Nabo – a framework to define leukemia-initiating cells and differentiation in single-cell RNA-sequencing data

Parashar Dhapola, Mohamed Eldeeb, Amol Ugale, Rasmus Olofzon, Eva Erlandsson, Shamit Soneji, David Bryder*, Göran Karlsson*

Division of Molecular Hematology, Lund Stem Cell Center, Lund University

*Contributed equally

Correspondence: Göran Karlsson, BMC B12, 221 84 Lund, Sweden
e-mail: goran.karlsson@med.lu.se
Tel: +46 730 866785

ABSTRACT

Single-cell transcriptomics facilitates innovative approaches to define and identify cell types within tissues and cell populations. An emerging interest in the cancer field is to assess the heterogeneity of transformed cells, including the identification of tumor-initiating cells based on similarities to their normal counterparts. However, such cell mapping is often confounded by the large effects on total gene expression programs introduced by strong perturbations such as an oncogenic event. Here, we present Nabo, a novel computational method that allows mapping of cells from one population to the most similar cells in a reference population, independently of confounding changes to gene expression programs initiated by perturbation. We validated this method on multiple datasets from different sources and platforms and show that Nabo achieves higher rates of accuracy than conventional classification methods. Nabo is available as an integrated toolkit for preprocessing, cell mapping,

differential gene expression identification, and visualization of single-cell RNA-Seq data. For exploratory studies, Nabo includes methods to help evaluate the reliability of cell mapping results. We applied Nabo on droplet-based single-cell RNA-Seq data of healthy and oncogene-induced (MLL-ENL) hematopoietic progenitor cells (GMLPs) differentiating in vitro. Despite a substantial cellular heterogeneity resulting from differentiation of GMLPs and the large transcriptional effects induced by the fusion oncogene, Nabo could pinpoint the specific cell stage where differentiation arrest occurs, which included an immunophenotypic definition of the tumor-initiating population. Thus, Nabo allows for relevant comparison between target and control cells, without being confounded by differences in population heterogeneity.

INTRODUCTION

It is increasingly accepted that the majority of tumors are organized in cellular hierarchies, where the cancer stem cells or tumor-initiating cells (TICs) drive tumor growth (Clevers, 2011). TICs possess unique stem cell characteristics such as self-renewal capacity, quiescence and drug resistance, which underlie metastasis and relapse. TICs are therefore a critical priority as targets for therapy. The development of improved immune-deficient mouse strains together with refinements to FACS-based protocols for prospective isolation of functionally distinct progenitor populations have paved the way for the definition of the immunophenotype of a plethora of TIC-containing cell populations (Al-Hajj et al., 2003; Boiko et al., 2010; Bonnet and Dick, 1997; Quintana et al., 2008; Schepers et al., 2012; Singh et al., 2004; Wang and Dick, 2005). Even though these efforts have been instrumental in shedding light on TIC biology, including the identification of leukemia-initiating

progenitor populations, they have also demonstrated that immunophenotypically defined populations are predominantly heterogeneous. Consequently, many times only a fraction of the cells are relevant for the research aim. Solving the cellular heterogeneity within TIC-containing fractions is particularly critical for molecular characterization and therapeutic target-identification, as transcriptional programs might otherwise be confounded by irrelevant cells.

During the last half a decade, a series of technological breakthroughs have radically improved our ability to quantify transcriptomes at the level of single cells (Hashimshony et al., 2012; Jaitin et al., 2014; Macosko et al., 2015; Picelli et al., 2014; Gierahn et al., 2017; Zheng et al., 2017). Clubbed under the generic term single-cell RNA-Seq (scRNA-Seq), this array of technologies has allowed for detailed analysis of cellular heterogeneity (Zeisel et al., 2015; Tirosh et al., 2016). Indeed, scRNA-Seq has led to the identification of new cell types (Grün et al., 2015; Ramsköld et al., 2012; Villani et al., 2017), deconstruction of developmental programs and lineage hierarchies (Treutlein et al., 2014; Trapnell et al., 2014; Shin et al., 2015; Blakeley et al., 2015; Moignard et al., 2015; Paul et al., 2015), spatial localization of cells in tissues (Achim et al., 2015; Satija et al., 2015), effects of high-throughput gene editing (Dixit et al., 2016) and cell reprogramming (Treutlein et al., 2016). A flurry of analysis methods has followed experimental innovations in this area. Most algorithms and software have focused on certain aspects of scRNA-Seq: data normalization (Vallejos et al., 2017), identification of cell clusters (Andrews and Hemberg, 2018), construction of lineage trajectories (Herring et al., 2018) and identification of differentially expressed genes (Jaakkola et al., 2017). Most recently,

data integration from multiple experiments has received necessary attention (Butler et al., 2018; Haghverdi et al., 2018; Kiselev et al., 2018).

Increasing efforts in the genomics area now focus on investigating changes in heterogeneity following a cellular or molecular perturbation. This is of particular interest in cancer research. Here, the effect on cellular heterogeneity by oncogenic transformation or drug treatment has a critical impact for defining TICs, understanding the molecular mechanisms behind therapy resistance, and identifying TIC-specific therapeutic targets. One conceptual strategy to analyze such data is to perform 'cell mapping', wherein one of the samples is considered to be a base or reference population (for example a healthy cell population). Individual cells from one or more perturbed populations, called target populations hereon (e.g. tumor cells) are then mapped/projected to the reference population with the objective to identify their most similar counterparts. The current methods of cell mapping (Kiselev et al., 2018), cell alignment (Butler et al., 2018) and batch correction (Haghverdi et al., 2018) rest on the critical assumption that the dissimilarity between the test cells and one or more reference subgroups is smaller than that between at least any one pair of reference subgroups. However, this assumption may not hold true in experimental settings wherein the expression variance of genes responsible for the cellular heterogeneity is smaller than the molecular response to the perturbation. This is often the case when comparing a cancer cell population to its heterogeneous population-of-origin and if cells have been perturbed to alter lineage determining transcriptional networks. Thus, to increase the impact of scRNA-Seq technologies, development of novel bioinformatics tools for improved cell mapping is warranted.

To address these challenges, we here introduce Nabo, a novel computational approach for cross-population cell mapping. Nabo provides a graph-theory based approach to statistically validate mapping and allows integrative comparison of multiple target populations to the same reference population. Using a variety of published as well as in-house generated datasets, we show that Nabo performs equally or better than currently available methods for conventional cell mapping. To demonstrate the power and impact of Nabo for cancer research, we perform scRNA-Seq analysis on the cancer stem cell-containing population from our MLL-ENL mouse model for acute myeloid leukemia (Ugale et al., 2014) and demonstrate how Nabo, unlike current state-of-the art cell-mapping methods, readily identifies leukemia-initiating cells within a heterogeneous population. Thus, Nabo represents a novel tool for relevant analysis of scRNA-Seq data in which a perturbation to an originally heterogeneous population results in a large molecular change to the target cells. As such, Nabo has a critical implementation in cancer research for detection of TICs.

RESULTS

Nabo maps cell populations across datasets with high accuracy

The generation of mouse models for leukemia by enforced expression of oncogenic fusion genes has been instrumental for our current conceptual understanding of the origin of cancer stem cells. Interestingly, while the oncogenic targeting of hematopoietic stem cells almost consistently results in leukemic transformation, other progenitor populations also has transformation potential (Eppert et al., 2011;

Heuser et al., 2011; Huntly et al., 2004; Krivtsov et al., 2006; Somervaille and Cleary, 2006; Ugale et al., 2014). Thus, the generation of cancer stem cells does not necessarily include high-jacking of the normal stem cells' molecular machinery, but could also occur due to re-activation of critical stem cell programs in progenitor populations (Eppert et al., 2011; Krivtsov et al., 2006). In fact, we have recently shown that the MLL-ENL fusion gene exclusively transforms heterogeneous progenitor populations with myeloid differentiation potential downstream of the hematopoietic stem cells (Ugale et al., 2014). Here, we used droplet-based scRNA-Seq technology to try to address whether MLL-ENL transformation expands a specific subpopulation within the heterogeneous GMLP population, which would potentially reveal the origin of the leukemic stem cells in MLL-ENL AML.

Purified GMLPs from transgenic mice carrying a doxycycline (dox) inducible MLL-ENL fusion gene were cultured for five days with or without dox and approximately 2,000 cells were subsequently processed for scRNA-Seq using the Chromium droplet-based platform. T-SNE visualization of the scRNA-Seq data (**supplementary figure 1A**) showed that the un-induced (WT) and induced (MLL-ENL) cells divided into two separate clusters, suggesting large molecular differences caused by the expression of the fusion gene. To rule out that this was not simply due to batch effects or other technical artefacts, we used Seurat's canonical correlation analysis (CCA) based approach of combining datasets. The t-SNE plot of WT and MLL-ENL cells post CCA alignment (**supplementary figure 1B**) did not show any obvious cell clusters and the population structure that was otherwise observed in WT cells alone was lost after CCA alignment. Together these analyses demonstrate the weakness of conventional bioinformatics tools in comparing single-cell RNA-seq data from a

heavily perturbed cell fraction to its heterogeneous population of origin. The dramatic molecular effect of the perturbation that a pre-leukemic lesion represents overshadows the gene expression variation that otherwise separates the different subpopulations within the heterogeneous progenitor population, and thus eliminates the advantages of the single-cell dimension from the analysis.

To specifically address these challenges, we designed a novel computational tool for cross-population cell mapping, called Nabo. When using Nabo to perform cell mapping, samples are divided into reference and test populations, where the test cells are projected or mapped over the reference population. Nabo defines a relationship between cells from the reference sample by creating a shared nearest neighbor (SNN) graph (Jarvis and Patrick, 1973). To create an SNN graph, the distance between each pair of cells is calculated based on the expression levels of genes. For each cell, an arbitrary number of cells that have the least distance to the given cell are identified and are regarded as that cells' neighbors. Subsequently, cells that have common(shared) neighbors are connected to each other and the strength of each connection is determined by the number of shared neighbors (**figure 1A**). After creation of the SNN graph, Nabo maps cells from one or more target samples onto this reference graph by identifying an arbitrary number of the most similar reference cells for each target cell (**figure 1B**). The test cells are then connected to the reference graph by identifying the shared neighbors. Importantly, whether from the same or different samples, the test cells are always mapped independently from each other and hence can never have connections between themselves (see Methods for further details) (**figure 1C**). This feature will circumvent separation between test and reference cells based on large molecular effects due to

the specific perturbation. Upon mapping, Nabo provides a 'mapping score' to each of the reference cells based on the number and strength of connections that were made to that reference cell from the target cells (**figure 1D**). Thus, the higher the mapping score of a cell, the higher its similarity to the target cell relative to other reference cells. The reference populations can be partitioned into clusters (**figure 1E**) and cluster-wise mapping scores can be identified to ascertain if the test cells were significantly similar to a subgroup of the reference population (**figure 1F**).

Using scRNA-Seq data from freshly isolated human peripheral blood mononuclear cells (66,000 peripheral blood mononuclear cells) as well as for purified CD19+ B cells, CD14+ monocytes and CD56+ natural killer (NK) cells (Zheng et al., 2017), Nabo correctly mapped cells of known identity onto a more heterogeneous group of cells containing multiple cell types (**figure 2A**). We found that the mapping score of the reference cells was significantly clustered ($p < 1e\text{-}50$; proportions z-test) to the cells expressing the corresponding marker gene (**supplementary figure 2A**). Reference cells expressing CD79A and MEIS1 (B cell marker genes) were mapped by CD19+ B cells with a specificity of 0.976 and 0.995. Similarly, for CD14+ Monocytes: 0.975 (CD14), 0.997 (FTL), 0.988 (LYZ) and for NK cells 0.993 (GNLY) and 0.994 (NKG7) mapping specificity was found. This indicated that Nabo could reliably project cells onto a heterogeneous population with high accuracy. Importantly, Nabo did not require any prior cluster knowledge to perform cell mapping.

