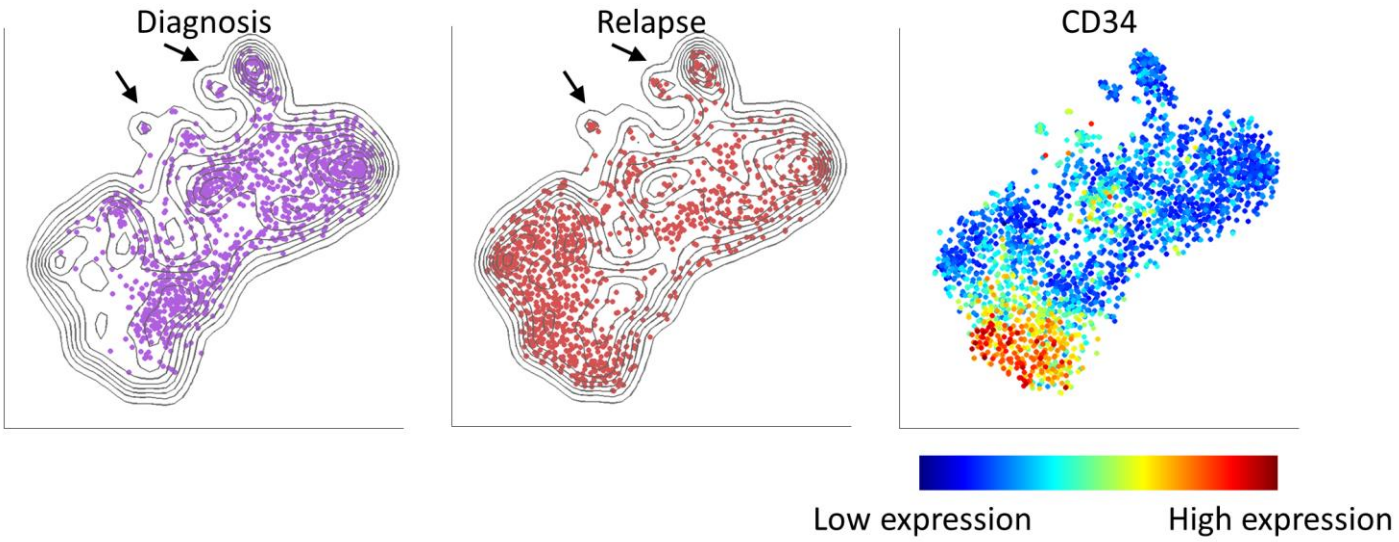
Supplementary Figure 8: Left: viSNE map of Marrow1, without CD4 and CD8. All non-T cell subtypes are identified well. There is an overlap between CD4 and CD8; however, information in the other channels allows a partial separation between them. Right: The same map, color coded by CD4 (top) and CD8 (bottom), again showing the partial separation between these subtypes despite the fact that both markers are missing from the mapping.



Supplementary Figure 9: Biaxial plots of the AML samples used in Figures 3 and 4 using the standard biaxial plots and marker combinations typically used to study AML [3].

