Unsupervised Clustering and Population Mapping of the Levine *et al.* (2015) 32-Dimensional Mass-Cytometry Benchmark

A Technical Report on Marker-Driven Clustering, Manual-Gate Agreement, Frequency Recovery, and the Confounding of CD34$^+$ Stem/Progenitor Cells

K-Dense Web
contact@k-dense.ai

July 10, 2026

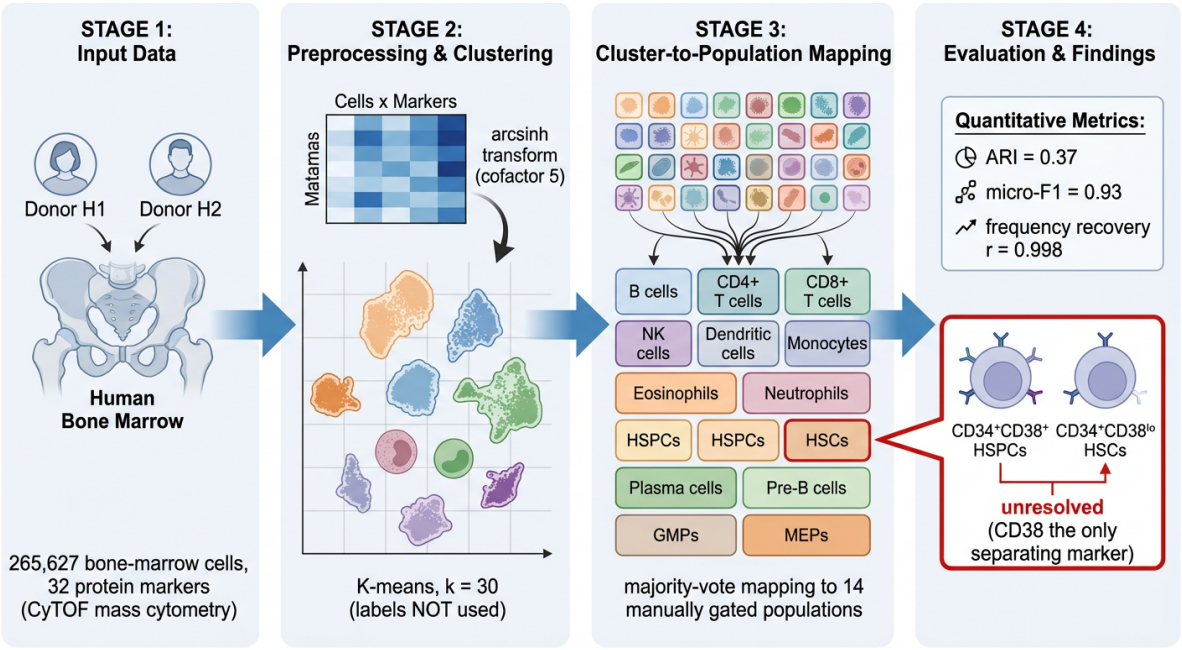

Figure 1: ***Graphical abstract.*** *The analysis workflow: 265,627 human bone-marrow mononuclear cells profiled across 32 protein markers by mass cytometry are arcsinh-transformed (cofactor 5), clustered with k-means ($k = 30$) without using the manual labels, and each cluster is mapped by majority vote to one of 14 manually gated populations. The mapping recovers most populations with high fidelity (micro-F1 = 0.93; frequency correlation $r = 0.998$) but cannot separate the two overlapping CD34$^+$ stem/progenitor populations, for which CD38 is essentially the only discriminating marker.*

Abstract

Manually gated mass-cytometry (CyTOF) benchmarks are the standard yardstick for evaluating unsupervised single-cell clustering. We analysed the Levine *et al.* (2015) 32-dimensional benchmark—human bone-marrow mononuclear cells (BMMCs) from two healthy donors distributed through the `HDCytoData` resource—comprising 265,627 cells, of which 104,184 carry manual gate labels spanning 14 immune/hematopoietic populations. We located the expression matrix and manual labels, applied the standard arcsinh transform (cofactor 5) to the 32 lineage markers, and clustered all cells with mini-batch k-means at $k = 30$ *without* using the manual labels as input. Each cluster was assigned to a population by majority vote over its labeled members. Against the manual gates, the partition achieved an adjusted Rand index (ARI) of 0.369 and a normalized mutual information (NMI) of 0.688; the mapped predictions

reached an overall accuracy / micro-F1 of 0.930 and a macro-F1 of 0.602. Recovered population frequencies tracked the manual frequencies almost perfectly (mean absolute error 0.41 percentage points; Pearson $r = 0.998$). The two populations the clustering most struggled to separate were the $CD34^{+}CD38^{+}CD123^{-}$ hematopoietic stem/progenitor cells (HSPCs) and the $CD34^{+}CD38^{lo}$ hematopoietic stem cells (HSCs): every HSC was absorbed into an HSPC-dominated cluster (HSC F1 = 0.0). Marker analysis shows the two populations share near-identical median expression across the entire lineage panel, with CD38 (and, secondarily, HLA-DR and CD45) providing the only appreciable separation—too little contrast for isotropic Euclidean clustering to resolve. We provide per-cell cluster assignments and complete code. The results reproduce the well-known finding that flat clustering excels on abundant, phenotypically distinct populations but fails on rare, continuous stem/progenitor compartments, motivating targeted sub-clustering or graph-based community detection.

Keywords: mass cytometry, CyTOF, unsupervised clustering, k-means, benchmark, hematopoietic stem cells, adjusted Rand index, F1 score

1 Introduction

1.1 Mass cytometry and high-dimensional single-cell phenotyping

Mass cytometry, or cytometry by time-of-flight (CyTOF), couples antibody staining to inductively coupled plasma time-of-flight mass spectrometry: antibodies are conjugated to purified transition-metal isotopes rather than fluorophores, and the per-cell metal abundances are read out as ion counts [1]. Because the metal reporters occupy narrow, well-separated mass channels, spectral overlap is minimal and 40 or more parameters can be measured simultaneously on each cell, far beyond the practical limit of conventional fluorescence flow cytometry [2]. The technique was first applied to human hematopoiesis by Bendall et al. [3], who mapped a continuous differentiation landscape in bone marrow and demonstrated that surface phenotype and intracellular signalling state can be profiled jointly at single-cell resolution. In the decade since, CyTOF has become a workhorse of systems immunology, oncology, and hematology, generating datasets in which every cell is a point in a 30–50-dimensional marker space [2].