To illustrate Nabo's ability to perform cell identity prediction, we took advantage of two publically available scRNA-seq datasets containing 6,000 (6K) and 33,000 (33K)

PBMCs obtained from a healthy donor. We created a reference graph of the 33K dataset, partitioned it into clusters and created a t-SNE layout of the cells for visualization (**supplementary figure 2B**). Using Nabo, we mapped the cells from the 6K dataset cells onto this reference graph, with the objective of identifying the heterogeneity of the 6K cells (**figure 2B**). The cluster identities generated by Nabo correlated to 90.6 % with clusters generated by regular Seurat analysis of the 6K dataset alone (**figure 2C**). Nabo outperformed Random Forest (κ: 0.395), a general classification algorithm, as well as scmap-cell (κ: 0.772) (Kiselev et al., 2018), a single-cell RNA-Seq cell projection tool. Mapping scores were then generated for the 33K cells by individual mapping of each cluster from the 6K data. Importantly, the mappings of distinct clusters from the 6K population were strongly restricted to the 33K cluster, with similar expression levels of marker genes (**figure 2D**). Quantification analysis revealed that 92.34% of all mapping scores were ascribed to cells with correct cluster identity (**supplementary figure 2C**).

Finally, we explored Nabo's ability to perform cell mapping when reference and target samples are from different studies and use different scRNA-Seq platforms. For this, we used 4 different published scRNA-Seq datasets from pancreatic islets of Langerhans (Baron et al., 2016; Muraro et al., 2016; Segerstolpe et al., 2016; Xin et al., 2016). We chose one of the datasets as the reference population (Baron et al.; InDrop sequencing) and, using Nabo, constructed an SNN graph of the same dataset and calculated a force directed layout of the graph for visualization (**figure 3A**). After mapping the other three datasets onto this reference graph (**figure 3B**), we found that Nabo was able to correctly map 69.3% (Xin et al, C1-IFC SMARTer platform (κ: 0.687)), 63.79% (Muraro et al., CEl-Seq2 platform (κ: 0.733)) and

88.36% (Segerstolpe et al SMART-Seq2 platform (κ: 0.927)) of the target cells to their appropriate cluster type as defined by the reference dataset (**figure 3C**). In comparison, scmap-cell correctly mapped 78.75% (κ: 0.665), 81.32% (κ: 0.764) and 50.65% (κ: 0.45) of cells in the three datasets. Thus, as per the Cohen's kappa scores, Nabo performed equally well- or better than scmap-cell in cell mapping using data obtained from different platforms. Together, these experiments demonstrate that Nabo performs equally or better than currently available methods for several different aspects of conventional cell mapping.

Identification of MLL-ENL leukemia-initiating cells using Nabo

Having an improved mapping tool at hand, we used Nabo on our scRNA-Seq data to compare the heterogeneity of MLL-ENL-transformed and WT GMLPs. A shared nearest neighbor graph of WT cells was created and partitioned into 12 clusters (**figure 4A**). MLL-ENL cells were projected onto this WT graph, providing a mapping score for each WT cell that signifies the reference cell's similarity to the MLL-ENL cells (**figure 4B**). MLL-ENL cells were classified as members of either one of the WT graph clusters or they remained unassigned if enough evidence to perform assignment was lacking. Intriguingly, WT cells with high mapping score were significantly ($p<1e-6$; Chi-squared test) closely connected on the graph where cluster 5 of WT cells had the absolute highest average mapping score (5.02) followed by cluster 11 (1.89) (**figure 4C**). Of all MLL-ENL cells that were assigned to a cluster, 70.44% were assigned to cluster 5. In contrast, scmap-cell was unable to assign the vast majority (98.73%) of MLL-ENL cells to any cluster, while scmap-cluster assigned most of the MLL-ENL cells (77.5%) to cluster 6. However, with increased stringency (see Methods), scmap-cluster failed to assign 81.56% of MLL-ENL cells

to any cluster (**figure 4D**). To test how the results of each tool were dependent on the clustering itself, we used Seurat to perform clustering of the reference population (**supplementary figure 3A-B**) and then predicted the identity of MLL-ENL cells again. Even with a different partitioning of the cells, Nabo, under two different stringency settings, assigned 56.18% and 45.6% of MLL-ENL cells to Seurat's cluster 5, which mostly comprised of the same cells as the cluster 5 from the SNN graph (**figure 4A**). Using this strategy, both scmap-cell and scmap-cluster (with high stringency settings) left 69.29% and 69.96% cells unassigned. Scmap-cell however, did assign 20.98% of cells to Seurat's cluster 5. With default settings, scmap-cluster assigned 45.51% of cells to cluster 5. Interestingly, scmap-cluster assigned only 1 cell to cluster 10 and no cell to cluster 11, which combined constituted the cells from the SNN cluster 6 (**figure 4A**), to which scmap-cluster had assigned 916 cells (77.5%) (**supplementary figure 3C**). The mapping performed by Nabo clearly indicates that the MLL-ENL cells bear a strong relative preference for cells that are from cluster 5. We also found that Nabo, unlike scmap, consistently predicted the association of MLL-ENL cells to WT clusters across different cluster partitioning.

As a control to address if the results obtained from Nabo-generated mapping were due to technical biases such as -cell-cycle effect or sequencing depth, we used low variance genes (LVGs) rather than high variance genes to create the reference graph and perform the mapping. We visualized the mapping scores on LVG reference graph using the same layout as for the actual reference graph. This revealed a substantially more even distribution of cells across the graph (**supplementary figure 4A**). Additionally, the distribution of LVG mapping scores across the clusters was relatively uniform, with cluster 5 having an insignificantly

higher mapping score distribution than others (**supplementary figure 4B**). However, in absolute terms, the mean mapping score in cluster 5 was 33.78 times lower in LVG mapping compared to original (HVG) mapping. Also, using LVG mapping, Nabo failed to assign 92.89% of cells to any cluster (**supplementary figure 4C**). To investigate the robustness of the mapping, we prevented MLL-ENL cells from mapping to reference cells that had received a mapping score higher than 1 in the actual mapping (10.75% cells). If WT cells with high mapping scores were not significantly more similar to MLL-ENL cells than to the rest of the WT cells, then in this control setup, called blocked mapping, a subset of these cells will receive similarly high mapping scores. While 58 of the reference cells received a mapping score higher than 5, only one cell passed this threshold in blocked mapping (**supplementary figure 4D**). Under these blocked mapping conditions, cluster 9 received the highest mapping scores (**supplementary figure 4E**). However, this was still 5.86 times lower than the mean mapping score of cluster 5 from the actual mapping (figure 4c). Also, 64.13% of cells were not assigned to any cluster by Nabo's classifier (**supplementary figure 4F**), which was 2.33 times higher than observed in the actual mapping (**figure 4C**). Overall, these two controls indicated that the mapping of MLL-ENL on WT cells was significant and non-trivial. As these controls are critical and useful methods for evaluating mapping results in exploratory studies, they have been assigned to Nabo for easy inclusion in an analysis workflow. Within Nabo, we have also included a heuristic to identify mapping specificity of each target node. Mapping specificity of a node will be higher if all the reference nodes it connects to are also close to each other. When applied to MLL-ENL cells, we found that the cells that projected to cluster 5 and cluster 11 (**figure 4A and 4C**) had higher average mapping specificity than other MLL-ENL cells (**supplementary figure 5A**).

These mapping specificities can also be viewed from a reference cell's perspective, by averaging the mapping specificity of all the target cells that map to a given reference cell (**supplementary figure 5B**).

To define the identity of WT cells in cluster 5, to which an overwhelming majority of the MLL-ENL cells had mapped, we used Nabo to identify marker genes specifically expressed within this cluster. The cumulative expression of these marker genes were then cross-referenced to known cell types from bulk transcriptome datasets obtained from the BloodSpot database (Bagger et al., 2016). Using the annotations of each cluster, a myeloid and a lymphoid differentiation trajectory could be inferred in the WT graph (**figure 4A**). Interestingly, within this trajectory, the MLL-ENL-mapped cluster 5 as well as cluster 9 represented subpopulations with the most primitive molecular signatures identified as LMPP- and Pre-GM-like, respectively (**supplementary figure 6**). This observation is in concordance with previous reports that suggest that overexpression of oncogenes in pluripotent/multipotent cells can prevent primitive cells from terminally differentiating and subsequently develop into tumors (Cozzio et al., 2003).

To further investigate the gene signature of 'MLL-ENL like WT cells', we focused only on a subset of cells within cluster 5 that received a mapping score greater than 1 (92 cells) and compared them to their nearest 89 cell neighbors (see Methods) (**supplementary figure 7A**). We found 96 genes with significantly higher expression (adjusted p-value < 0.05) in MLL-ENL like WT cells, including FLT3, BCL2, SOX4 and AFF3 (**supplementary figure 7B**). Many of these genes were not just higher in test cells compared to control cells, but also when compared to rest of the cells on

the graph (**supplementary figure 7C**). Interestingly, WT cells with high mapping scores demonstrated a primitive gene expression signature when compared to the rest of WT cells (**supplementary figure 8A**), but a more differentiated signature compared to MLL-ENL cells that mapped to these WT cells (**supplementary figure 8B**). On the other hand, MLL-ENL cells displayed a higher expression of genes that are normally expressed in hematopoietic stem cells (**supplementary figure 8C**). These results strongly indicate that MLL-ENL induction perturbs a primitive subpopulation of GMLPs and induces a gene expression program that is normally found upstream in the lineage hierarchy.

To further verify that the MLL-ENL induced cells were actually in a less differentiated state than the WT GMLPs, we chose to map the MLL-ENL-transformed GMLPs against total cKit+ BM cells (Säwen et al., 2018), which consists of a wider range of hematopoietic progenitors (**supplementary figure 9A**). First, we mapped 500 highly purified hematopoietic stem cells (HSCs) (LSK CD150+ CD48-; (Säwen et al., 2018)) to the reference graph (**figure 5A**). Interestingly, these 500 cells mapped almost exclusively to a rare group of 7 cells within the cKit+ reference graph, demonstrating the stringency of Nabo and establishing the location of the most primitive cell type within the trajectory of the graph. We then mapped the WT GMLPs and MLL-ENL induced GMLPs on the cKit+ cells (**figure 5B-C**). To quantify the differentiation state of the reference cells, we assigned the differentiation potential for each cell (Weinreb et al., 2018) (**figure 5D**) and inferred a pseudo-time axis of the data (**figure 5E**). This revealed that the sorted HSCs were most upstream, followed by MLL-ENL induced GMLPs and then WT GMLPs (**figure 5E**), thereby confirming our earlier observations (**figure 5A-C**). When annotating the cell clusters of the cKit+ reference

population, we found that WT GMLPs mapped cells had signatures of GMPs and differentiated granulocytes, while MLL-ENL mapped cells had highest similarity to gene signatures of preGMs and LMPPs (**supplementary figure 9B**). Thus, Nabo is a powerful tool to predict the identity of TICs from scRNA-Seq data.

Nabo has a broad implementation for mapping perturbed populations of different cell systems

Similar to oncogenic transformation, overexpression or deletion of transcription factors may cause large-scale transcriptional changes and disrupt the heterogeneity of a cell population. Such effects will create substantial differences between the normal and perturbed cells, resulting in their independent clustering. To evaluate if Nabo would be a useful tool for these kinds of experimental settings, we used a scRNA-seq dataset comparing differentiating WT murine embryonic stem cells (ESCs) with ESCs deficient for the transcription factor YY1 (Weintraub et al., 2017).

Using tSNE visualization, it was previously observed that YY1- cells clustered away from the YY1+ cells, resulting in a loss of single-cell resolution (Weintraub et al., 2017). In an attempt to compare the heterogeneity between YY1+ and YY1- cells, we used Nabo and created a reference graph for YY1+ cells that was visualized using a force directed layout after partitioning the graph into six clusters (**figure 6A**). After mapping the YY1- cells (**figure 6B**), we found that the mapping scores were mainly concentrated in the clusters corresponding to clusters 4, 5 and 6 (**figure 6C**). Visualizing marker genes for pluripotency and primary germ layers **(figure 6D**) revealed that clusters 5 and 6 were dominated by endodermal and mesodermal

molecular signatures, respectively. As the mapping scores were largely low or absent in cells with pluripotency and ectodermal lineage marker expression, we inferred that YY1 depletion in murine ESCs either results in accelerated differentiation towards mesodermal and endodermal lineages, or in the loss of pluripotent and ectodermal progenitor cells. This finding is consistent with the notion that YY1 is critically important for ectodermal development (Satijn et al., 2001) and demonstrate the utility of Nabo to define cellular heterogeneity following different types of perturbation in a broad range of cell types.