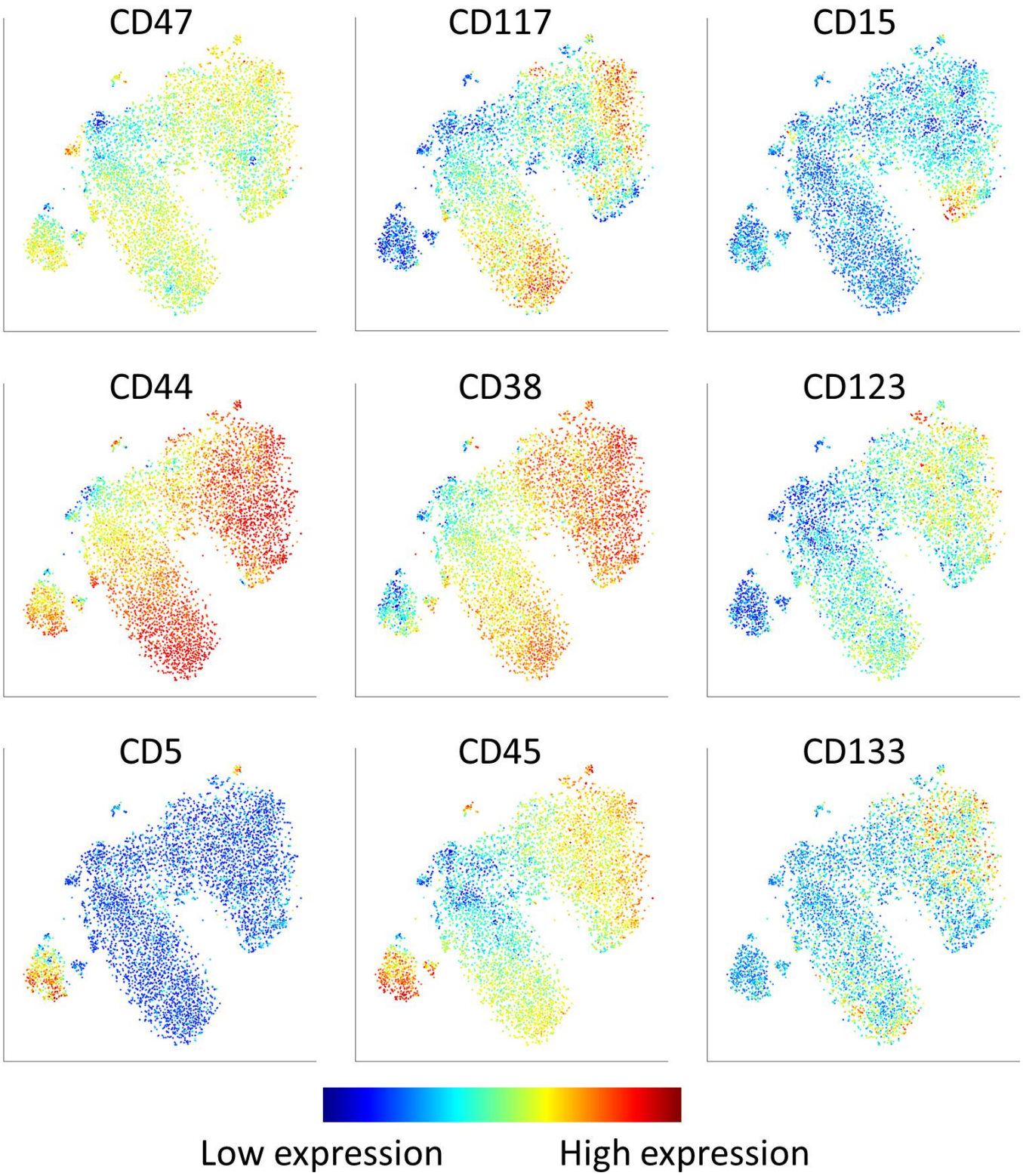


Supplementary Figure 10: Additional characterization of viSNE map of AML 1 (as in Fig. 3B), cells colored by the expression levels of the marker in the panel's title.

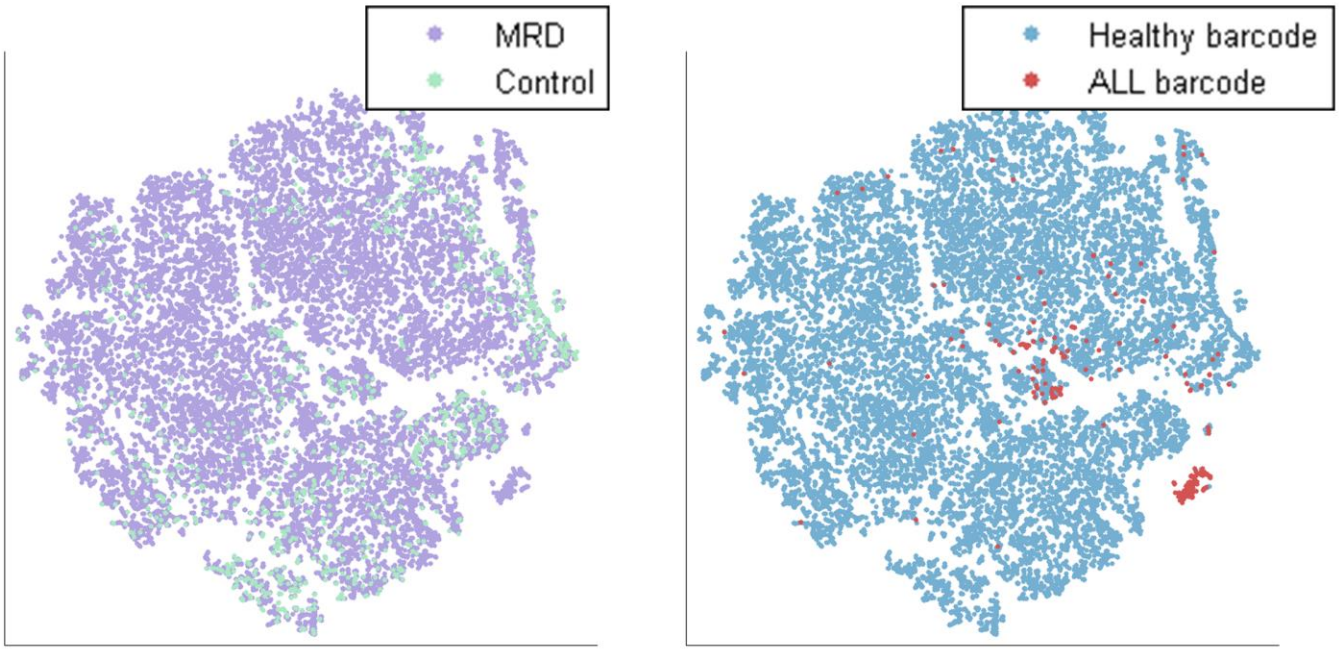


Supplementary Figure 11: viSNE separates an additional diagnosis-relapse pair (patient AML A). Left: viSNE map of diagnosis (left, purple) and relapse (right, red) samples of an additional patient. We see two small healthy populations (black arrows) and a large cancer region which includes both separate and overlapping