The central analytical task for such data is *population identification*: partitioning millions of cells into biologically meaningful subsets. Historically this was done by *manual gating*—an expert sequentially drawing polygons on two-dimensional biaxial scatter plots to isolate known cell types. Manual gating is laborious, subjective, and scales poorly with dimensionality, motivating a large family of automated methods: dimensionality-reduction visualisations such as viSNE [4] and UMAP [5], tree-based summaries such as SPADE [6], self-organising-map clustering such as FlowSOM [7], and graph-based community detection such as PhenoGraph [8]. All of these are *unsupervised*: they recover structure from the marker expression alone, without access to any labels.

1.2 Why manually gated benchmarks matter

To know whether an unsupervised method recovers "real" biology, its output must be compared against a trusted reference. Manually gated CyTOF datasets provide exactly this: a subset of cells for which an expert has assigned a population label, allowing quantitative agreement to be measured. The most influential such benchmark is the 32-marker bone-marrow dataset published by Levine et al. [8] (their "benchmark data set 2"), in which healthy BMMCs from two donors were manually gated into 14 immune and hematopoietic populations. This dataset was central to the comparative study of Weber and Robinson [9], which evaluated many clustering tools against manual gates using the F1 score and Hungarian matching, and it is now distributed in analysis-ready form through the `HDCytoData` Bioconductor package [10]. Because

the dataset spans abundant lymphoid populations (CD4/CD8 T cells, B-cell stages, NK cells, monocytes) alongside rare $CD34^+$ stem and progenitor compartments, it stresses a clustering method across the full difficulty spectrum—from cleanly separated cell types to overlapping, continuous developmental states [11].

1.3 Objectives of this report

This report documents an end-to-end unsupervised clustering analysis of the Levine_32dim benchmark and its evaluation against the manual gates. Following the task specification, we (i) locate the expression matrix and manual labels; (ii) cluster the cells from marker expression alone, *without* using the manual labels as input; (iii) map clusters to populations; and (iv) report the clustering method and number of clusters, quantitative agreement with the manual gates (adjusted Rand index and a matched cluster-to-population F1), the recovered frequency of each population, and—with supporting marker evidence—the two manually gated populations the clustering most struggles to separate. Per-cell cluster assignments and all code accompany the report. Our specific contributions are:

- A reproducible pipeline (arcsinh transform, z-scoring, mini-batch k-means at $k = 30$, majority-vote mapping) with fixed random seeds and version pinning.
- A complete quantitative evaluation: ARI = 0.369, NMI = 0.688, micro-F1 = 0.930, macro-F1 = 0.602, and per-population precision/recall/F1.
- A near-perfect frequency-recovery analysis (mean absolute error 0.41 percentage points; $r = 0.998$).
- A marker-level diagnosis of the single most confounded population pair—$CD34^+CD38^+CD123^-$ HSPCs versus $CD34^+CD38^{lo}$ HSCs—and a discussion of methodological remedies.

2 Materials and Methods

2.1 Dataset and provenance

We used the *Levine_32dim* benchmark from Levine et al. [8], obtained via the source referenced by `HDCytoData` [10]. The archive (37 683 044.0000 bytes; SHA-256 `ff683a...eda57`) contains 30 FCS files—15 per donor (H1 and H2), corresponding to 14 manually gated populations plus one file (`NotDebrisSinglets`) holding the ungated ("unassigned") cells. Files were read with `flowio`; benign off-by-one DATA-offset flags were forced and the reconstructed cell counts validated exactly against the published benchmark. The combined single-cell table contains **265,627** cells (H1: 191,351; H2: 74,276). Of these, **104,184** carry a manual population label (the "labeled subset"; H1: 72,463; H2: 31,721) and 161,443 are unassigned. Each cell has 39 FCS channels, of which **32** are "type" (lineage) markers used for clustering: CD45RA, CD133, CD19, CD22, CD11b, CD4, CD8, CD34, Flt3, CD20, CXCR4, CD235ab, CD45, CD123, CD321, CD14, CD33, CD47, CD11c, CD7, CD15, CD16, CD44, CD38, CD13, CD3, CD61, CD117, CD49d, HLA-DR, CD64, and CD41. The 14 populations and their overall cell counts are listed in Table 1. There were no missing values in the marker columns.

Table 1: *The 14 manually gated populations of the Levine_32dim benchmark and their labeled-cell counts and frequencies (as a fraction of the 104,184 labeled cells). Populations span abundant lymphoid/myeloid types and rare CD34$^+$ stem/progenitor compartments.*

Population	Cells	Freq. (%)	Compartment
CD4 T cells	26,366	25.31	Lymphoid (T)
Monocytes	21,099	20.25	Myeloid
CD8 T cells	20,108	19.30	Lymphoid (T)
Mature B cells	16,520	15.86	Lymphoid (B)
Pre B cells	6,135	5.89	Lymphoid (B)
CD16$^-$ NK cells	3,905	3.75	Lymphoid (NK)
CD34$^+$CD38$^+$CD123$^-$ HSPCs	3,295	3.16	Stem/progenitor
CD16$^+$ NK cells	2,248	2.16	Lymphoid (NK)
pDCs	1,238	1.19	Dendritic
Basophils	1,207	1.16	Myeloid
CD34$^+$CD38^{lo} HSCs	916	0.88	Stem/progenitor
Pro B cells	513	0.49	Lymphoid (B)
Plasma B cells	330	0.32	Lymphoid (B)
CD34$^+$CD38$^+$CD123$^+$ HSPCs	304	0.29	Stem/progenitor
Total labeled	**104,184**	**100.0**	—

2.2 Preprocessing: arcsinh transform (cofactor 5)

Raw CyTOF ion counts are heavily right-skewed and heteroscedastic, spanning several orders of magnitude with a dense mass of near-zero (baseline-subtracted) values. The community-standard variance-stabilising transform is the inverse hyperbolic sine,

$$y \;=\; \operatorname{arcsinh}\!\left(\frac{x}{c}\right), \qquad c = 5, \tag{1}$$

applied independently to each of the 32 type markers. The cofactor $c = 5$ is the conventional choice for mass cytometry: it behaves approximately linearly for small $|x|$ (so negative, baseline-subtracted counts are handled gracefully without clipping) and logarithmically for large x, compressing the dynamic range while preserving the negative/positive population structure. Because arcsinh is defined on the whole real line, the ∼27% of marker entries that are negative required no special handling; we verified zero non-finite values after transformation. Diagnostic before/after histograms (Fig. 2) confirm the intended effect: the raw signal is converted into approximately bimodal distributions with clear negative and positive modes. A sensitivity check overlaying $c = 5$ against $c = 10$ (Fig. 3) showed cofactor 5 giving slightly wider positive-peak separation, and it was retained as the standard.