DISCUSSION

Recent advances in RNA sequencing methodology has allowed for measurements of global gene expression programs in individual cells as a proxy for their function. Thus, scRNA-Seq experiments offer an unprecedented possibility to dissect cellular heterogeneity in tissues or purified cell-fractions, that is now extensively used to visualize cellular hierarchies and compositions throughout the entire human body. Most approaches for analysis of scRNA-seq data have focused on statistical methods to discriminate cells by clustering into groups or trajectories. However, for scRNA-Seq to become useful when approaching changes in heterogeneity during situations of major perturbations of molecular programs, computational methods that allow for comparison of relevant information between datasets are critical. One way to integrate such data is through cell mapping, which allows identification of cell-cell relationships even when the data is spread across experiments, alternative sequencing platforms and studies. With Nabo, we aimed to develop an accessible and interpretable platform to perform cell mapping. The control datasets indicated

that Nabo has equal or improved accuracy than existing methods in conventional classification of target cells based on reference cell clusters. Nabo supplements two other recently published algorithms that aim to integrate scRNA-Seq datasets, mmCorrect (mutual nearest neighbors correction) (Haghverdi et al., 2018) and Seurat's CCA (canonical covariate analysis) based cell alignment (Butler et al., 2018). Both of these algorithms are geared towards removal of batch effects between the datasets. The implicit assumption made by these algorithms is that the differences between the populations being compared is mostly technical in nature. Nabo is not based on such assumptions. Instead, Nabo allow for relevant comparison of heterogeneity between populations even if they have large differences in gene expression profiles due to a biological component, such as overexpression of an oncogene. MNNcorrect is more useful when a user wants to obtain batch corrected expression values, while Seurat's CCA can be useful in scenarios where the user wants an integrated low-dimensional embedding of two or more datasets. Both of these approaches, however, lead to changes of cell embeddings of the reference population for each new population/sample included. For example, the t-SNE layout of cells will change when new samples are included. Nabo solves this by providing a quantitative measure that can be ascribed to each cell of the reference sample. In this way, no matter which and how many target populations are projected, the reference cells can always be visualized in their original space.

In cell mapping and other predictive analytics approaches that use scRNA-seq data, one major challenge has been the validation of the results. We demonstrated two innovative approaches that Nabo uses to evaluate the mapping reliability that we believe could be particularly useful for experiments where little prior knowledge

about the mapping exists. The implementation of Nabo was done with two objectives: high accuracy and low memory requirements, so that the software can be run on ordinary desktops and laptops with modest hardware requirements. The tradeoff for this implementation was a runtime that increases quadratically with cell number. We were able to process PBMC data with 66,000 cells within 6hrs with just 8 GB of RAM required. Most scRNA-Seq experiments carried out today are well within that range.

Importantly, we also demonstrated an exclusive capacity of Nabo to ascribe mapping scores to reference cells. This allows for quantitative assessment of mapping and subsequent identification of specific groups of cells in the reference population that are similar to the test cells, which is advantageous when trying to determine the exact identity of TICs. Unlike current methods, Nabo is designed to ignore the massive transcriptional changes associated with a strong perturbant such as the onset of a strong oncogene and allow for cell mapping entirely based on expression of molecular signatures associated with heterogeneity. Using Nabo on scRNA-Seq data acquired from the TIC-containing GMLP population from our MLL-ENL mouse model of AML, we could identify a distinct target population characterized by a primitive and multipotent molecular signature. Interestingly, this population could be discriminated from other more differentiated GMLP populations by the expression of fms-like tyrosine kinase 3 (Flt3). Together, these results validate the usefulness of Nabo as a tool for scRNA-Seq analysis of tumor populations with the aim of defining the changes in heterogeneity caused by oncogenic transformation and subsequent TIC identification.

METHODS

Nabo overview

Nabo uses HDF5 file format to store the data on the disk. The data is stored in gene-wise and cell-wise manner to allow quick subselection across both axes. For mapping, the reference dataset is first normalized (library size normalization) and the selected features are subjected to standard scaling. The data is normalized and scaled on the fly as it is being loaded from the disk. The scaled data is subjected to PCA reduction using an out-of-core (incremental) implementation of PCA. The Euclidean distances are computed between each pair of cells in the PCA space to identify k-nearest neighbor of each reference cell. The shared nearest neighbors are identified for each pair of KNN neighbors and if they have non-zero shared neighbors then an edge is added between the two cells with weight equal to the ratio of number of shared neighbors (*s*) to s – maximum possible shared neighbors. The cells to be projected are too library scale normalized but their features are scaled as per the mean and standard deviation of features in the reference dataset. This scaled target data is then projected into the PCA space trained on reference data. The distance of each target cell to every reference cell is calculated using a modified Canberra metric. The metric is modified such that if in a given dimension the distance between target (*t*) and reference (*r*) cell value is greater than $f * r$, where f is a predefined factor between a range of 0-1, then distance is set as 1 (highest value). This means that if the target cell has a very high value in a given dimension, then that dimension would automatically cause the distance to saturate.

Mapping score calculation

Mapping score is calculated using the `*get_mapping_score*` function of the *Graph* class. The score for a reference cell is calculated by calculating the weighted sum of all of its incoming projections. Mapping scores are always normalized to the number of projected cells, to allow comparison between two mapping score distributions as long as the other parameters remain similar.

Classification of cells

The classification of target cells is performed by the `*classify_targets*` function of the *Graph* class. The target cells are classified to one of the reference clusters using voting method. The projections (edges) of a given target cell to reference cells are grouped based on the cluster identity of the reference cells. A weighted sum of projections is calculated for each of the groups. If a single group has a weighted sum that is higher than predefined threshold (default: 50% of total weighted sum), then the cell is classified to that cluster otherwise the target cell remains unassigned to any cluster. As an additional filter, projections can be discarded during calculation of cluster-wise weighted sums based on individual weight of each edge connection between a target cell and a reference cell. Furthermore, a target cell can directly be classified as unassigned if it has fewer projections than a pre-set cutoff. These two tunable parameters ensure that users have fine control on classification accuracy.

Differential gene expression calculation

Mann-Whitney U test (as available in the scipy.stats package in Python) is used to identify genes that are differentially expressed between two groups of cells. P values are corrected for multiple hypotheses testing using the Benjamini/Hochberg method (as implemented in the statsmodels package). If the number of cells in the control

are higher than the number of cells in the test group, the same number of cells as in the test group are selected from the control group. Such subselection of cells is performed after sorting the control cells in order to retrieve top *nc* values (from highest to lowest), where *nc* in number of test cells. This strategy helps improve the specificity of the results. To identify the genes driving a mapping specificity, reference cells with mapping score higher than a given threshold are selected (test cells); thereafter all the cells that are at the node distance of *n* (defined by the user) from the test cells, in the reference graph, are marked as control nodes against which genes upregulated in test nodes are identified. The higher values of *n* will cause identification of larger differences in transcriptomes of mapped and unmapped cells, while smaller values of *n* will highlight the smaller differences between the control and test cells.

Hematopoietic gene signature identification

Gene sets were queried against the 'normal mouse hematopoiesis' dataset that contains cell types from the BloodSpot database (Bagger et al., 2016). The median value of the gene set in each cell type was determined and these values were min-max scaled across cell types. The results were visualized as area plots in polar coordinates to enable quick assessment for presence of cell type signature in a gene set. +/-1 standard deviation is also calculated by usage of biological replicate data of cell types and also visualized to provide a quick assessment of noise in the results.

Data processing

All the datasets were subjected to cell filtering, HVG identification, Louvain clustering, UMAP and/or tSNE embedding generation and marker gene identification; all using Seurat version 3.0.1. Parameters used of each individual dataset can be found in the online repository. The HVGs identified by Seurat were used to perform reference graph generation of the respective datasets. The same list of HVGs was also used to train classifiers and scmap.

Single-cell RNA-Seq of in vitro cultured GMLPs and MLL-ENL induction

Granulocyte-macrophage-lymphoid progenitors (GMLPs) were enriched from BM of WT and iMLL/ENL mice by depletion of mature cells using biotinylated antibodies against lineage markers (CD4, CD8a, B220, CD11b, Gr1and Ter-119) and anti-biotin conjugated magnetic beads, according to manufacturer□s instructions (Miltenyi Biotech, Germany). GMLPs were sorted as Lin-Sca1+c-Kit+CD48+CD150- on a FACS Aria II or III cell sorter (Becton Dickinson, San José, CA). Propidium iodide (Invitrogen, Carlsbad, CA) was used to exclude dead cells. The 10,000 sorted GMLPs were maintained in OptiMEM (Invitrogen, Carlsbad, CA) supplemented with 10% FCS, 0.1 mM β-mercaptoethanol (Invitrogen, Carlsbad, CA), 1x Penicillin/Streptomycin (Invitrogen, Carlsbad, CA), SCF (10 ng/ml), IL3 (5 ng/ml), G-CSF (5 ng/ml) (all from Peprotech Inc., Rocky Hill, NJ) and 1 µg/ml doxycycline (Sigma-Aldrich, St. Louis, MO). After 4 days, cells were harvested and subjected to single cell (SC) RNA sequencing. ScRNA-Seq data was generated on the 10X platform (10X Genomics) according to the manufacturer's instructions.

Antibody clones and suppliers are as follow:

Antibody	**Clone**	**Supplier**
Sca1	D7	Biolegend
Ckit	2B8	Ebioscience
CD150	TC15-12F12.2	Biolegend

CD48	HM48-1	Pharmingen

Data and code availability

All code used to analyze the data and generate the figures can be found here: http://github.com/parashardhapola/nabo_manuscript. This repository also contains the count matrices of the datasets in MTX or CSV format. Nabo format HDF5 files can also be found in the same repository. Source code of Nabo is available here: http://github.com/parashardhapola/nabo. API and tutorials for usage of Nabo can be found here: nabo.readthedocs.io. YY1 dataset was downloaded from GSE103574 (samples: GSM2774584 and GSM2774585). Pancreatic cells datasets were downloaded from these GEO repositories: GSE84133 (Baron et. al.), GSE85241 (Muraro et. al.), GSE81608 (Xin et .al.). Data for Segerstolpe et. al. was downloaded from Array Express archive E-MTAB-5061. PBMC 68K, 33K and 6K cell dataset was obtained from 10x genomics data portal as count matrices (MTX format) generated using Cell Ranger version 1.1.0. Data for murine HSCs and cKIT+ cells was obtained from GSE122473.

FIGURE LEGENDS

Figure 1: **Workflow of cell mapping using Nabo**. (**A**) SNN graph of reference population. (**B-C**) Projection of test cells over the reference SNN graph. (**D**) Reference cells sized based on their mapping score. (**E**) Reference cells colored based on their cluster identity. (**F**) Reference cells colored based on cluster identity and sized as per mapping score.

Figure 2: **The mapping accuracy of Nabo**. (**A**) Mapping of three purified cell populations, CD19+ B cells, CD14+ monocytes and CD56+ NK cells on fresh PBMC cells. In the top t-SNE plot of the reference population, i.e. PBMCs, the cells are connected to each other based on the SNN graph of PBMCs. Middle panel of t-SNE plots shows cells sized based on the mapping scores obtained from mapping of each individual purified population. The cells are colored in blue and the edges in grey. Absence of blue cells indicates that no mapping score was assigned to those reference cells. The bottom panel of t-SNE plots shows the reference cells colored based on expression of marker genes of the respective mapped populations; darker blue color indicates higher expression. (**B**) Mapping of individual clusters of cells from one dataset to another. The t-SNE layout of cells from the 6K dataset has been shown with cells colored based on their cluster identity. The clusters are labeled based on expression of canonical markers. The bottom t-SNE plot shows the reference population, i.e. 33K PBMC dataset. The cells are connected to each other as for the SNN graph. As no clustering was done on this dataset, all cells are colored blue. (**C**) Comparison of cell types in the 6K dataset identified by Nabo and Seurat. Each data point in the scatter plot is sized to indicate number of cells. (**D**) The top panel contains the t-SNE plots of the 6K dataset, with individual cells colored based on the expression of defined marker genes for the indicated cell type. The middle plot depicts a t-SNE layout of the 33K dataset sized based on mapping scores following mapping of the respective 6K cluster. The gray lines indicate the edges in the SNN graph of the 33K dataset. The bottom panel of t-SNE plots shows expression of marker gens in the 33K dataset in same order as the top panel.

Figure 3: **Mapping of pancreatic cells across datasets**. (**A**) A force directed layout of pancreatic cells obtained from Baron et al. showing clusters of cells labelled based on marker expression. This cell population was used as reference population for mapping from other datasets. (**B**) Each panel shows a force directed layout of reference cells, with cells scaled in size based on mapping scores obtained by mapping of specific cell types from a given study. NA panel indicate that the cell type was not present in that study. (**C**) Barplots showing the number of cells whose cell type was predicted either correctly, incorrectly or remained unassigned with respect to cell types reported in the original study.