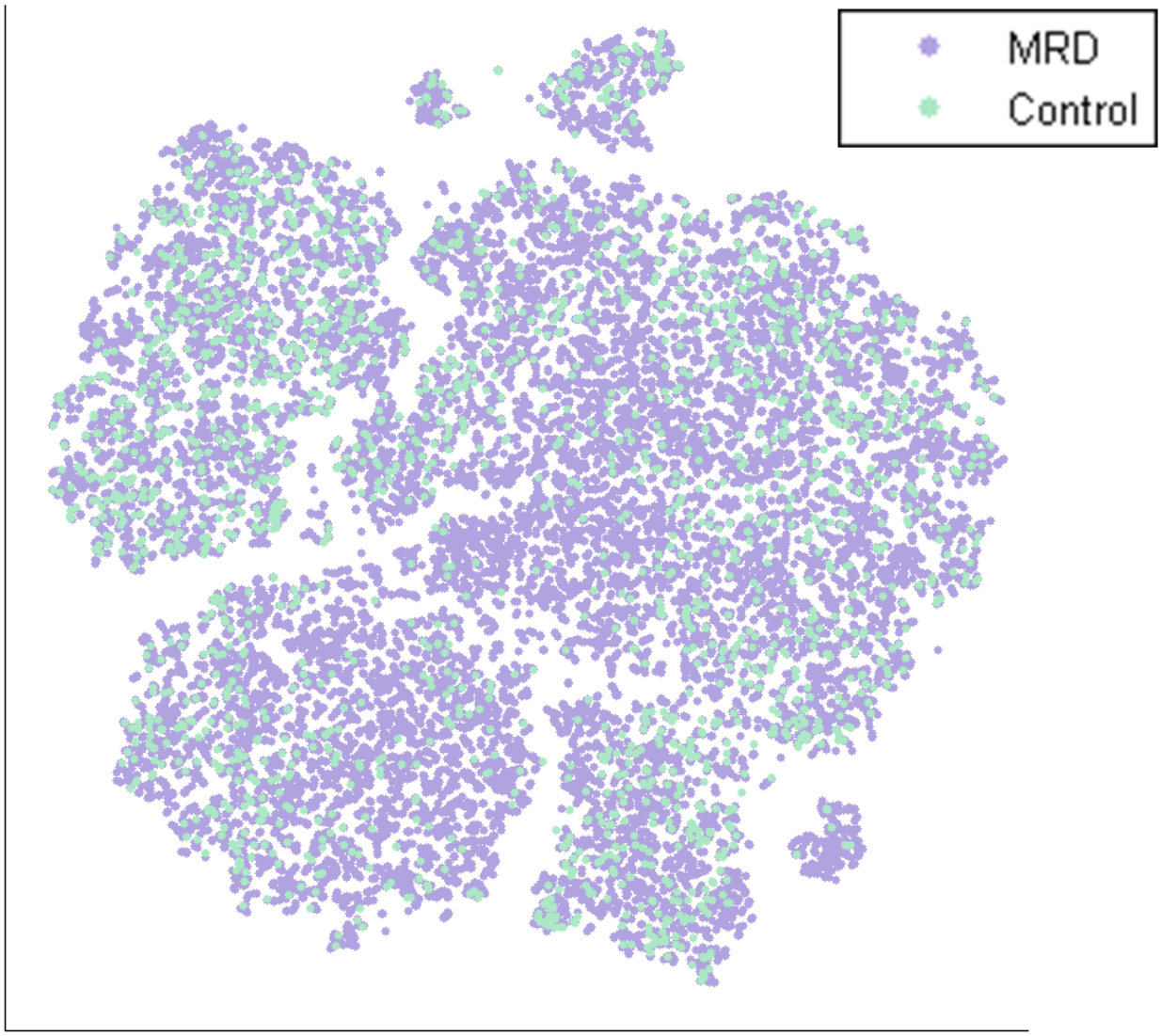
regions between diagnosis (purple) and relapse (red). Right: The same map, colored by CD34 expression levels. We see a gradient of CD34 that begins in the diagnosis sample and reaches its peak in the relapse region.