2.3 Unsupervised clustering

The 32 arcsinh-transformed markers were z-scored with a `StandardScaler` fit on the full 265,627-cell dataset, so that high-variance markers do not dominate the Euclidean metric. We then clustered all cells with **mini-batch k-means** [12, 13], a scalable stochastic approximation to Lloyd's algorithm, using `scikit-learn` [14]. We set the number of clusters to $k = 30$. This is a deliberate *over-clustering* choice: it lies within the standard CyTOF range of 20–40 and provides more clusters than the 14 target populations, so that rare or heterogeneous populations have a chance of receiving their own cluster before any cluster-to-population mapping is imposed. Hyperparameters were `batch_size` = 10,000, `n_init` = 10, `max_no_improvement` = 20, and `random_state` = 42. Crucially, **the manual labels were not used at any point during**

scaling or clustering; they enter only in the post-hoc mapping and evaluation. Clustering all cells took ∼6 s.

2.4 Cluster-to-population mapping

The full-dataset cluster assignments were joined by `Cell_ID` to the 104,184 labeled cells. Each cluster was then assigned to the manual population holding the *majority* of its labeled members (majority-vote mapping):

$$\mathrm{map}(k) \;=\; \arg\max_{p} \big|\{\, i : c_i = k \;\wedge\; y_i = p \,\}\big|, \tag{2}$$

where c_i is the cluster of cell i and y_i its manual label. Of the 30 clusters, 29 contained at least one labeled cell and were mapped; one cluster (cluster 8) contained only unassigned cells in the labeled join and was therefore unmapped. Per-cluster *purity* (the majority fraction) was recorded for every cluster. Majority-vote mapping is a many-to-one assignment: multiple clusters may map to the same population (expected under over-clustering), but each labeled cell receives exactly one predicted population, enabling a standard multi-class evaluation.

2.5 Evaluation metrics

Two complementary families of metrics were computed on the labeled subset.

Partition-level agreement (mapping-invariant) quantifies how well the raw cluster partition matches the manual-gate partition, irrespective of labels:

- *Adjusted Rand Index* (ARI) [15]: the Rand index of pairwise co-assignment agreement, corrected for chance so that a random partition scores ≈ 0 and a perfect match scores 1. ARI penalises the intentional splitting of one population across several clusters, so an over-clustered solution is expected to score well below its classification accuracy.
- *Normalized Mutual Information* (NMI): the mutual information between cluster and label assignments, normalised to $[0, 1]$; unlike ARI it is not penalised by pure sub-splitting of a population.

Matched classification metrics treat the mapped population as a prediction and the manual gate as ground truth. We report per-population precision, recall, and F1, together with macro- (unweighted mean over the 14 populations), micro- (equivalently, overall accuracy), and support-weighted averages. The matching underlying these scores is the majority-vote cluster-to-population map defined above (a many-to-one assignment), *not* a one-to-one Hungarian matching; this is stated explicitly because the choice of matching materially affects the interpretation of the F1 scores (Section 3.3).

Frequency recovery compares each population's manual frequency (fraction of labeled cells) with its recovered frequency (fraction of labeled cells mapped to it), reporting per-population absolute error, the mean/max absolute error, and the Pearson correlation between the two frequency vectors.

Resolution analysis. A row-normalised confusion matrix was computed over the 14 populations. For every unordered population pair (A, B) we defined a *mutual-confusion* score $= \mathrm{frac}(A \to B) + \mathrm{frac}(B \to A)$, i.e. the sum of the fractions of each population's cells predicted as the other, and ranked all pairs. For the top pair we extracted per-marker median arcsinh expression and ranked markers by the absolute difference of medians to identify the most similar (confounding) and most discriminating markers.

2.6 Dimensionality-reduction visualisation

For qualitative inspection only, UMAP [5] was fit on a 20,000-cell random subsample of the labeled cells (`n_neighbors` = 15, `min_dist` = 0.1, Euclidean metric, seed 42) and coloured by manual gate and by cluster (Fig. 4). The UMAP embedding was not used for clustering.

2.7 Reproducibility

All steps use fixed seeds (`random_state` = 42). Software: Python 3.12.10; NumPy 2.4.6; pandas 3.0.3; scikit-learn 1.9.0; umap-learn 0.5.12; flowio 1.4.0; pyarrow 25.0.0. Per-cell assignments (`per_cell_cluster_assignments.csv`, 265,627 rows; and the labeled subset, 104,184 rows) and the analysis scripts (`preprocess_data.py`, `cluster_cells.py`, `evaluate_mapping.py`) are provided with this report.

3 Results

3.1 Preprocessing recovers bimodal marker distributions

The arcsinh transform (cofactor 5) converted the raw, right-skewed ion-count distributions into approximately bimodal distributions with clearly resolved negative and positive modes (Fig. 2), the expected variance-stabilising behaviour that makes Euclidean distances between cells meaningful. Lineage markers such as CD45, CD3, CD19, and CD11b show the crisp "off/on" structure that manual gating relies on. The cofactor sensitivity check (Fig. 3) confirmed that $c = 5$ gives marginally wider positive-peak separation than $c = 10$, supporting the standard choice.