Figure 4: **Mapping of MLL-ENL expressing GMLPs**. (**A**) A force directed layout of an SNN graph depicting normal GMLPs, with clusters assigned to numerals and colors. The different cell stages inferred from gene expression patterns is indicated. (**B**) SNN graph of normal GMLPs where cells have been sized based on their mapping score obtained after projection of MLL-ENL induced GMLPs. (**C**) Boxplots showing the distribution of mapping scores across the clusters. The red line in each box indicates the median value. (**D**) Barplots showing the number of MLL-ENL cells that were assigned to each reference graph cluster. The minimum weight for assignment using Nabo was set at either 0 or 0.1 (default). In case of scmap-cluster, the weight fraction was set either at 0.7 (default) or 0.9.

Figure 5: **Assessing differentiation of MLL-ENL cells**. Mapping of (**A**) purified HSC (**B**) WT GMLPs (**C**) and MLL-ENL induced GMLPs on the SNN graph of cKit+ cells. Cells have been size scaled proportional to the mapping score of respective target population. (**D**) The inferred differentiation potential c-kit+ cells. (**E**) The

differentiation potential of cells is shown on the x axis and the y axis shows the average mapping score on cells grouped into bin sizes of 100 cells.

Figure 6: **Nabo interrogation of the effects of YY1 depletion on murine embryonic stem cells.** (**A**) SNN graph of murine embryonic stem cells (mESCs) where cells have been colored based on their cluster identity. (**B**) SNN graph of mESCs, where node size has been scaled proportional to the mapping scores obtained on projection of YY1 depleted mESCs. (**C**) Distribution of mapping scores across the mESC clusters. (**D**) SNN graph of YY1+ ES cells with expression of marker genes for each germ layer as well as pluripotency associated genes highlighted.

Supplementary figure 1: (**A**) t-SNE plot showing cells from WT GMLPs (in blue) and MLL-ENL induced GMLPs (in red) as separate clusters. (**B**) t-SNE plot of WT and MLL-ENL induced GMLPs obtained after applying Seurat's CCA algorithm.

Supplementary figure 2: (**A**) The fraction of either mapped or unmapped cells expressing the given marker gene. The mapping was performed using the indicated cell population in the PBMC dataset. (**B**) Reference graph of the 33K data set, with cell labels obtained post clustering using Seurat. Cell labels were placed based on the expression of canonical marker genes. (**C**) Fraction of mapping scores that was either within or outside the 33K dataset Seurat clusters upon mapping of each corresponding 6K cluster. (**D**) A t-SNE visualization of the 6K PBMC dataset showing the cells (in red) whose cluster identity was differently predicted by Nabo and Seurat.

Supplementary figure 3: (**A**) A t-SNE plot of normal GMLPs obtained using Seurat. The cells have been colored based on cluster association, which is also indicated by numerals. (**B**) The SNN graph (from Figure 4A) of normal GMLPs, using the same cluster identities as shown in A. (**C**) Barplots showing the number of MLL-ENL cells that were assigned to each cluster obtained from Seurat. The minimum weight for assignment using Nabo was set at either 0 or 0.1 (default). In case of scmap-cluster the weight fraction was set either at 0.7 (default) or 0.9.

Supplementary figure 4: (**A**) SNN graph of normal GMLPs constructed using low variance genes sampled from the same expression range as the highly variable genes (HVGs). The cell positions have been set to be the same as in the SNN graph obtained using the HVGs (**figure 4A**). The cell size has been scaled to indicate the mapping score obtained when MLL-ENL induced GMLPs were projected on this graph. (**B**) The cluster-wise distribution of mapping scores shown in A. The same cluster identities were used in the HVG graph. (**C**) The number of MLL-ENL cells assigned to each of the cluster. (**D**) SNN graph of WT GMLPs following blocked mapping. (**E**) The cluster-wise distribution of mapping scores shown in D. The same cluster identities were used in the HVG graph. (**F**) The number of MLL-ENL cells assigned to each of the cluster after the projection shown in D.

Supplementary figure 5: (**A**) Force directed layout of SNN graph of normal GMLPs along with the force directed layout of MLL-ENL cells. The reference cells are in gray and the MLL-ENL cells are colored based on their mapping specificity. The mapping specificity is indicated in dark purple (highest) and light yellow (least specificity). (**B**)

The average mapping specificity of each reference cell based on the specificity of each MLL-ENL cell that mapped onto those reference cells. Smaller and lighter colored cells have higher mapping specificies and those in large size and dark green color have lower specificity. The cells shown in grey were not mapped to.

Supplementary figure 6: (**A**) Radar plots showing the predicted cell type for each of the normal GMLP cluster. The three area polygons in the radar plot show the mean and +/- 1 SD values.

Supplementary figure 7: (**A**) Force directed SNN graph of normal GMLPs. The red dots represent cells with mapping scores higher than 1 and are from cluster 5. The blue dots represent cells which are at a path distance of 2 from red cells. The red cells were compared with blue cells to identify differentially expressed genes associated with mapped cells. (**B**) The expression of the top 30 most differentially expressed genes. The cells marked in dark red have higher expression and the ones marked in yellow have the less expression. (**C**) Notched boxplots showing the distribution of selected genes in the three groups.

Supplementary figure 8: Radar plots showing the gene expression-based lineage affiliation of cells. Genes used to generate each plot were those differentially expressed (upregulated) obtained by comparing: (**B**) Normal GMLPs from cluster 5 that received mapping scores over 1 compared to other cells at path distance of 2. (**B**) WT GMLPs from cluster 5 that received mapping scores over 1 compared to MLL-ENL cells that mapped to these cells. (**C**) The opposite approach of B. (**D**) All MLL-ENL cells compared to all normal GMLPs.

Supplementary figure 9: (**A**) SNN graph of c-kit+ cells, where cells have been colored based on the cluster identity. (**B**) Radar plots showing the lineage bias of each cluster, inferred from gene expression signature of each cluster.

ACKNOWLEDGMENTS

We thank Mikael Sommarin, Johan Rodhe and Oscar Legetth at Lund Stem Cell Center for their help with code and documentation review. This work was supported by grants from the Swedish Cancer Society, The Ragnar Söderberg Foundation, the Knut and Alice Wallenberg Foundation, the Swedish Research Council, the Swedish Society for Medical Research, and the Swedish Childhood Cancer Foundation.

AUTHOR CONTRIBUTIONS

G.K, D.B, and P.D conceived and designed the study; D.B, A.U, M.E, and E.E designed and performed mouse experiments and sequencing; P.D and S.S designed and performed the bioinformatics- and computational analyses; P.D and R.O packaged Nabo codebase and prepared documentation; G.K, D.B, and P.D analyzed and interpreted data; G.K and P.D prepared the figures and wrote the manuscript with input from all authors.

REFERENCES

Achim, K., Pettit, J.-B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., Marioni, J.C., 2015. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat. Biotechnol. 33, 503–509. https://doi.org/10.1038/nbt.3209

Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J., Clarke, M.F., 2003. Prospective identification of tumorigenic breast cancer cells. Proc. Natl. Acad. Sci. U.S.A. 100, 3983–3988. https://doi.org/10.1073/pnas.0530291100

Andrews, T.S., Hemberg, M., 2018. Identifying cell populations with scRNASeq. Mol. Aspects Med. 59, 114–122. https://doi.org/10.1016/j.mam.2017.07.002

Bagger, F.O., Sasivarevic, D., Sohi, S.H., Laursen, L.G., Pundhir, S., Sønderby, C.K., Winther, O., Rapin, N., Porse, B.T., 2016. BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. Nucleic Acids Res. 44, D917-924. https://doi.org/10.1093/nar/gkv1101

Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M., Melton, D.A., Yanai, I., 2016. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst 3, 346-360.e4. https://doi.org/10.1016/j.cels.2016.08.011

Blakeley, P., Fogarty, N.M.E., del Valle, I., Wamaitha, S.E., Hu, T.X., Elder, K., Snell, P., Christie, L., Robson, P., Niakan, K.K., 2015. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development 142, 3151–3165. https://doi.org/10.1242/dev.123547

Boiko, A.D., Razorenova, O.V., van de Rijn, M., Swetter, S.M., Johnson, D.L., Ly, D.P., Butler, P.D., Yang, G.P., Joshua, B., Kaplan, M.J., Longaker, M.T., Weissman, I.L., 2010. Human melanoma-initiating cells express neural crest nerve growth factor receptor CD271. Nature 466, 133–137. https://doi.org/10.1038/nature09161

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

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R., 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. 36, 411–420. https://doi.org/10.1038/nbt.4096

Clevers, H., 2011. The cancer stem cell: premises, promises and challenges. Nat. Med. 17, 313–319. https://doi.org/10.1038/nm.2304

Cozzio, A., Passegué, E., Ayton, P.M., Karsunky, H., Cleary, M.L., Weissman, I.L., 2003. Similar MLL-associated leukemias arising from self-renewing stem cells and short-lived myeloid progenitors. Genes Dev. 17, 3029–3035. https://doi.org/10.1101/gad.1143403

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T.M., Lander, E.S., Weissman, J.S., Friedman, N., Regev, A., 2016. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell 167, 1853-1866.e17. https://doi.org/10.1016/j.cell.2016.11.038

Eppert, K., Takenaka, K., Lechman, E.R., Waldron, L., Nilsson, B., van Galen, P., Metzeler, K.H., Poeppl, A., Ling, V., Beyene, J., Canty, A.J., Danska, J.S., Bohlander, S.K., Buske, C., Minden, M.D., Golub, T.R., Jurisica, I., Ebert, B.L., Dick, J.E., 2011. Stem cell gene expression programs influence clinical outcome in human leukemia. Nat. Med. 17, 1086–1093. https://doi.org/10.1038/nm.2415

Gierahn, T.M., Wadsworth, M.H., Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., Shalek, A.K., 2017. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat. Methods 14, 395–398. https://doi.org/10.1038/nmeth.4179

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., van Oudenaarden, A., 2015. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 525, 251–255. https://doi.org/10.1038/nature14966

Haghverdi, L., Lun, A.T.L., Morgan, M.D., Marioni, J.C., 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. 36, 421–427. https://doi.org/10.1038/nbt.4091

Hashimshony, T., Wagner, F., Sher, N., Yanai, I., 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep 2, 666–673. https://doi.org/10.1016/j.celrep.2012.08.003

Herring, C.A., Chen, B., McKinley, E.T., Lau, K.S., 2018. Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems. Cell Mol Gastroenterol Hepatol 5, 539–548. https://doi.org/10.1016/j.jcmgh.2018.01.023

Heuser, M., Yun, H., Berg, T., Yung, E., Argiropoulos, B., Kuchenbauer, F., Park, G., Hamwi, I., Palmqvist, L., Lai, C.K., Leung, M., Lin, G., Chaturvedi, A., Thakur, B.K., Iwasaki, M., Bilenky, M., Thiessen, N., Robertson, G., Hirst, M., Kent, D., Wilson, N.K., Göttgens, B., Eaves, C., Cleary, M.L., Marra, M., Ganser, A., Humphries, R.K., 2011. Cell of origin in AML: susceptibility to MN1-induced transformation is regulated by the MEIS1/AbdB-like HOX protein complex. Cancer Cell 20, 39–52. https://doi.org/10.1016/j.ccr.2011.06.020

Huntly, B.J.P., Shigematsu, H., Deguchi, K., Lee, B.H., Mizuno, S., Duclos, N., Rowan, R., Amaral, S., Curley, D., Williams, I.R., Akashi, K., Gilliland, D.G., 2004. MOZ-TIF2, but not BCR-ABL, confers properties of leukemic stem cells to committed murine hematopoietic progenitors. Cancer Cell 6, 587–596. https://doi.org/10.1016/j.ccr.2004.10.015

Jaakkola, M.K., Seyednasrollah, F., Mehmood, A., Elo, L.L., 2017. Comparison of methods to detect differentially expressed genes between single-cell populations. Brief. Bioinformatics 18, 735–743. https://doi.org/10.1093/bib/bbw057

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., Amit, I., 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343, 776–779. https://doi.org/10.1126/science.1247651

Jarvis, R.A., Patrick, E.A., 1973. Clustering Using a Similarity Measure Based on Shared Near Neighbors. IEEE Transactions on Computers C–22, 1025–1034. https://doi.org/10.1109/T-C.1973.223640