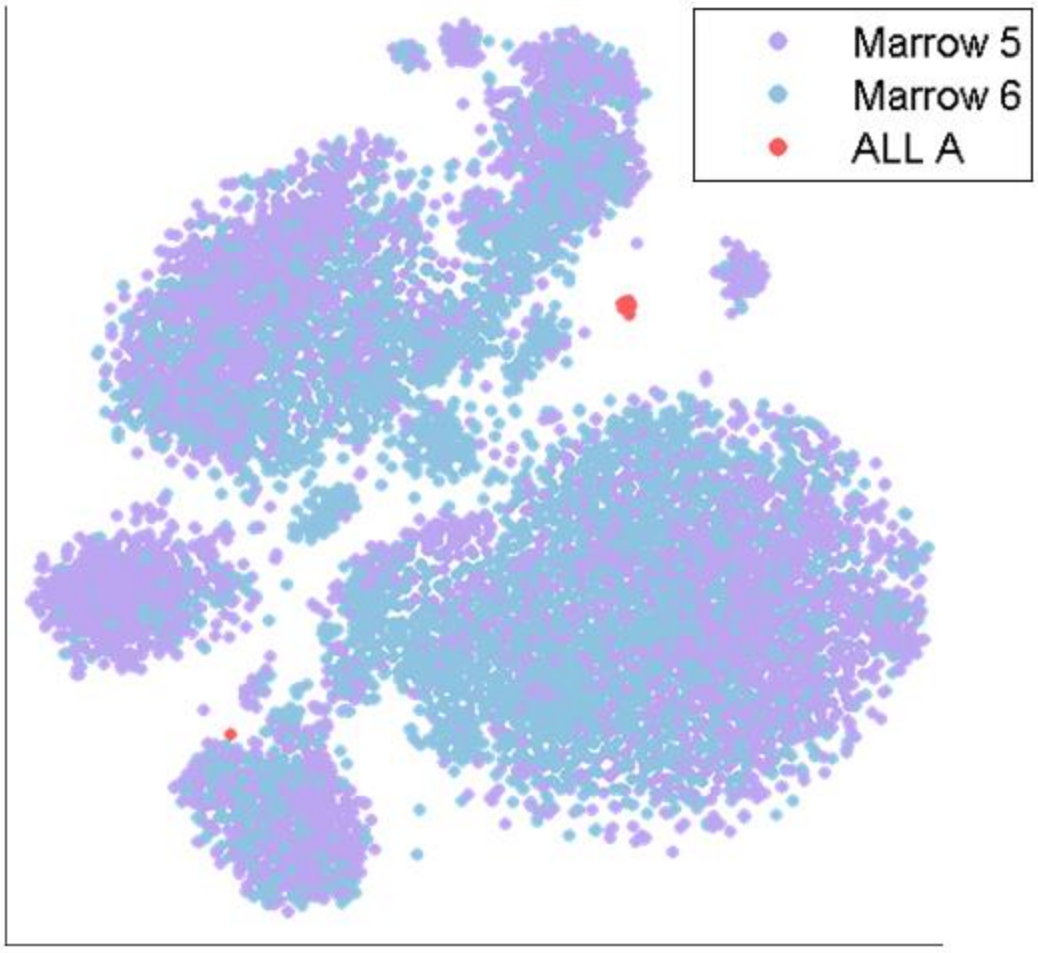
Supplementary Figure 12: viSNE map of diagnosis and relapse AML B samples (same map as in Fig. 5), colored by expression level of indicated markers.



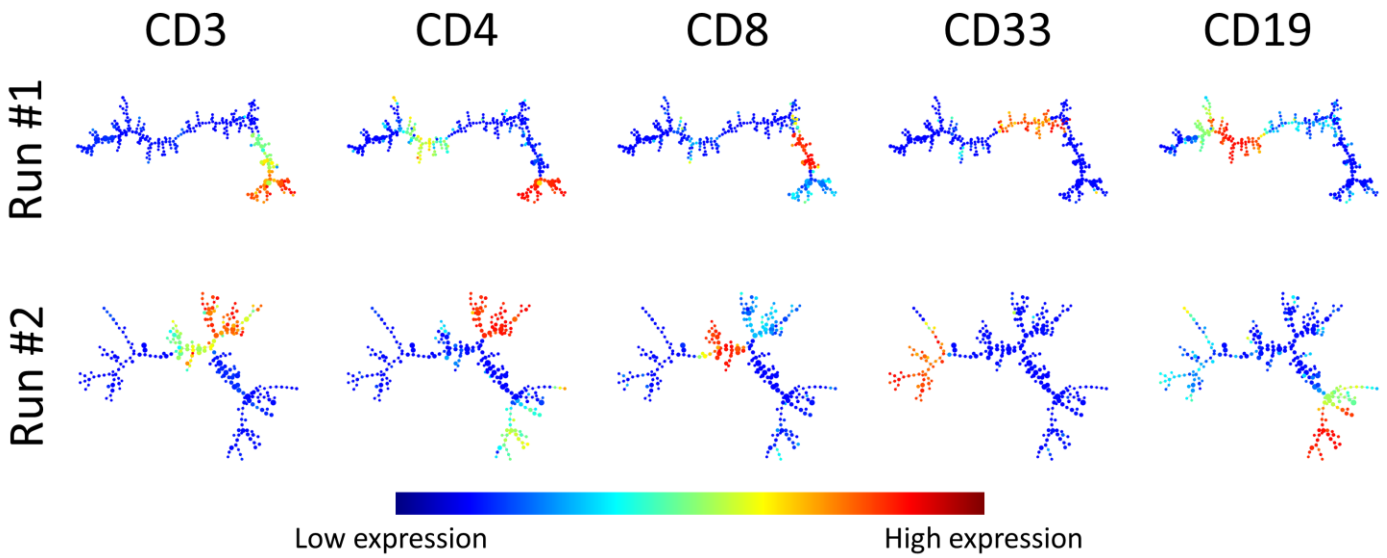
Supplementary Figure 13: MRD detection works with different channels. This figure follows the same structure of Figures 6A and 6C. The eight channels used here are CD38, CD34, CD10, CD45, CD33, HLA-DR, surface IgM and a "dump" channel that combines CD235, CD62 and CD66b. As in the healthy samples (Fig. 2A), viSNE is robust to the panel of measured markers.