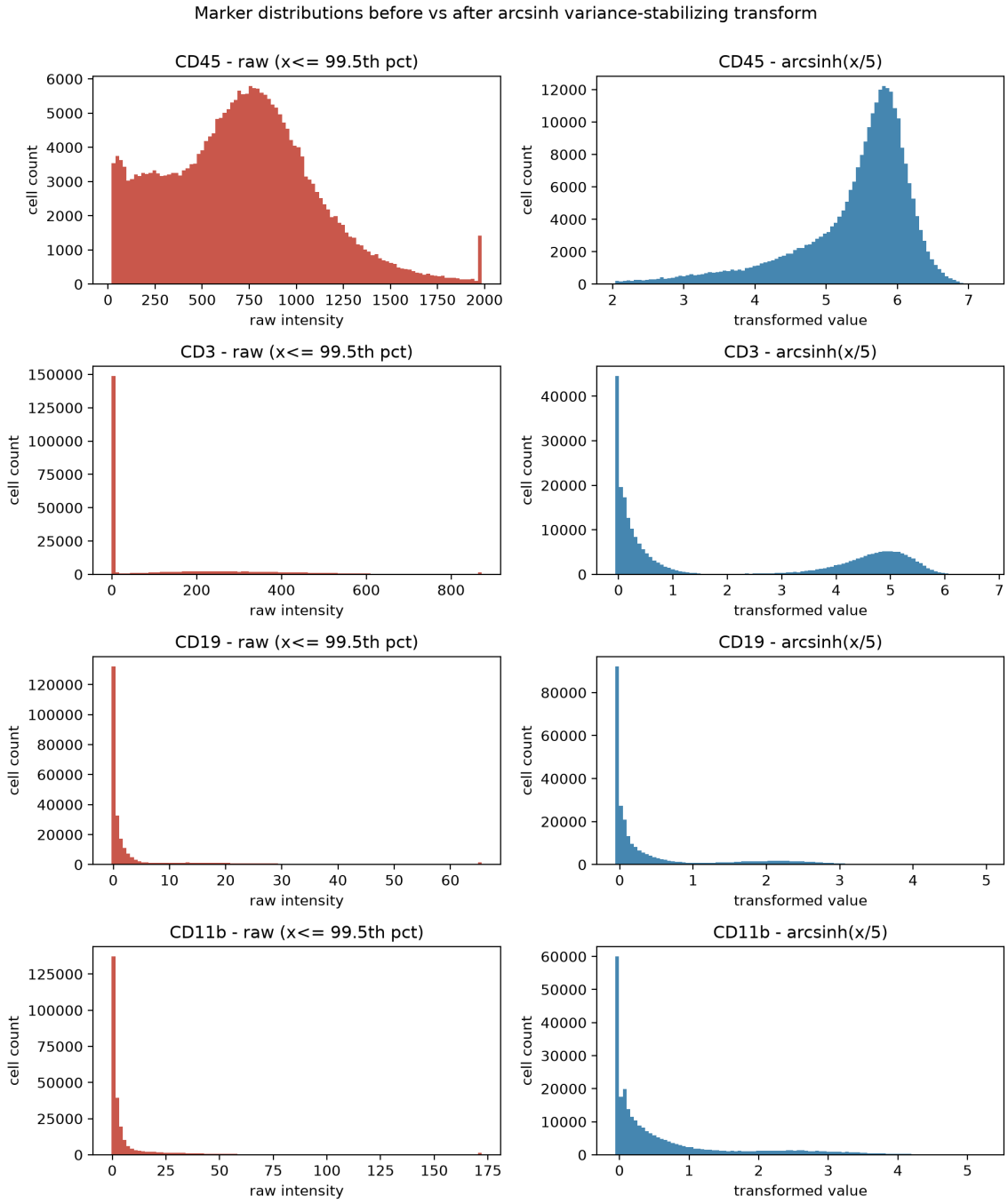

Figure 2: ***Effect of the arcsinh (cofactor 5) transform.*** *Before (raw ion counts) and after (arcsinh) distributions for four representative markers (CD45, CD3, CD19, CD11b). The transform compresses the extreme right tail and resolves the signal into approximately bimodal negative/positive modes, stabilising variance for Euclidean clustering.*

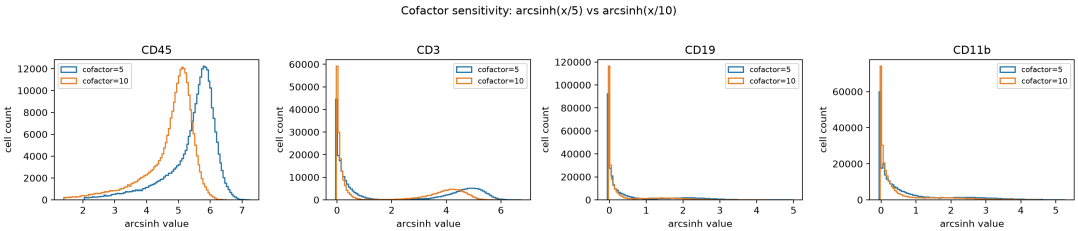

Figure 3: ***Cofactor sensitivity.*** *Overlaid transformed distributions for cofactor $c = 5$ (CyTOF standard) versus $c = 10$. Cofactor 5 yields slightly wider separation of the positive peak and was retained.*

3.2 Clustering method and number of clusters

The reported clustering is **mini-batch k-means with $k = 30$** on the 32 z-scored, arcsinh-transformed markers, run on all 265,627 cells without the manual labels. The 30 clusters ranged in size from 912 to 25,843 cells (full dataset). On the labeled subset, 29 of the 30 clusters contained labeled cells and were mapped to populations; cluster 8 held only unassigned cells and was left unmapped. The UMAP embedding (Fig. 4) shows visually that the unsupervised clusters and the manual gates carve out the same island structure: abundant lymphoid and myeloid populations form well-separated territories that the clusters recover cleanly, whereas the small CD34$^+$ progenitor island in the centre is fragmented.

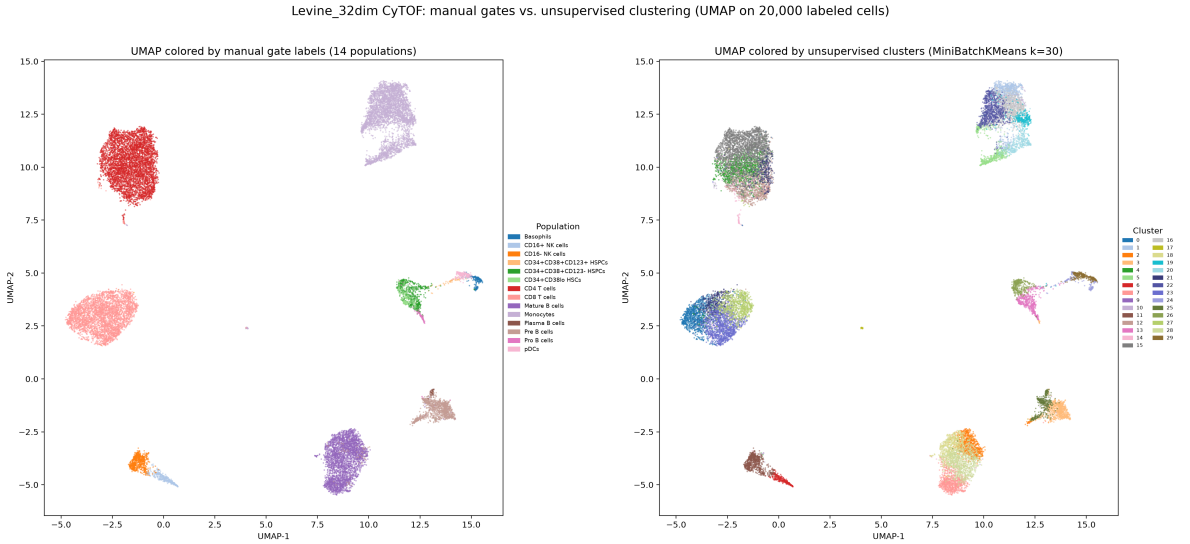

Figure 4: ***UMAP of 20,000 labeled cells, coloured by manual gate (left) versus unsupervised cluster (right).*** *The two panels share the same island topology, confirming that mini-batch k-means ($k = 30$) recovers the manually gated populations. Over-clustering splits several large populations (e.g. CD4/CD8 T cells, monocytes) into multiple same-population clusters, and the small central CD34$^+$ stem/progenitor island is not cleanly separated into its constituent gates.*

3.3 Clustering agreement against the manual gates

Partition-level agreement between the 30-cluster solution and the 14 manual gates was **ARI** $= 0.369$ and **NMI** $= 0.688$ (Table 2). These values reproduce exactly the post-hoc sanity check computed during clustering, confirming consistency. The moderate ARI is expected and not a sign of poor clustering: ARI penalises the deliberate splitting of a single population across multiple clusters (over-clustering), so an intentionally over-partitioned solution is structurally

capped well below 1 even when each cluster is highly pure. The higher NMI (0.688) reflects that the clusters are individually informative about the labels.