Kiselev, V.Y., Yiu, A., Hemberg, M., 2018. scmap: projection of single-cell RNA-seq data across data sets. Nat. Methods 15, 359–362. https://doi.org/10.1038/nmeth.4644

Krivtsov, A.V., Twomey, D., Feng, Z., Stubbs, M.C., Wang, Y., Faber, J., Levine, J.E., Wang, J., Hahn, W.C., Gilliland, D.G., Golub, T.R., Armstrong, S.A., 2006. Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. Nature 442, 818–822. https://doi.org/10.1038/nature04980

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A., McCarroll, S.A., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., Diamanti, E., Nishikawa, S.-I., Piterman, N., Kouskoff, V., Theis, F.J., Fisher, J., Göttgens, B., 2015. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat. Biotechnol. 33, 269–276. https://doi.org/10.1038/nbt.3154

Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J.P., van Oudenaarden, A., 2016. A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst 3, 385-394.e3. https://doi.org/10.1016/j.cels.2016.09.002

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F.K.B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B.T., Tanay, A., Amit, I., 2015. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 163, 1663–1677. https://doi.org/10.1016/j.cell.2015.11.013

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., Sandberg, R., 2014. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc 9, 171–181. https://doi.org/10.1038/nprot.2014.006

Quintana, E., Shackleton, M., Sabel, M.S., Fullen, D.R., Johnson, T.M., Morrison, S.J., 2008. Efficient tumour formation by single human melanoma cells. Nature 456, 593–598. https://doi.org/10.1038/nature07567

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., Schroth, G.P., Sandberg, R., 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. 30, 777–782. https://doi.org/10.1038/nbt.2282

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A., 2015. Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. 33, 495–502. https://doi.org/10.1038/nbt.3192

Satijn, D.P., Hamer, K.M., den Blaauwen, J., Otte, A.P., 2001. The polycomb group protein EED interacts with YY1, and both proteins induce neural tissue in Xenopus embryos. Mol. Cell. Biol. 21, 1360–1369. https://doi.org/10.1128/MCB.21.4.1360-1369.2001

Säwen, P., Eldeeb, M., Erlandsson, E., Kristiansen, T.A., Laterza, C., Kokaia, Z., Karlsson, G., Yuan, J., Soneji, S., Mandal, P.K., Rossi, D.J., Bryder, D., 2018. Murine HSCs contribute actively to native hematopoiesis but with reduced differentiation capacity upon aging. Elife 7. https://doi.org/10.7554/eLife.41258

Schepers, A.G., Snippert, H.J., Stange, D.E., van den Born, M., van Es, J.H., van de Wetering, M., Clevers, H., 2012. Lineage tracing reveals Lgr5+ stem cell activity in mouse intestinal adenomas. Science 337, 730–735. https://doi.org/10.1126/science.1224676

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., Smith, D.M., Kasper, M., Ämmälä, C., Sandberg, R., 2016. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab. 24, 593–607. https://doi.org/10.1016/j.cmet.2016.08.020

Shin, J., Berg, D.A., Zhu, Y., Shin, J.Y., Song, J., Bonaguidi, M.A., Enikolopov, G., Nauen, D.W., Christian, K.M., Ming, G., Song, H., 2015. Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. Cell Stem Cell 17, 360–372. https://doi.org/10.1016/j.stem.2015.07.013

Singh, S.K., Hawkins, C., Clarke, I.D., Squire, J.A., Bayani, J., Hide, T., Henkelman, R.M., Cusimano, M.D., Dirks, P.B., 2004. Identification of human brain tumour initiating cells. Nature 432, 396–401. https://doi.org/10.1038/nature03128

Somervaille, T.C.P., Cleary, M.L., 2006. Identification and characterization of leukemia stem cells in murine MLL-AF9 acute myeloid leukemia. Cancer Cell 10, 257–268. https://doi.org/10.1016/j.ccr.2006.08.020

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A.S., Hughes, T.K., Ziegler, C.G.K., Kazer, S.W., Gaillard, A., Kolb, K.E., Villani, A.-C., Johannessen, C.M., Andreev, A.Y., Van Allen, E.M., Bertagnolli, M., Sorger, P.K., Sullivan, R.J., Flaherty, K.T., Frederick, D.T., Jané-Valbuena, J., Yoon, C.H., Rozenblatt-Rosen, O., Shalek, A.K., Regev, A., Garraway, L.A., 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189–196. https://doi.org/10.1126/science.aad0501

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., Rinn, J.L., 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386. https://doi.org/10.1038/nbt.2859

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., Quake, S.R., 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375. https://doi.org/10.1038/nature13173

Treutlein, B., Lee, Q.Y., Camp, J.G., Mall, M., Koh, W., Shariati, S.A.M., Sim, S., Neff, N.F., Skotheim, J.M., Wernig, M., Quake, S.R., 2016. Dissecting direct reprogramming

from fibroblast to neuron using single-cell RNA-seq. Nature 534, 391–395. https://doi.org/10.1038/nature18323

Ugale, A., Norddahl, G.L., Wahlestedt, M., Säwén, P., Jaako, P., Pronk, C.J., Soneji, S., Cammenga, J., Bryder, D., 2014. Hematopoietic stem cells are intrinsically protected against MLL-ENL-mediated transformation. Cell Rep 9, 1246–1255. https://doi.org/10.1016/j.celrep.2014.10.036

Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., Marioni, J.C., 2017. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat. Methods 14, 565–571. https://doi.org/10.1038/nmeth.4292

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P.L., Rozenblatt-Rosen, O., Lane, A.A., Haniffa, M., Regev, A., Hacohen, N., 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 356. https://doi.org/10.1126/science.aah4573

Wang, J.C.Y., Dick, J.E., 2005. Cancer stem cells: lessons from leukemia. Trends Cell Biol. 15, 494–501. https://doi.org/10.1016/j.tcb.2005.07.004

Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M., Klein, A.M., 2018. Fundamental limits on dynamic inference from single-cell snapshots. Proc. Natl. Acad. Sci. U.S.A. 115, E2467–E2476. https://doi.org/10.1073/pnas.1714723115

Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., Guo, Y.E., Hnisz, D., Jaenisch, R., Bradner, J.E., Gray, N.S., Young, R.A., 2017. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell 171, 1573-1588.e28. https://doi.org/10.1016/j.cell.2017.11.008

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Murphy, A.J., Yancopoulos, G.D., Lin, C., Gromada, J., 2016. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. Cell Metab. 24, 608–615. https://doi.org/10.1016/j.cmet.2016.08.018

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., Linnarsson, S., 2015. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142. https://doi.org/10.1126/science.aaa1934

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H., 2017. Massively parallel digital transcriptional profiling of single cells. Nat Commun 8, 14049. https://doi.org/10.1038/ncomms14049

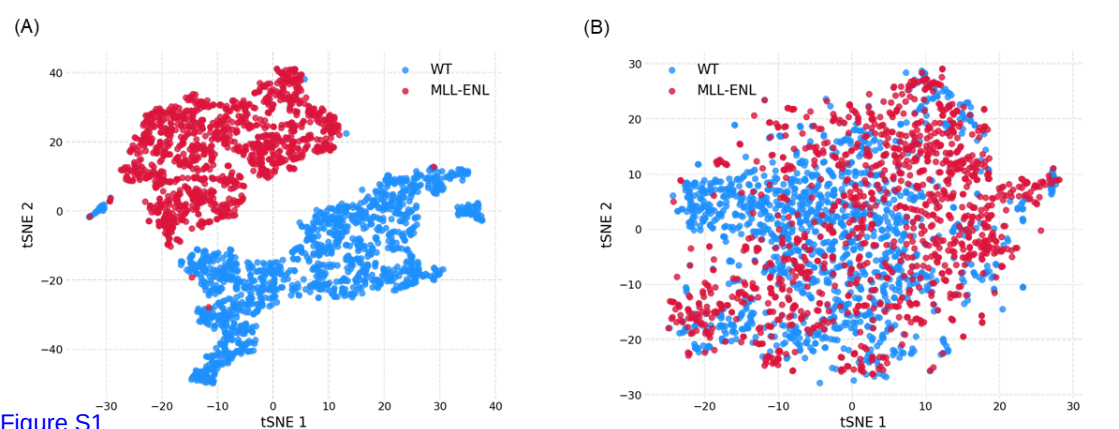

(A)
WT
MLL-ENL
tSNE 2
tSNE 1
40
20
0
−20
−40
−30
−20
−10
0
10
20
30
40
(B)
WT
MLL-ENL
tSNE 2
tSNE 1
30
20
10
0
−10
−20
−30
−20
−10
0
10
20
30

Figure S1

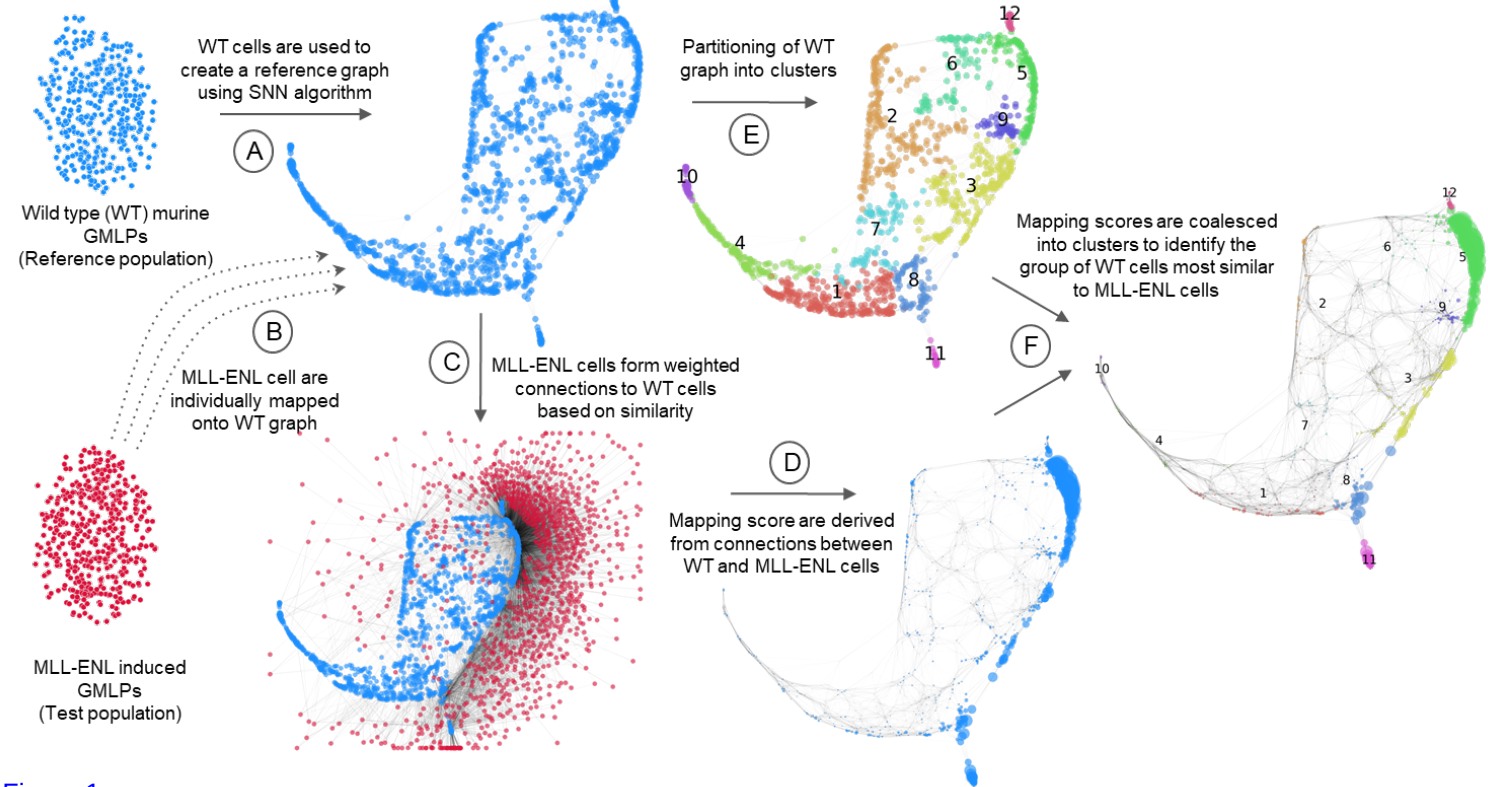

Wild type (WT) murine GMLPs (Reference population)
WT cells are used to create a reference graph using SNN algorithm
A
B
MLL-ENL cell are individually mapped onto WT graph
MLL-ENL induced GMLPs (Test population)
C
MLL-ENL cells form weighted connections to WT cells based on similarity
D
Mapping score are derived from connections between WT and MLL-ENL cells
E
Partitioning of WT graph into clusters
F
Mapping scores are coalesced into clusters to identify the group of WT cells most similar to MLL-ENL cells