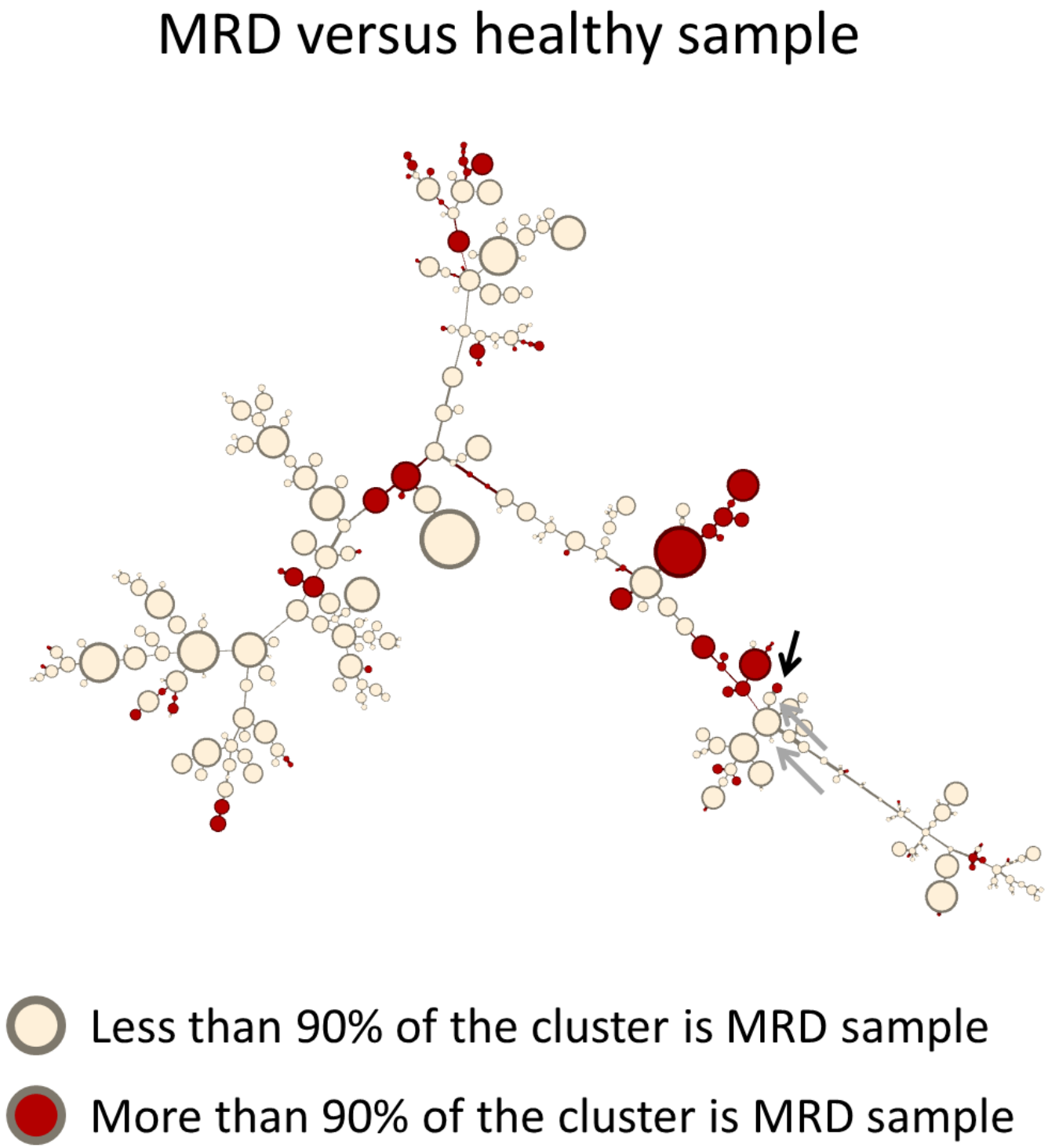
Supplementary Figure 14: MRD algorithm does not detect a suspect region in a healthy sample. A viSNE map of healthy vs. healthy run of the MRD detection algorithm, colored by sample. The "MRD" sample is composed of healthy cells only. Unlike Figure 6 and Supplementary Figure 13, we do not see a region which has distinctly cells from one sample, demonstrating that the subsampling method and viSNE do not spuriously create suspect regions.