Under the majority-vote cluster-to-population matching, the mapped predictions achieved an **overall accuracy / micro-F1 of 0.930** and a **macro-F1 of 0.602** (Table 2). The large gap between micro- and macro-F1 is the key structural result: micro-F1 is dominated by the abundant populations, which are recovered almost perfectly, whereas macro-F1 weights all 14 populations equally and is dragged down by three rare populations that receive F1 = 0 (see below). Per-cluster purity was high for most clusters—the majority of clusters exceeded 90% purity (e.g. clusters mapping to monocytes, mature B cells, and NK cells reached 99–100%)—while the clusters covering the $CD34^+$ compartment, basophils, and pDCs were markedly less pure (48–77%).

Table 2: *Overall agreement metrics on the 104,184 labeled cells. ARI and NMI are partition-level (mapping-invariant); accuracy and the F1 averages use the majority-vote cluster-to-population matching. Micro-F1 equals overall accuracy for a single-label multi-class problem.*

Metric	Value	Metric	Value
Adjusted Rand Index (ARI)	0.369	Macro precision	0.596
Normalized Mutual Info (NMI)	0.688	Macro recall	0.617
Overall accuracy / micro-F1	0.930	Macro F1	0.602
Micro precision / recall	0.930	Weighted F1	0.921

Per-population precision, recall, and F1 are given in Table 3. Ten of the 14 populations were recovered with F1 $\geq$ 0.40, and seven of the most abundant with F1 $\geq$ 0.90 (monocytes 0.987; mature B cells 0.974; CD4 T cells 0.950; CD8 T cells 0.936; $CD16^-$ NK cells 0.925; $CD16^+$ NK cells 0.910; pre B cells 0.901). Three rare populations received F1 = 0: the two smallest $CD34^+$ compartments ($CD34^+CD38^+CD123^+$ HSPCs, n = 304; $CD34^+CD38^{lo}$ HSCs, n = 916) and plasma B cells (n = 330), together with pro B cells (n = 513). In every case, no cluster's *majority* belonged to these populations, so the majority-vote map never predicts them, forcing precision and recall—and hence F1—to zero. This is a direct and expected consequence of a many-to-one majority-vote matching applied to rare populations that are subsumed by larger, phenotypically adjacent neighbours; it is not evidence that the cells are unclustered, but that they co-cluster with a more abundant relative.

Table 3: *Per-population precision, recall, and F1 under majority-vote mapping, ordered by support. F1 = 0 arises when no cluster's majority is that population (the map never predicts it). Support is the number of manually gated cells.*

Population	Precision	Recall	F1	Support
CD4 T cells	0.954	0.946	0.950	26,366
Monocytes	0.976	0.999	0.987	21,099
CD8 T cells	0.931	0.941	0.936	20,108
Mature B cells	0.966	0.982	0.974	16,520
Pre B cells	0.901	0.901	0.901	6,135
$CD16^-$ NK cells	0.906	0.944	0.925	3,905
$CD34^+CD38^+CD123^-$ HSPCs	0.704	0.974	0.817	3,295
$CD16^+$ NK cells	0.997	0.836	0.910	2,248
pDCs	0.531	0.770	0.629	1,238
Basophils	0.480	0.347	0.403	1,207
$CD34^+CD38^{lo}$ HSCs	0.000	0.000	0.000	916
Pro B cells	0.000	0.000	0.000	513
Plasma B cells	0.000	0.000	0.000	330
$CD34^+CD38^+CD123^+$ HSPCs	0.000	0.000	0.000	304

3.4 Recovered population frequencies

Despite the resolution failures on rare populations, the *recovered frequencies* of the major populations were remarkably accurate (Fig. 5, Table 4). Across all 14 populations the mean absolute frequency error was **0.41 percentage points**, the maximum absolute error was 1.21 percentage points (for the $CD34^{+}CD38^{+}CD123^{-}$ HSPCs, which absorb the HSCs), and the Pearson correlation between manual and recovered frequency vectors was $r = 0.998$. For the abundant populations the recovery is essentially exact: CD4 T cells 25.31% vs. 25.11%, monocytes 20.25% vs. 20.72%, CD8 T cells 19.30% vs. 19.51%, mature B cells 15.86% vs. 16.13%, and pre B cells 5.889% vs. 5.890%. The populations with zero recovered frequency are precisely those with F1 = 0: their cells are redistributed into a neighbouring population's frequency rather than lost, which is why the errors remain small in absolute terms (each of these populations is $< 1\%$ of cells).