Figure 1

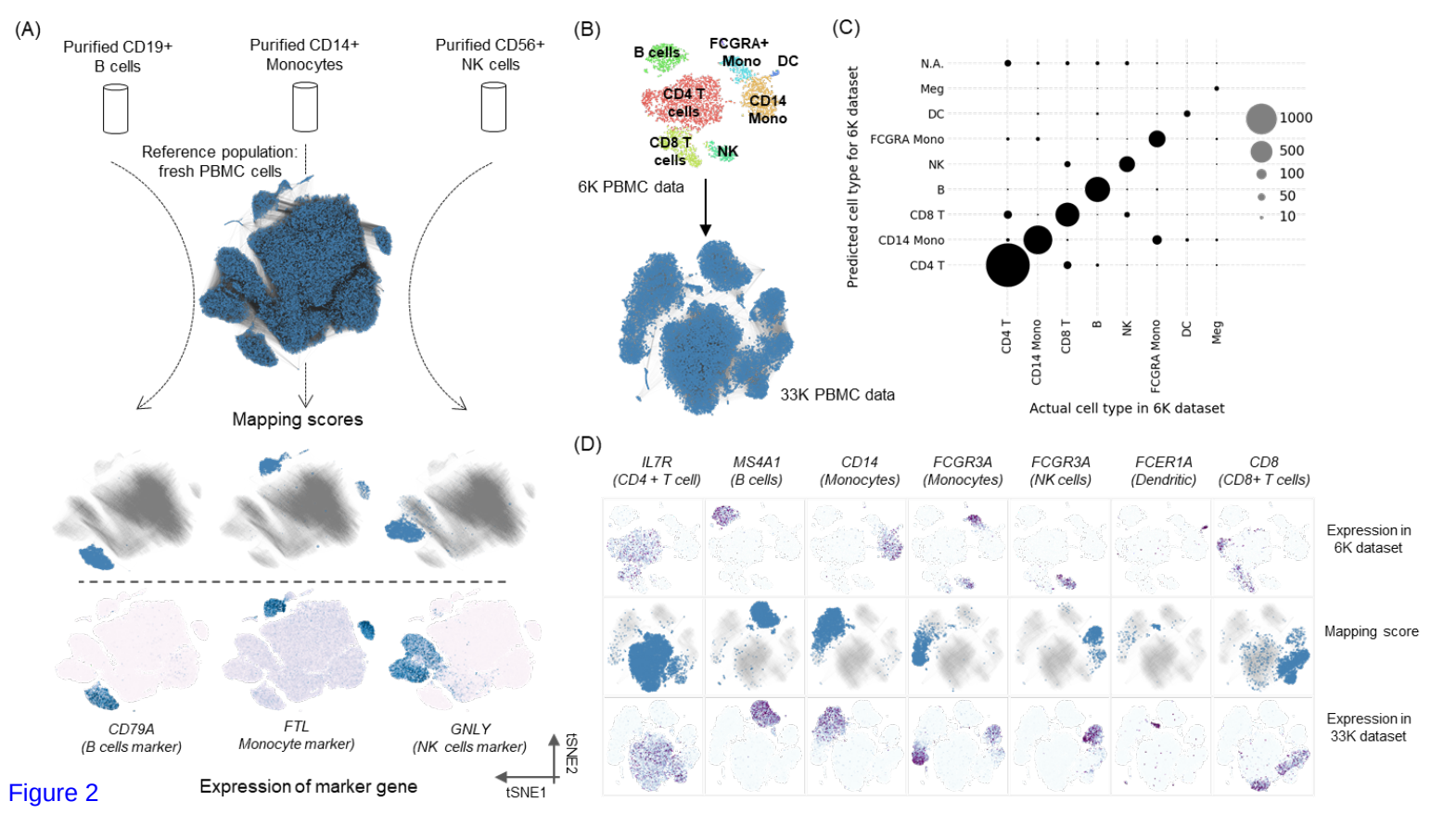

(A)
Purified CD19+ B cells
Purified CD14+ Monocytes
Purified CD56+ NK cells
Reference population: fresh PBMC cells
Mapping scores
CD79A (B cells marker)
FTL Monocyte marker)
GNLY (NK cells marker)
Expression of marker gene
tSNE2
tSNE1
(B)
B cells
FCGRA+ Mono
DC
CD4 T cells
CD14 Mono
CD8 T cells
NK
6K PBMC data
33K PBMC data
(C)
Predicted cell type for 6K dataset
N.A.
Meg
DC
FCGRA Mono
NK
B
CD8 T
CD14 Mono
CD4 T
1000
500
100
50
10
Actual cell type in 6K dataset
(D)
IL7R (CD4 + T cell)
MS4A1 (B cells)
CD14 (Monocytes)
FCGR3A (Monocytes)
FCGR3A (NK cells)
FCER1A (Dendritic)
CD8 (CD8+ T cells)
Expression in 6K dataset
Mapping score
Expression in 33K dataset

Figure 2

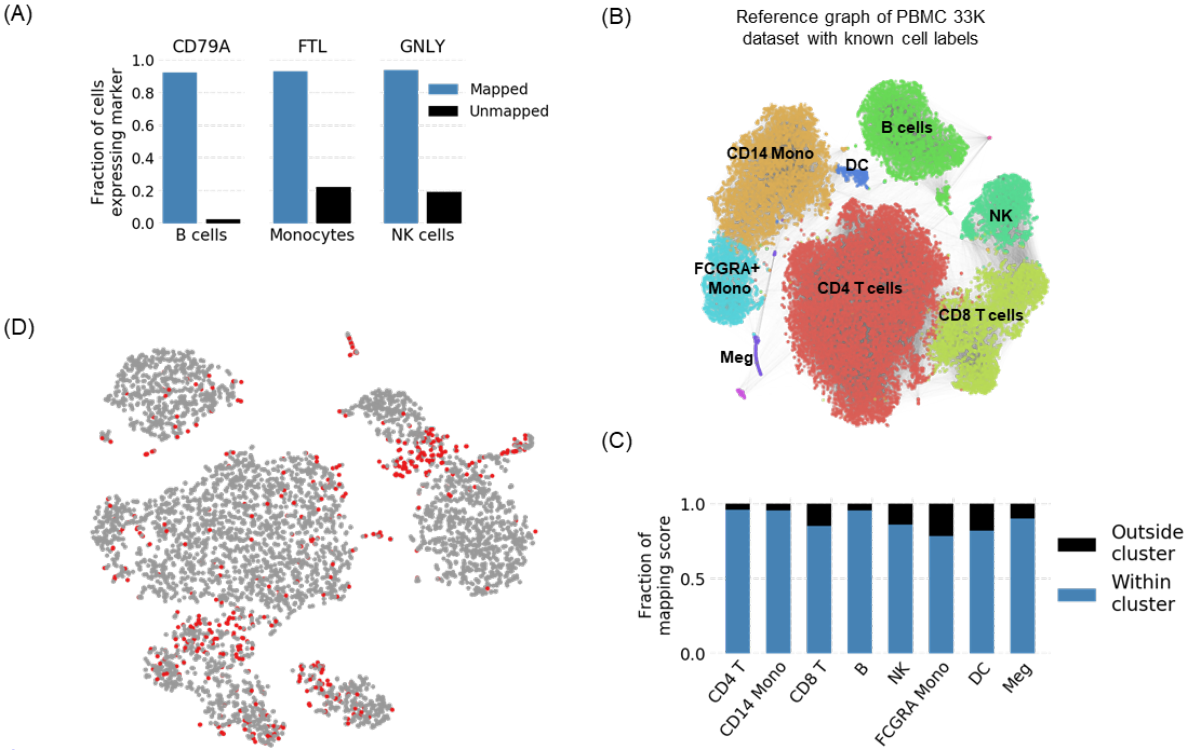

(A)
CD79A
FTL
GNLY
Fraction of cells expressing marker
1.0
0.8
0.6
0.4
0.2
0.0
Mapped
Unmapped
B cells
Monocytes
NK cells
(B)
Reference graph of PBMC 33K dataset with known cell labels
B cells
CD14 Mono
DC
NK
FCGRA+ Mono
CD4 T cells
CD8 T cells
Meg
(C)
Fraction of mapping score
1.0
0.5
0.0
Outside cluster
Within cluster
CD4 T
CD14 Mono
CD8 T
B
NK
FCGRA Mono
DC
Meg
(D)

Figure S2

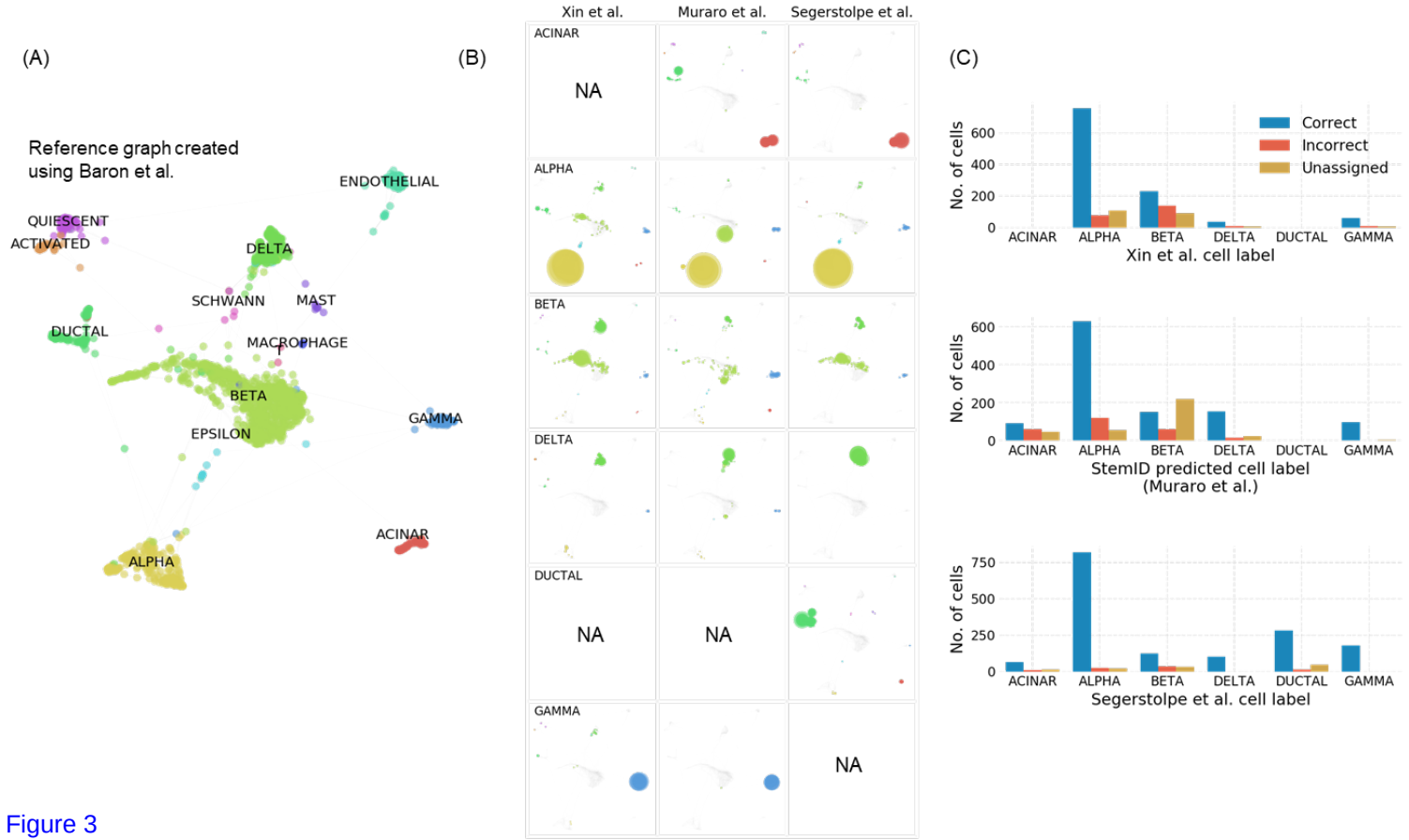

(A)
Reference graph created using Baron et al.
ENDOTHELIAL
QUIESCENT
ACTIVATED
DELTA
SCHWANN
MAST
DUCTAL
MACROPHAGE
BETA
GAMMA
EPSILON
ACINAR
ALPHA
(B)
Xin et al.
Muraro et al.
Segerstolpe et al.
ACINAR
NA
ALPHA
BETA
DELTA
DUCTAL
NA
NA
GAMMA
NA
(C)
No. of cells
Correct
Incorrect
Unassigned
ACINAR
ALPHA
BETA
DELTA
DUCTAL
GAMMA
Xin et al. cell label
StemID predicted cell label (Muraro et al.)
Segerstolpe et al. cell label