Supplementary Figure 15: viSNE can detect and separate a rare subpopulation. Ten thousand cells from two healthy samples (Marrow5 and Marrow6, purple and blue respectively) were combined computationally with twenty cells from an ALL sample (ALL A, red). Despite being a very small proportion of the combined sample (0.2%), viSNE still manages to identify and separate the ALL cells.

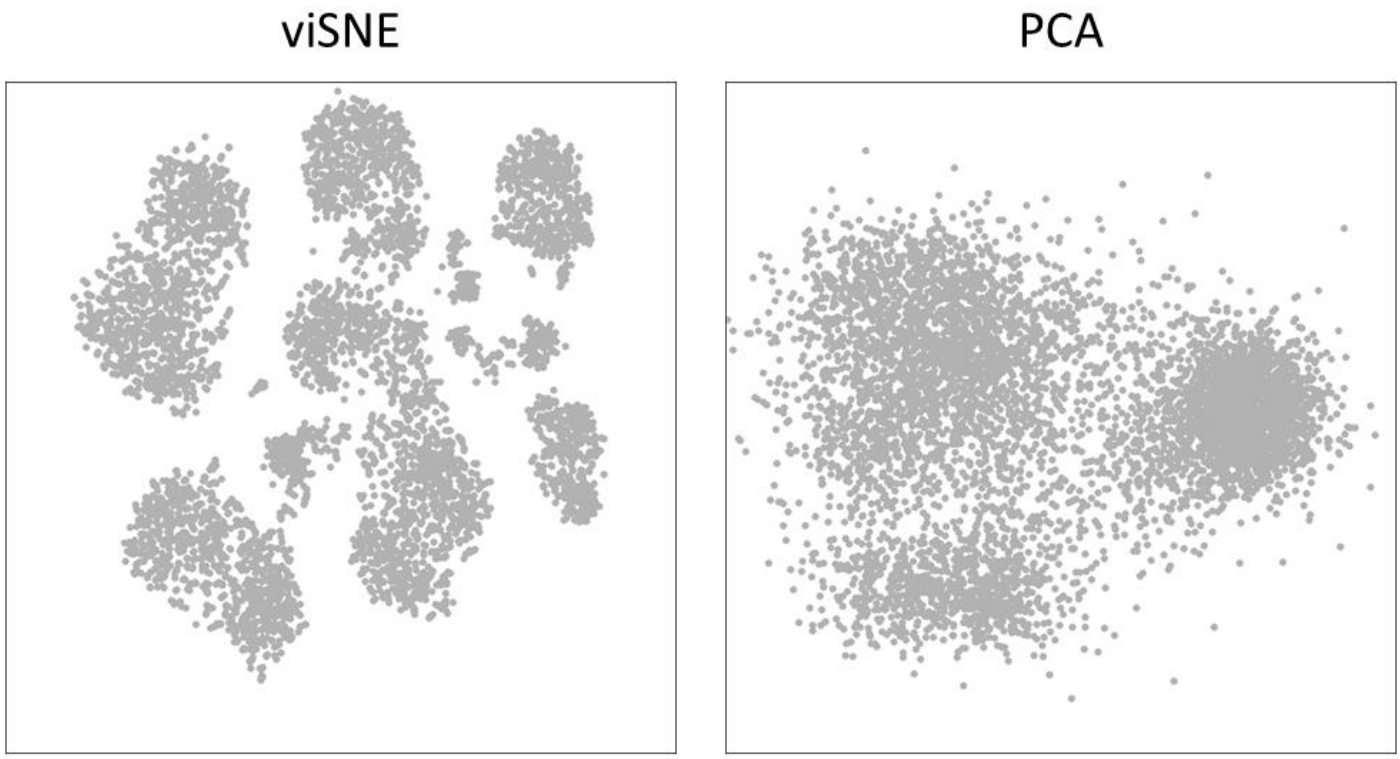


Supplementary Figure 16: Two SPADE runs of Marrow1, colored by mean marker expression levels for each cluster. In this plot each point is a cluster of cells and edges represent the minimal spanning tree identified by SPADE. Each row represents one of the two SPADE runs and each column representes the marker whose mean value was used to color the clusters. Clusters group by immune subtype, but longer range distances are less conserved between runs, resulting in considerable differences between the SPADE tree from the two runs.