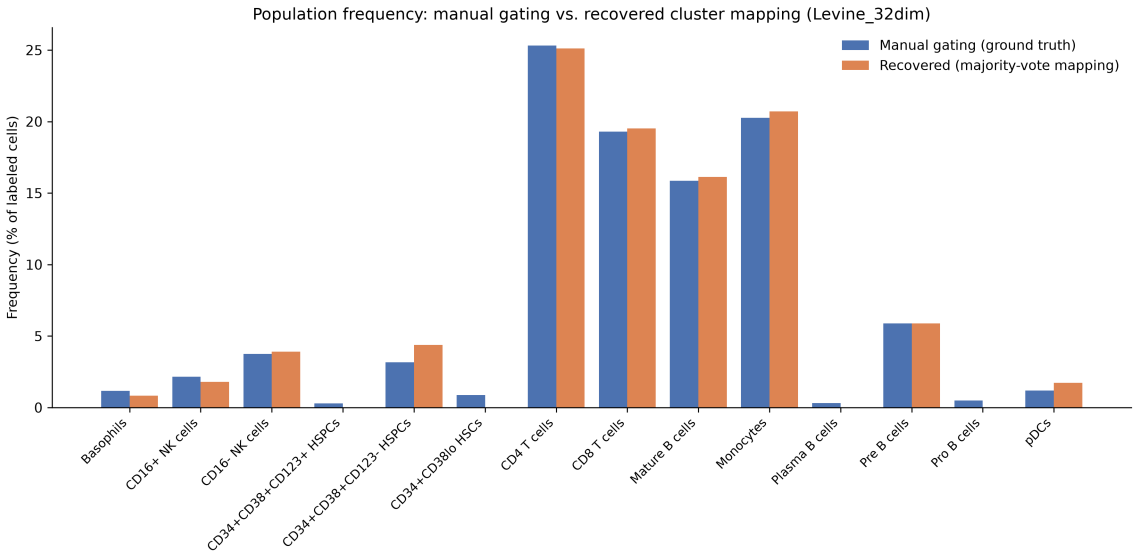

Figure 5: ***Population frequency: manual gating vs. recovered cluster mapping.*** *Manual (blue) and recovered (orange) frequencies as a percentage of the 104,184 labeled cells. Abundant populations are recovered almost exactly; the $CD34^{+}CD38^{lo}$ HSCs, $CD34^{+}CD38^{+}CD123^{+}$ HSPCs, plasma B, and pro B populations recover to ≈ 0 because their cells are absorbed into neighbouring populations, inflating those neighbours slightly.*

Table 4: *Manual versus recovered population frequencies (percentage of labeled cells) and absolute error. Summary: mean absolute error = 0.41 pp, maximum = 1.21 pp, Pearson r = 0.998.*

Population	Manual (%)	Recovered (%)	Abs. error (pp)
CD4 T cells	25.31	25.11	0.20
Monocytes	20.25	20.72	0.47
CD8 T cells	19.30	19.51	0.21
Mature B cells	15.86	16.13	0.27
Pre B cells	5.889	5.890	0.001
$CD16^-$ NK cells	3.75	3.90	0.16
$CD34^+CD38^+CD123^-$ HSPCs	3.16	4.38	1.21
$CD16^+$ NK cells	2.16	1.81	0.35
pDCs	1.19	1.72	0.53
Basophils	1.16	0.84	0.32
$CD34^+CD38^{lo}$ HSCs	0.88	0.00	0.88
Pro B cells	0.49	0.00	0.49
Plasma B cells	0.32	0.00	0.32
$CD34^+CD38^+CD123^+$ HSPCs	0.29	0.00	0.29

3.5 The two populations hardest to separate

Ranking all population pairs by mutual-confusion score identified a clear worst case (Table 5): the **$CD34^+CD38^+CD123^-$ HSPCs** and the **$CD34^+CD38^{lo}$ HSCs**, with a mutual-confusion score of **0.996**. The confusion is almost entirely one-directional: 99.6% of manually gated HSCs are routed into clusters whose majority is HSPCs, while essentially no HSPCs are routed to HSCs. As a result the HSPCs are recovered with F1 = 0.817 (they dominate their clusters and even absorb the HSCs), but the HSCs collapse to F1 = 0.0—no cluster is ever majority-HSC. Concretely, of 916 gated HSCs, 912 land in HSPC-mapped clusters; the two HSPC-dominated clusters (13 and 26) contain 511 and 401 HSCs respectively alongside their HSPC majorities.

Table 5: *Top five most-confused population pairs by mutual-confusion score* = $\mathrm{frac}(A \to B) + \mathrm{frac}(B \to A)$. *The $CD34^+$ stem/progenitor pair is by far the hardest to separate.*

Population A	Population B	$A \to B$	$B \to A$	Mutual
$CD34^+CD38^+CD123^-$ HSPCs	$CD34^+CD38^{lo}$ HSCs	0.000	0.996	0.996
Plasma B cells	Pre B cells	0.976	0.000	0.976
Basophils	pDCs	0.640	0.227	0.867
$CD34^+CD38^+CD123^-$ HSPCs	Pro B cells	0.000	0.696	0.696
Basophils	$CD34^+CD38^+CD123^+$ HSPCs	0.000	0.513	0.513

Marker evidence. The two populations are near-identical across the lineage panel (Fig. 6). Ranking the 32 markers by the absolute difference of their median arcsinh expression between HSPCs and HSCs (Table 6), the eight *most similar* markers—CD19, CD22, CD14, CD64, CD20, CD11b, CD4, CD7—differ by < 0.01 arcsinh units, i.e. both populations are uniformly negative for the entire B-cell, monocyte, and T/NK lineage panel, exactly as expected for two lineage-negative $CD34^+$ progenitor compartments. The differences that do exist are concentrated in a handful of markers: **CD38** (median 3.41 in HSPCs vs. 1.20 in HSCs; $\Delta = 2.21$)—by construction the defining axis of the two gates—followed distantly by HLA-DR ($\Delta = 1.00$), CD45 ($\Delta = 0.60$), CD34 ($\Delta = 0.42$, both strongly positive), CD47, CD321, and CD13. In other words, of 32 markers only *one* carries substantial discriminating signal (CD38), and after z-scoring that single dimension is far too weak against the 31 near-identical dimensions for isotropic k-means

to place the two populations in separate clusters. The clustering therefore merges them, and majority vote assigns the merged clusters to the more abundant HSPCs.

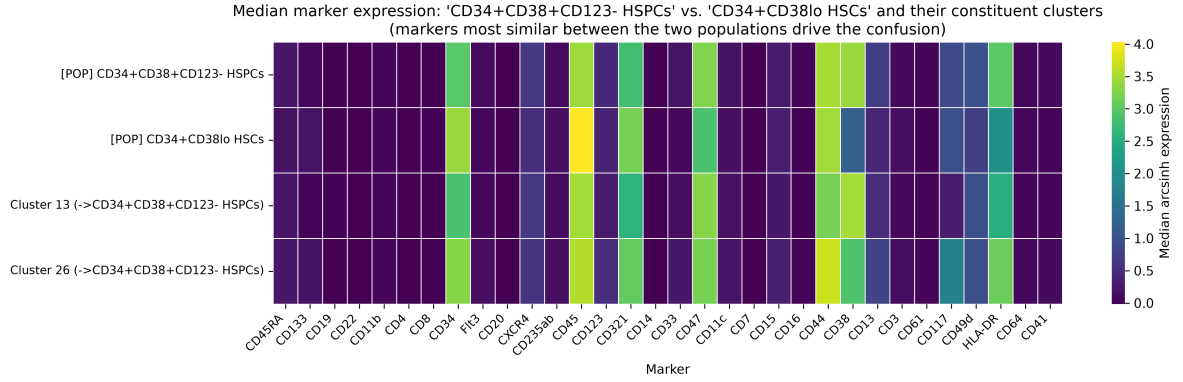

Figure 6: ***Median marker expression of the two hardest-to-separate populations and their constituent clusters.*** *Rows: the two populations ($CD34^+CD38^+CD123^-$ HSPCs and $CD34^+CD38^{lo}$ HSCs) and the two clusters (13, 26) that capture the $CD34^+$ compartment. Columns: the 32 markers (median arcsinh expression). The two population rows are visually near-identical except at CD38 (and, more subtly, HLA-DR and CD45); every lineage marker is uniformly negative, leaving k-means almost no Euclidean contrast to separate the populations.*