Figure 3

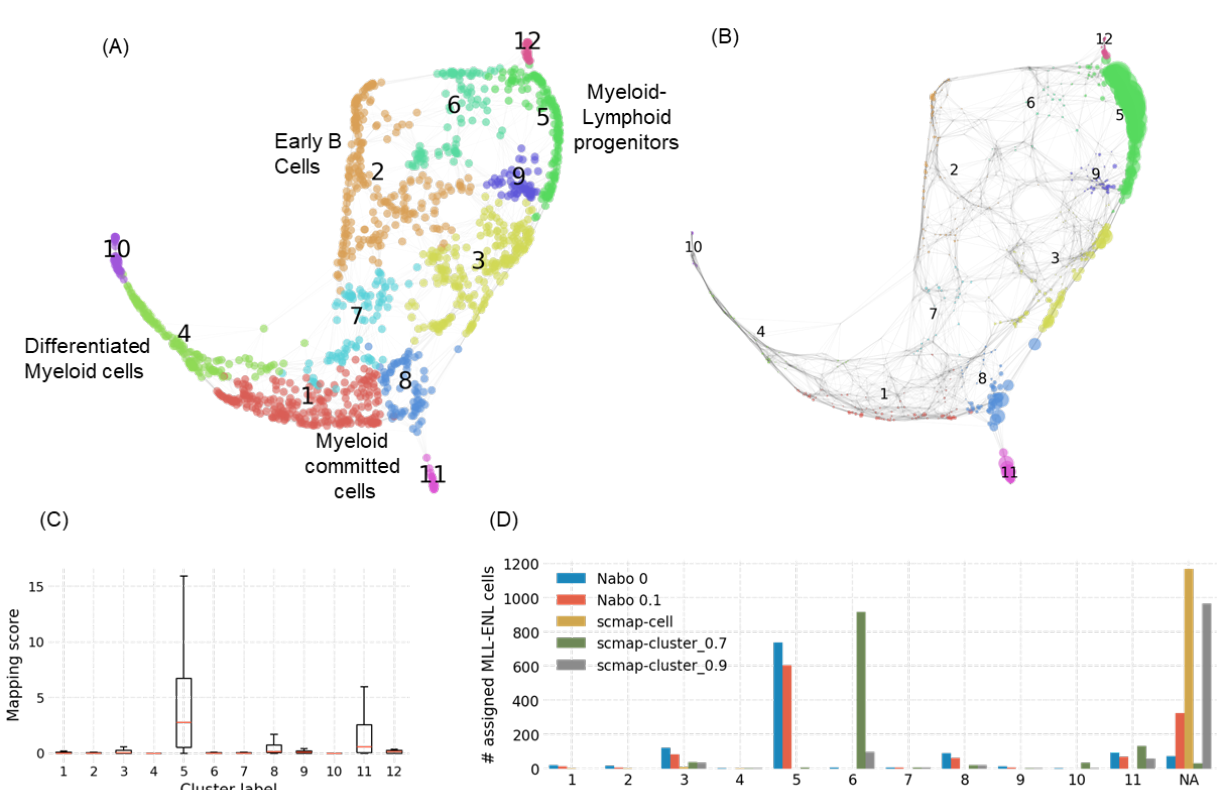

(A)
(B)
(C)
(D)
Early B Cells
Myeloid-Lymphoid progenitors
Differentiated Myeloid cells
Myeloid committed cells
Mapping score
Cluster label
assigned MLL-ENL cells
Cluster number
Nabo 0
Nabo 0.1
scmap-cell
scmap-cluster_0.7
scmap-cluster_0.9

Figure 4

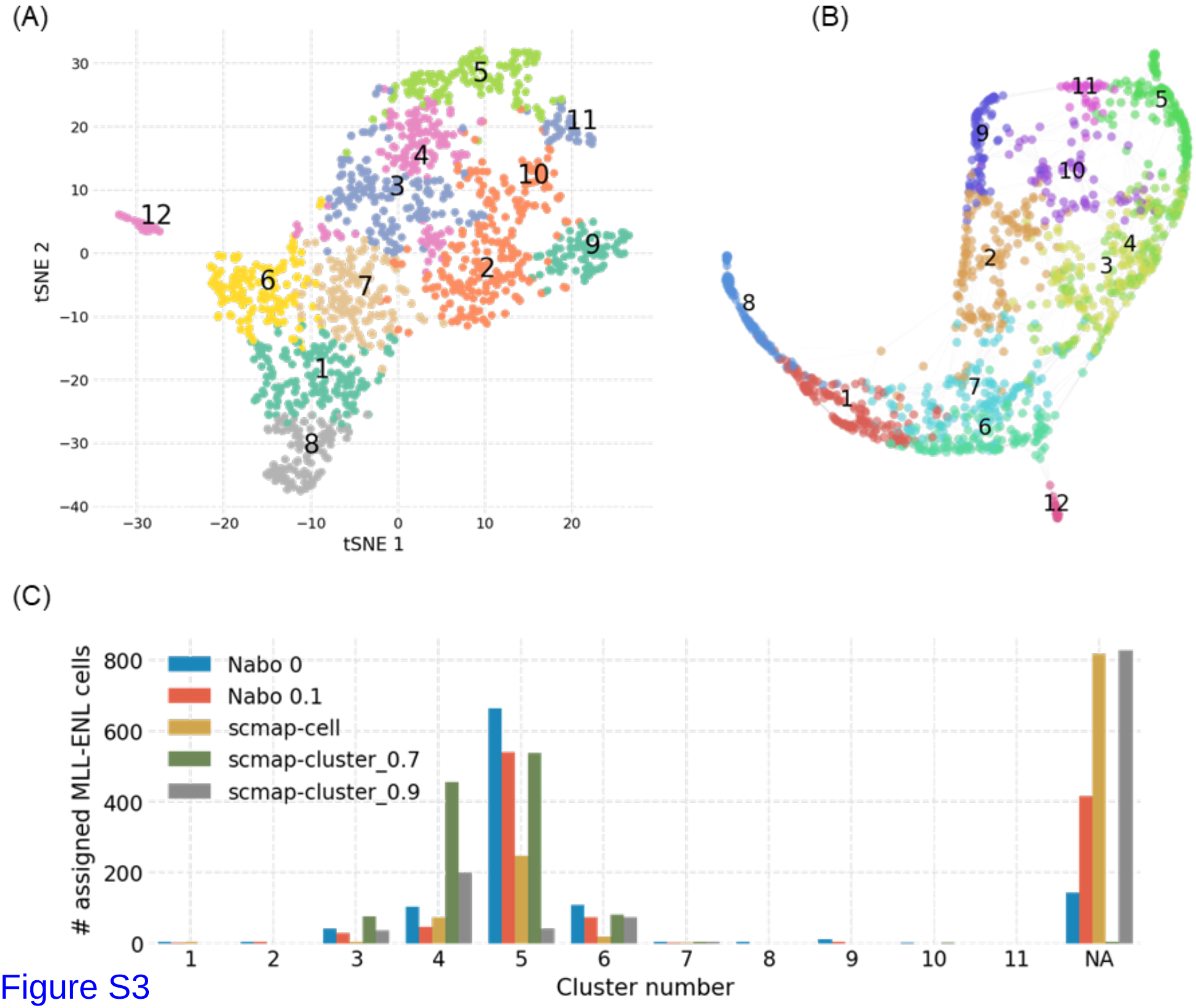

(A)
(B)
(C)
tSNE 1
tSNE 2
assigned MLL-ENL cells
Cluster number
Nabo 0
Nabo 0.1
scmap-cell
scmap-cluster_0.7
scmap-cluster_0.9
NA

Figure S3

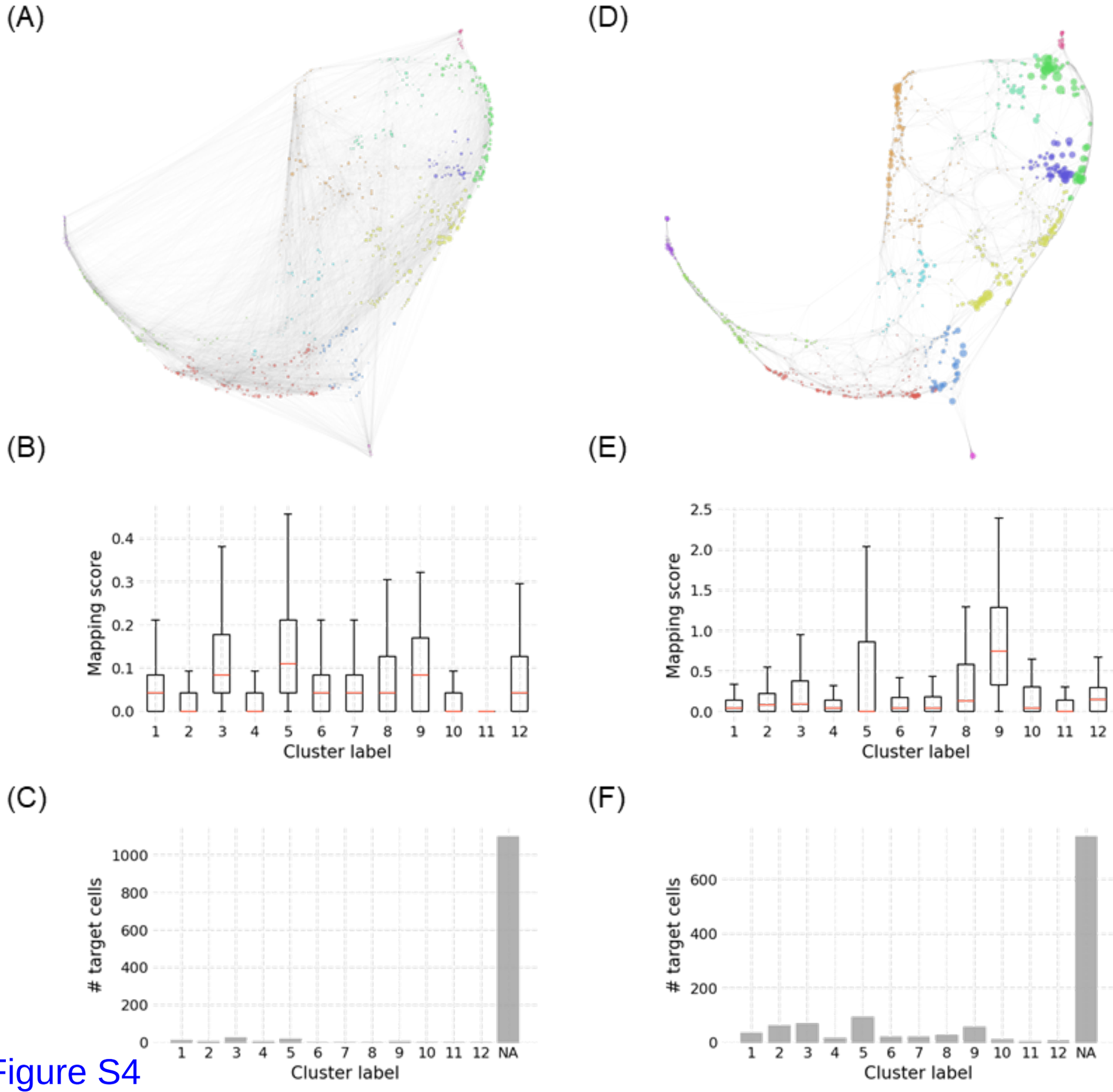

(A)
(D)
(B)
Mapping score
0.0
0.1
0.2
0.3
0.4
1 2 3 4 5 6 7 8 9 10 11 12
Cluster label
(E)
Mapping score
0.0
0.5
1.0
1.5
2.0
2.5
1 2 3 4 5 6 7 8 9 10 11 12
Cluster label
(C)
target cells
0
200
400
600
800
1000
1 2 3 4 5 6 7 8 9 10 11 12 NA
Cluster label
(F)
target cells
0
200
400
600
1 2 3 4 5 6 7 8 9 10 11 12 NA
Cluster label