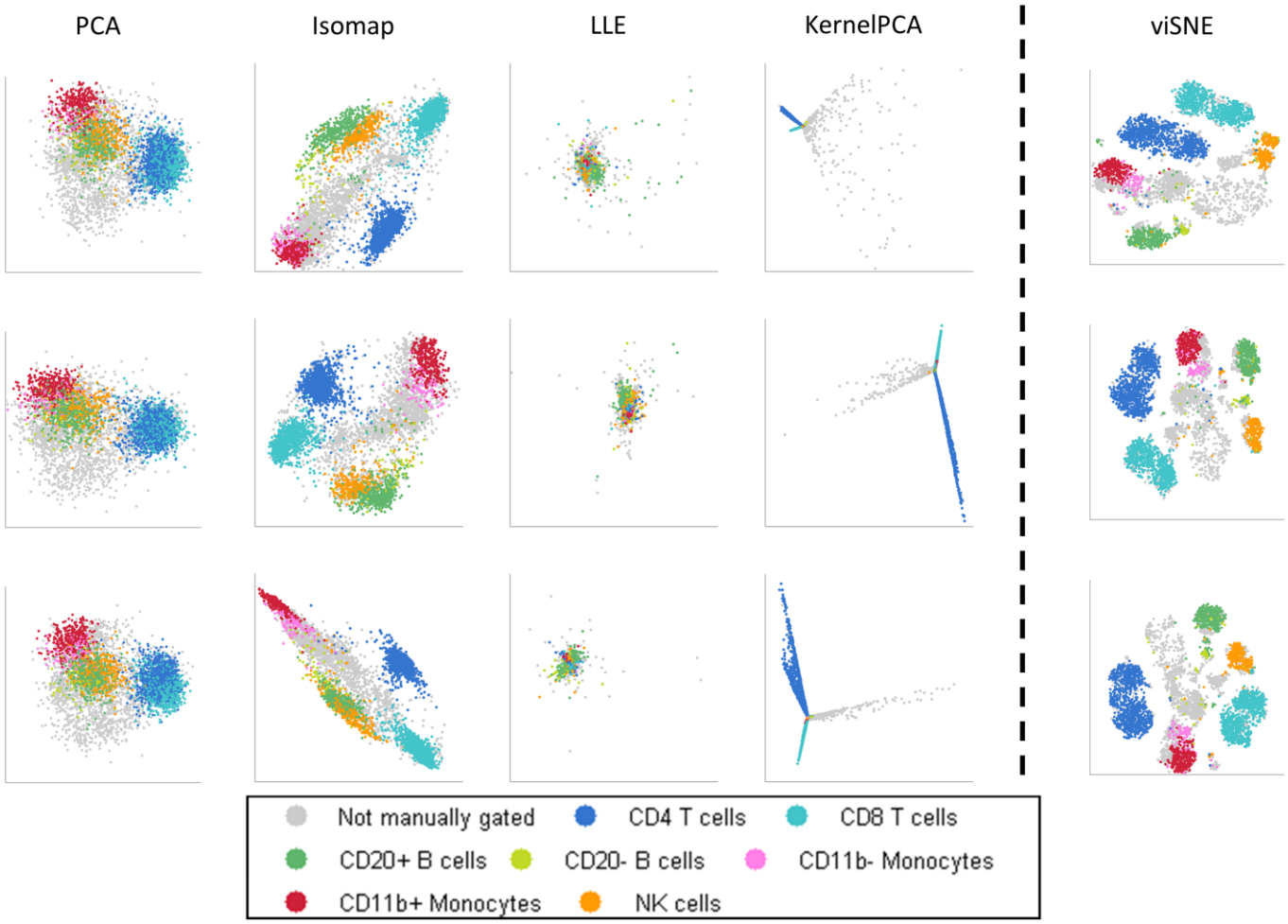


Supplementary Figure 17: SPADE was applied to the same synthetic MRD sample used in Figure 6. The ALL-barcoded cells (the target cells) are spread over 3 different clusters (marked with arrows); the cluster with the highest percentage of ALL-barcoded cells is marked with a black arrow- approximately 90% of its cells are from the MRD sample (the remaining 10% are from the healthy control sample). However, there are 73 clusters