Table 6: *Markers ranked by $|\Delta\text{median}|$ (absolute difference of median arcsinh expression) between the two struggling populations. Only CD38 shows large separation; the top-ranked "most similar" markers are all uniformly negative lineage markers. (Selected rows; full 32-marker table in* `evaluation_results.json`*.)*

Marker	HSPC median	HSC median	$\lvert\Delta\rvert$
Most discriminating			
CD38	3.409	1.197	2.212
HLA-DR	3.005	2.000	1.005
CD45	3.430	4.030	0.601
CD34	2.976	3.399	0.423
CD47	3.263	2.865	0.398
CD321	2.799	3.192	0.393
CD13	0.699	0.404	0.295
Most similar (confounding)			
CD7	−0.006	−0.014	0.008
CD4	−0.010	−0.003	0.006
CD11b	0.048	0.054	0.006
CD20	−0.011	−0.007	0.004
CD64	0.071	0.068	0.003
CD14	−0.011	−0.013	0.002
CD22	−0.009	−0.008	0.0004
CD19	−0.005	−0.005	0.0003

4 Discussion

4.1 Interpretation: flat clustering succeeds on distinct types, fails on continua

The overall picture is that mini-batch k-means at $k = 30$ recovers the Levine_32dim populations very well where the biology provides discrete, well-separated phenotypes, and fails precisely

where the biology is continuous. The abundant lymphoid and myeloid populations—T cells, B-cell stages, NK cells, monocytes—occupy distinct regions of marker space defined by strong, near-binary lineage markers (CD3, CD19/CD20, CD16, CD14), and the clusters recover them with F1 ≥ 0.90 and near-exact frequencies ($r = 0.998$). The moderate ARI (0.369) is an artefact of intentional over-clustering rather than a defect: pure clusters that split one population into several pieces are penalised heavily by ARI even though they are biologically correct, which is why the micro-F1 (0.930) is a fairer summary of population recovery here.

The failures cluster around *developmental continua*. The single hardest pair—$CD34^+CD38^+CD123^-$ HSPCs versus $CD34^+CD38^{lo}$ HSCs—is a textbook example. Human hematopoietic stem and progenitor cells have long been enriched by the $CD34^+CD38^{-/lo}$ phenotype [16, 17], but CD38 is a *continuous* marker that rises gradually as HSCs differentiate into downstream progenitors; the "HSC" and "HSPC" gates therefore sit on adjacent segments of a single continuum separated by an operator-chosen threshold on essentially one axis [11, 18]. Our marker analysis makes this quantitative: of 32 markers, only CD38 shows large median separation between the two gates ($\Delta = 2.21$ arcsinh units), with everything else—including the entire lineage-negative panel—nearly identical (Table 6). In the z-scored 32-dimensional space that mini-batch k-means optimises, a single discriminating dimension embedded among 31 near-identical dimensions contributes negligibly to the Euclidean objective, so the algorithm has no incentive to split the compartment along CD38. The rarer population (HSCs, 0.88% of labeled cells) is consequently swallowed by the more abundant one (HSPCs, 3.16%), yielding HSC F1 = 0 and the extreme one-directional confusion we observe. The same mechanism explains the other zero-F1 populations: plasma B cells are absorbed by pre B cells (they lie on the B-maturation continuum), and pro B cells fall into the $CD34^+$ progenitor clusters.

Importantly, this is a limitation of *flat, isotropic, distance-based clustering*, not of the data or of CyTOF: the discriminating information (CD38, HLA-DR, CD45) is present and the manual gater used it. It is simply drowned out when all 32 markers are weighted equally in a global Euclidean metric.

4.2 Methodological improvements

Several well-established strategies would be expected to recover the confounded compartments.

Targeted sub-clustering (metaclustering). Rather than asking one global clustering to resolve every scale of structure at once, one can first isolate the $CD34^+$ compartment (e.g. the union of clusters 13 and 26) and re-cluster it at higher resolution using only the markers that define its internal structure (CD38, CD45RA, CD90, CD117, HLA-DR). Freed from the 31 uninformative dimensions, even k-means can split the compartment along CD38. This hierarchical "cluster-then-refine" approach is the standard remedy for rare, nested populations and underlies the metaclustering step of FlowSOM [7].

Graph-based community detection. Methods that build a k-nearest-neighbour graph and optimise modularity—Louvain [19] and its well-connected successor Leiden [20], as used in PhenoGraph [8]—adapt their effective resolution to local density and do not assume convex, isotropic clusters. With an appropriate resolution parameter they can carve out small, dense progenitor communities that k-means merges. FlowSOM's self-organising map likewise preserves local topology and, in the benchmark of Weber and Robinson [9], combined strong accuracy with rare-population sensitivity and very fast runtimes; X-shift was found especially effective for rare populations.

Supervised marker weighting or metric learning. Because the confounding is a metric problem, weighting markers by their discriminative value (up-weighting CD38/HLA-DR/CD45 within the $CD34^+$ compartment) or learning a Mahalanobis metric would restore separation

without changing the algorithm family.

Differential-abundance frameworks. Where the goal is downstream comparison rather than exact population labelling, high-resolution clustering followed by moderated testing—as in `diffcyt` [21]—is designed precisely to retain and test rare populations that flat clustering would merge.

Additive markers. Finally, the biology itself argues for richer panels: modern work resolves the primitive human HSC compartment with CD90, CD45RA, CD49f, and EPCR/CD201 rather than CD38 alone [18, 22]. No unsupervised method can separate what the panel does not measure; the Levine_32dim panel intentionally provides only CD38 (plus CD34, CD45RA, CD117, Flt3, CD133) for the stem compartment, which caps the achievable resolution regardless of algorithm.

4.3 Limitations

This analysis used a single clustering family (k-means) at a single resolution ($k = 30$) with one random seed; while seeds were fixed for reproducibility, we did not sweep k or compare algorithms, so the numbers characterise this configuration rather than the best achievable. The majority-vote mapping is a many-to-one assignment and therefore cannot, by construction, predict a population that never dominates a cluster; a one-to-one (Hungarian) matching would assign *some* cluster to each of the 14 populations and would report non-zero (though low) F1 for the rare compartments, at the cost of degrading a dominant population's score—an alternative accounting of the same underlying merger. Evaluation was performed on the labeled subset only (as is standard for this benchmark), and the 161,443 unassigned cells were clustered but not scored. Finally, agreement is measured against manual gates, which are themselves an imperfect, operator-dependent reference—particularly for the CD38 threshold that separates the very populations we identify as hardest.

4.4 Conclusion

Unsupervised mini-batch k-means at $k = 30$, applied to the arcsinh-transformed 32-marker Levine_32dim benchmark without any label information, recovers the manually gated bone-marrow populations with high fidelity for all abundant, phenotypically distinct cell types (micro-F1 = 0.930; frequency correlation $r = 0.998$) while achieving a chance-corrected partition agreement of ARI = 0.369 and NMI = 0.688. The method's one clear failure is biologically expected and precisely localised: the $CD34^{+}CD38^{+}CD123^{-}$ HSPCs and $CD34^{+}CD38^{lo}$ HSCs are two adjacent segments of a single CD38 continuum that differ in essentially one marker, and isotropic Euclidean clustering cannot separate them, merging the rarer HSCs into the HSPCs (HSC F1 = 0). The remedy is not a better distance but a better *structure*: targeted sub-clustering of the $CD34^{+}$ compartment or graph-based/topology-preserving methods (Leiden, FlowSOM), optionally with marker weighting, and—ultimately—panels that measure more than CD38 to define the stem compartment. Per-cell cluster assignments and all code are provided to reproduce every number in this report.

Data and Code Availability

The Levine_32dim benchmark is publicly available via the `HDCytoData` Bioconductor package [10]. This report is accompanied by: per-cell cluster assignments for all 265,627 cells (`per_cell_cluster_assignments.csv`) and for the 104,184 labeled cells (`per_cell_cluster_assignments_la`

the full machine-readable evaluation (`evaluation_results.json`), including the complete confusion matrix, per-cluster mapping and purity, all 32 marker-difference values, and metric definitions; and the analysis scripts `preprocess_data.py`, `cluster_cells.py`, and `evaluate_mapping.py`. All results use fixed random seed 42 and the software versions listed in Methods.

References

[1] Dmitry R. Bandura, Vladimir I. Baranov, Olga I. Ornatsky, Alexei Antonov, Robert Kinach, Xudong Lou, Serguei Pavlov, Sergey Vorobiev, John E. Dick, and Scott D. Tanner. Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical Chemistry*, 81(16): 6813–6822, 2009. doi: 10.1021/ac901049w.

[2] Matthew H. Spitzer and Garry P. Nolan. Mass cytometry: Single cells, many features. *Cell*, 165(4):780–791, 2016. doi: 10.1016/j.cell.2016.04.019.

[3] Sean C. Bendall, Erin F. Simonds, Peng Qiu, El-ad D. Amir, Peter O. Krutzik, Rachel Finck, Robert V. Bruggner, Rachel Melamed, Angelica Trejo, Olga I. Ornatsky, Robert S. Balderas, Sylvia K. Plevritis, Karen Sachs, Dana Pe'er, Scott D. Tanner, and Garry P. Nolan. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, 2011. doi: 10.1126/science.1198704.

[4] El-ad D. Amir, Kara L. Davis, Michelle D. Tadmor, Erin F. Simonds, Jacob H. Levine, Sean C. Bendall, Daniel K. Shenfeld, Smita Krishnaswamy, Garry P. Nolan, and Dana Pe'er. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, 31(6):545–552, 2013. doi: 10.1038/nbt.2594.

[5] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426.

[6] Peng Qiu, Erin F. Simonds, Sean C. Bendall, Kenneth D. Gibbs Jr., Robert V. Bruggner, Michael D. Linderman, Karen Sachs, Garry P. Nolan, and Sylvia K. Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology*, 29(10):886–891, 2011. doi: 10.1038/nbt.1991.

[7] Sofie Van Gassen, Britt Callebaut, Mary J. Van Helden, Bart N. Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015. doi: 10.1002/cyto.a.22625.

[8] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe'er, and Garry P. Nolan. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015. doi: 10.1016/j.cell.2015.05.047.

[9] Lukas M. Weber and Mark D. Robinson. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, 89(12):1084–1096, 2016. doi: 10.1002/cyto.a.23030.

[10] Lukas M. Weber and Charlotte Soneson. HDCytoData: Collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats. *F1000Research*, 8:1459, 2019. doi: 10.12688/f1000research.20210.2.

[11] Sergei Doulatov, Faiyaz Notta, Elisa Laurenti, and John E. Dick. Hematopoiesis: A human perspective. *Cell Stem Cell*, 10(2):120–136, 2012. doi: 10.1016/j.stem.2012.01.006.

[12] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

[13] David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, pages 1177–1178, 2010. doi: 10.1145/1772690.1772756.

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2 (1):193–218, 1985. doi: 10.1007/BF01908075.

[16] Mickie Bhatia, Jean C. Y. Wang, Ulrike Kapp, Dominique Bonnet, and John E. Dick. Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice. *Proceedings of the National Academy of Sciences of the United States of America*, 94 (10):5320–5325, 1997. doi: 10.1073/pnas.94.10.5320.

[17] Ravindra Majeti, Christopher Y. Park, and Irving L. Weissman. Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. *Cell Stem Cell*, 1(6): 635–645, 2007. doi: 10.1016/j.stem.2007.10.001.

[18] Danilo Pellin, Mariana Loperfido, Cristina Baricordi, Samuel L. Wolock, Annita Montepeloso, Olga K. Weinberg, Alessandra Biffi, Allon M. Klein, and Luca Biasco. Human CD34+ hematopoietic stem cell hierarchy: How far are we with its delineation at the most primitive level? *Blood*, 142(6):509–518, 2023. doi: 10.1182/blood.2022018071.

[19] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

[20] Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019. doi: 10.1038/s41598-019-41695-z.

[21] Lukas M. Weber, Malgorzata Nowicka, Charlotte Soneson, and Mark D. Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications Biology*, 2:183, 2019. doi: 10.1038/s42003-019-0415-5.

[22] Fernando Anjos-Afonso, Florian Buettner, Syed A. Mian, Hefin Rhys, Jesus Perez-Lloret, Manuel Garcia-Albornoz, Namrata Rastogi, Linda Ariza-McNaughton, and Dominique Bonnet. Single cell analyses identify a highly regenerative and homogenous human CD34+ hematopoietic stem cell population. *Nature Communications*, 13:2048, 2022. doi: 10.1038/s41467-022-29675-w.