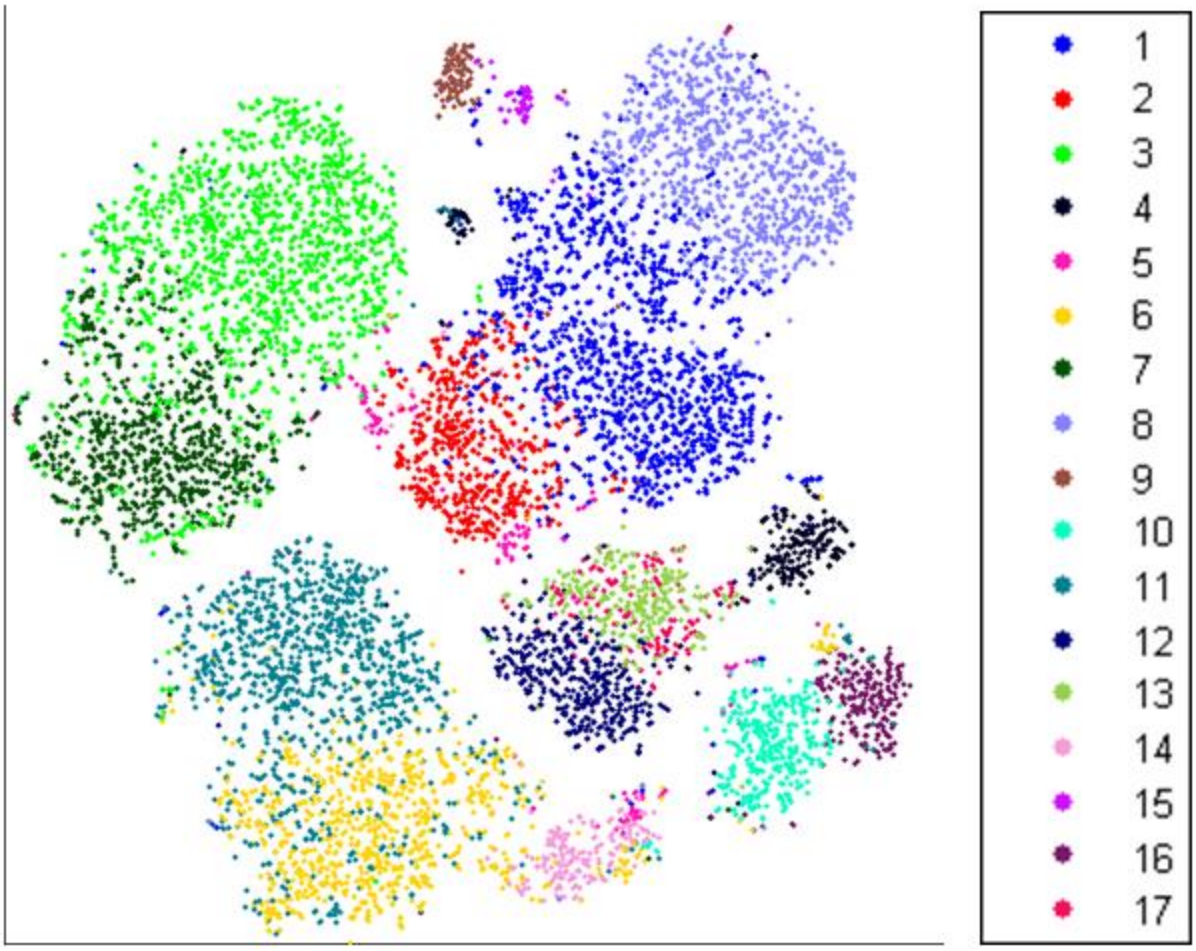
that fit the 90% MRD criterion (red clusters), and the 3 ALL-barcoded-rich clusters do not stand out relative to the other 72 high-MRD clusters or the SPADE tree as a whole.



Supplementary Figure 18: Comparison of PCA to viSNE. We applied viSNE (left) and PCA (right) to the same cells from Marrow1. PCA is limited to a linear projection and therefore the first two principal components (x and y axis respectively) fail to separate between the different immune subtypes. This demonstrates the non-linear nature of hematopoietic space and the importance of non-linear approaches to explore this space.



Supplementary Figure 19: Comparison of four dimensionality reduction algorithms (PCA, Isomap, LLE and Kernel PCA) [4-7] to viSNE over three subsamples of Marrow1. These are the same subsamples used in Supplementary Figure 2. Cells are color coded by immune subsets, as in Figure 1B. We found similar results for additional NLDR algorithms from the toolkit [8].



Supplementary Figure 20: FLOCK clustering of mass cytometry data [9], as visualized by viSNE, each cell is colored by its cluster id. FLOCK separates the major subtypes. However, it suffers from over-clustering and breaks most cell subtypes into multiple clusters. The viSNE maps helps regroup these back together for interpretation.

Supplementary Table 1: Antibody sources, metal isotope and staining concentration for all of the antibodies used throughout the various experiments.

Supplementary Table 2: Experiment details, per figure. Figure and section refers to the location of the figure in the text. In Dataset, Marrow stands for healthy bone marrow, ALL for acute lymphoid leukemia, AML for

acute myeloid leukemia and MRD for the *in vitro* MRD experiment. Cells subsampled is the number of cells subsampled from the whole dataset. Markers used is the list of surface markers measured in the dataset.

Supplementary Data 1: Marrow1 toy example. The FCS file includes the mature B cells and CD11b+ monocytes from Marrow1, their intensity over three channels (CD45, CD4 and CD8) and two- and three-dimensional viSNE maps. Three non-canonical markers can separate monocytes and myeloids. The FCS file can be opened and examined in *cyt*. Examination of the 3D plot (CD45 vs. CD4 vs. CD38) shows that the two immune subtypes are separated in 3D, but not in any 2D or 1D plot.

Supplementary Data 2-4: 3D viSNE maps of Marrow1, AML 1 and AML 2, respectively. The FCS files include all of the cells from these samples (measured over all channels) and two- and three- dimensional viSNE maps for each sample. Examining the map in 3D reveals further separation between immune subtypes. The structure captured by these maps is more intricate than in the 2D map.