Figure S4

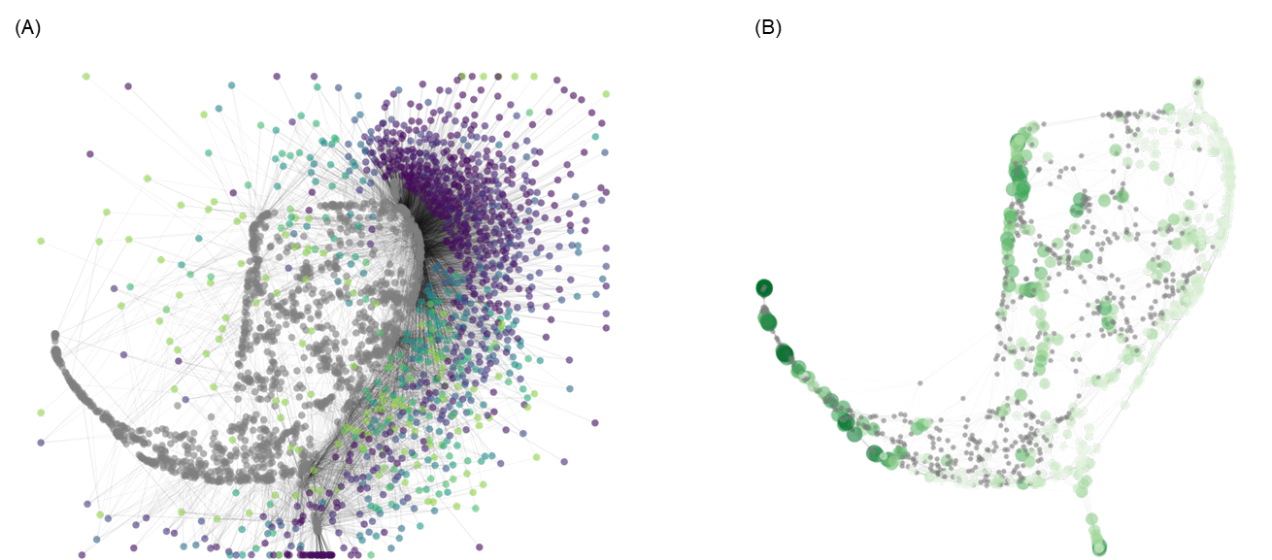
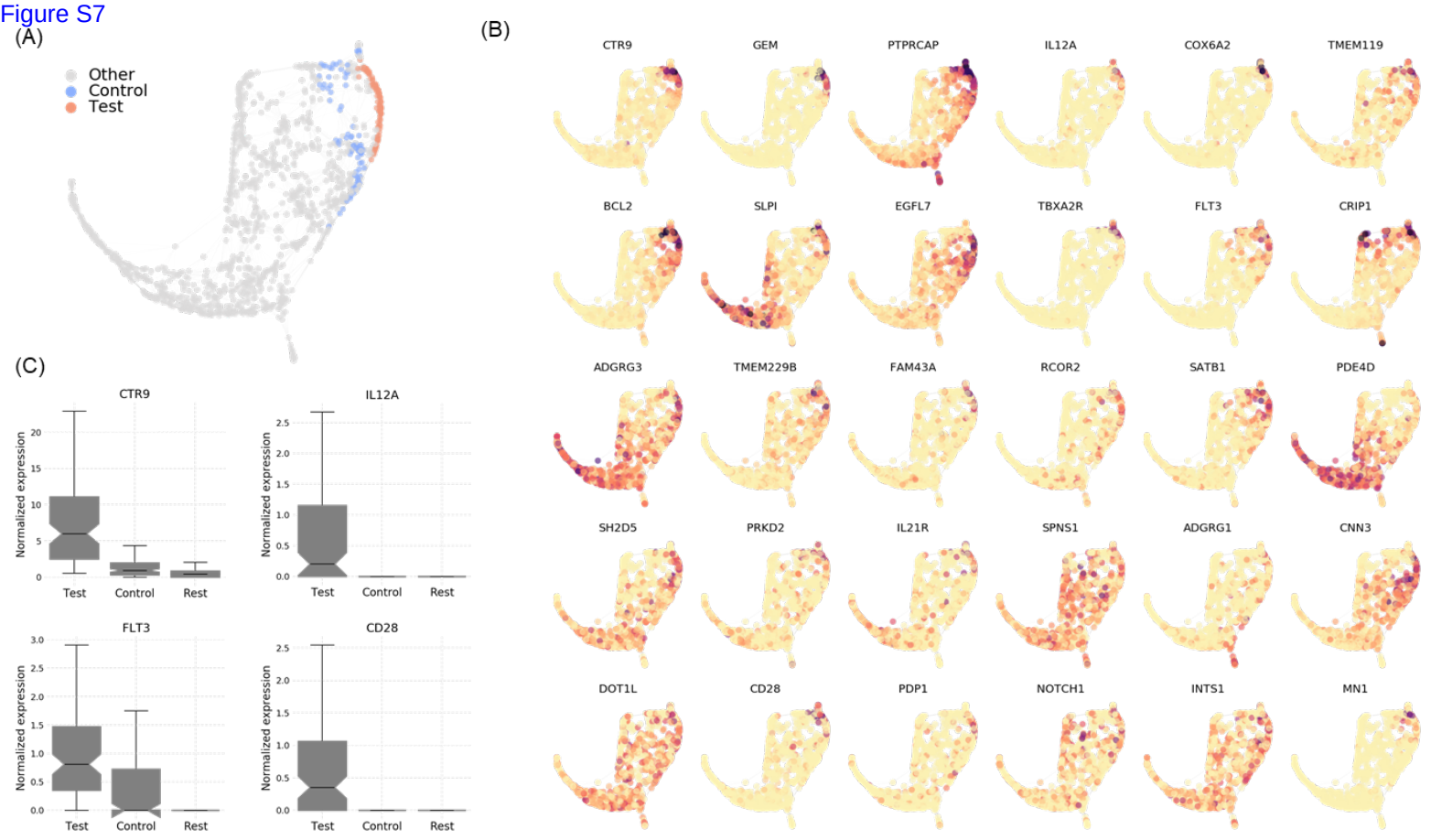


Figure S5


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Monocyte
Granulocytes
GMP
preGM
LMPP
ST HSC
LT HSC
Nucleated Erythrocytes
ProE
CFUE
PreCFUE
MkE
MkP
CD8 activated
CD8 naive
CD4 activated
NKmature
NK cells
CD4
B Cell
ProB
PreB
ETP
CLP

Figure S6

Figure S7


(A)
Other
Control
Test
(B)
CTR9
GEM
PTPRCAP
IL12A
COX6A2
TMEM119
BCL2
SLPI
EGFL7
TBXA2R
FLT3
CRIP1
ADGRG3
TMEM229B
FAM43A
RCOR2
SATB1
PDE4D
SH2D5
PRKD2
IL21R
SPNS1
ADGRG1
CNN3
DOT1L
CD28
PDP1
NOTCH1
INTS1
MN1
(C)
CTR9
IL12A
FLT3
CD28
Normalized expression
Test
Control
Rest

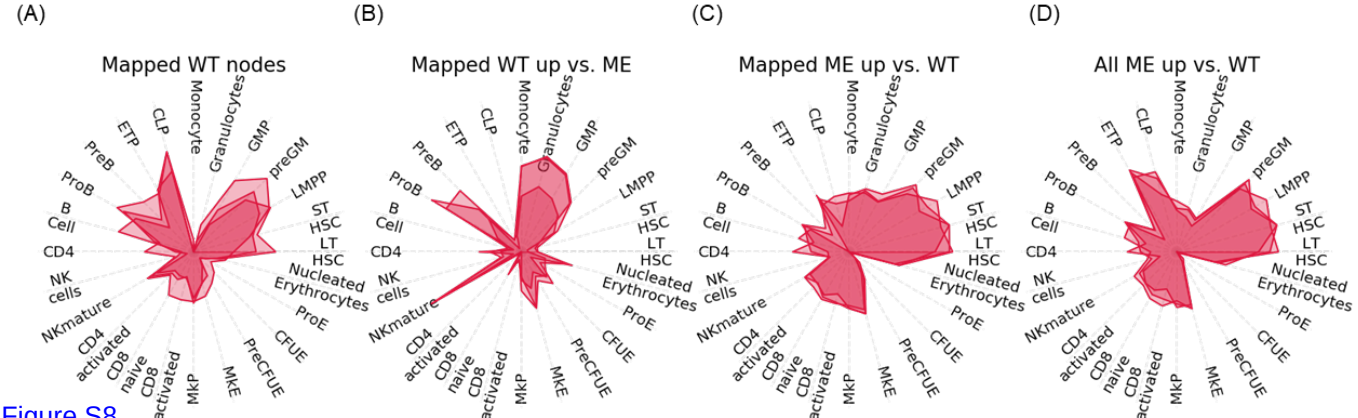

(A)
Mapped WT nodes
(B)
Mapped WT up vs. ME
(C)
Mapped ME up vs. WT
(D)
All ME up vs. WT
Monocyte
Granulocytes
GMP
preGM
LMPP
ST HSC
LT HSC
Nucleated Erythrocytes
ProE
CFUE
PreCFUE
MkE
MkP
CD8 activated
CD8 naive
CD4 activated
NKmature
NK cells
CD4
B Cell
ProB
PreB
ETP
CLP

Figure S8

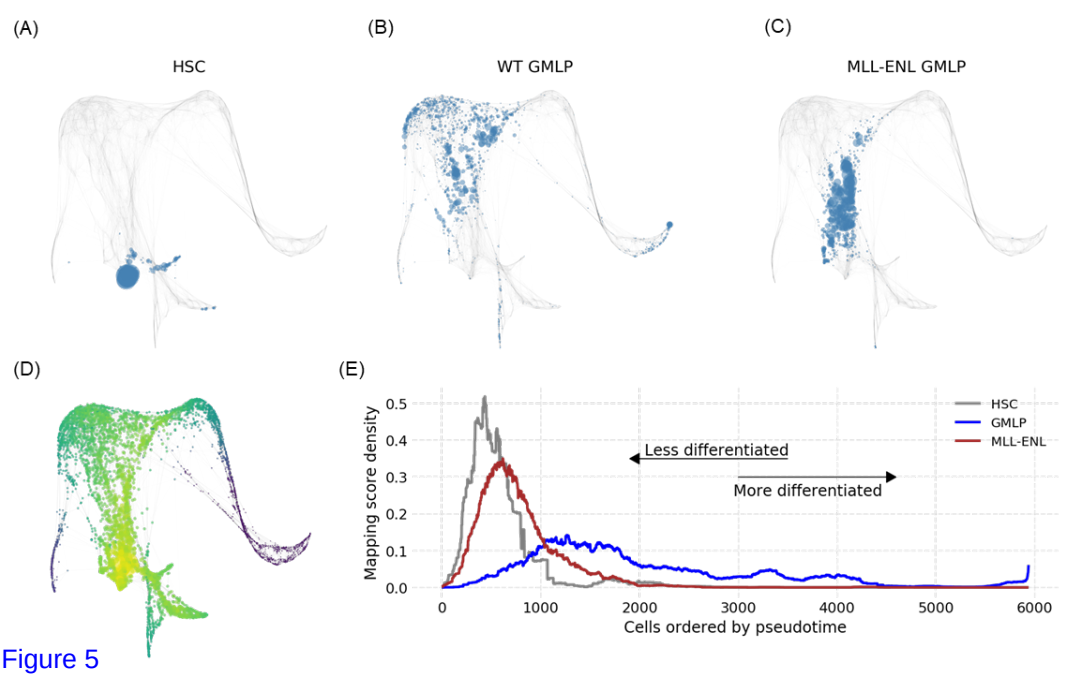

(A)
HSC
(B)
WT GMLP
(C)
MLL-ENL GMLP
(D)
(E)
Mapping score density
0.0
0.1
0.2
0.3
0.4
0.5
0
1000
2000
3000
4000
5000
6000
Cells ordered by pseudotime
HSC
GMLP
MLL-ENL
Less differentiated
More differentiated

Figure 5

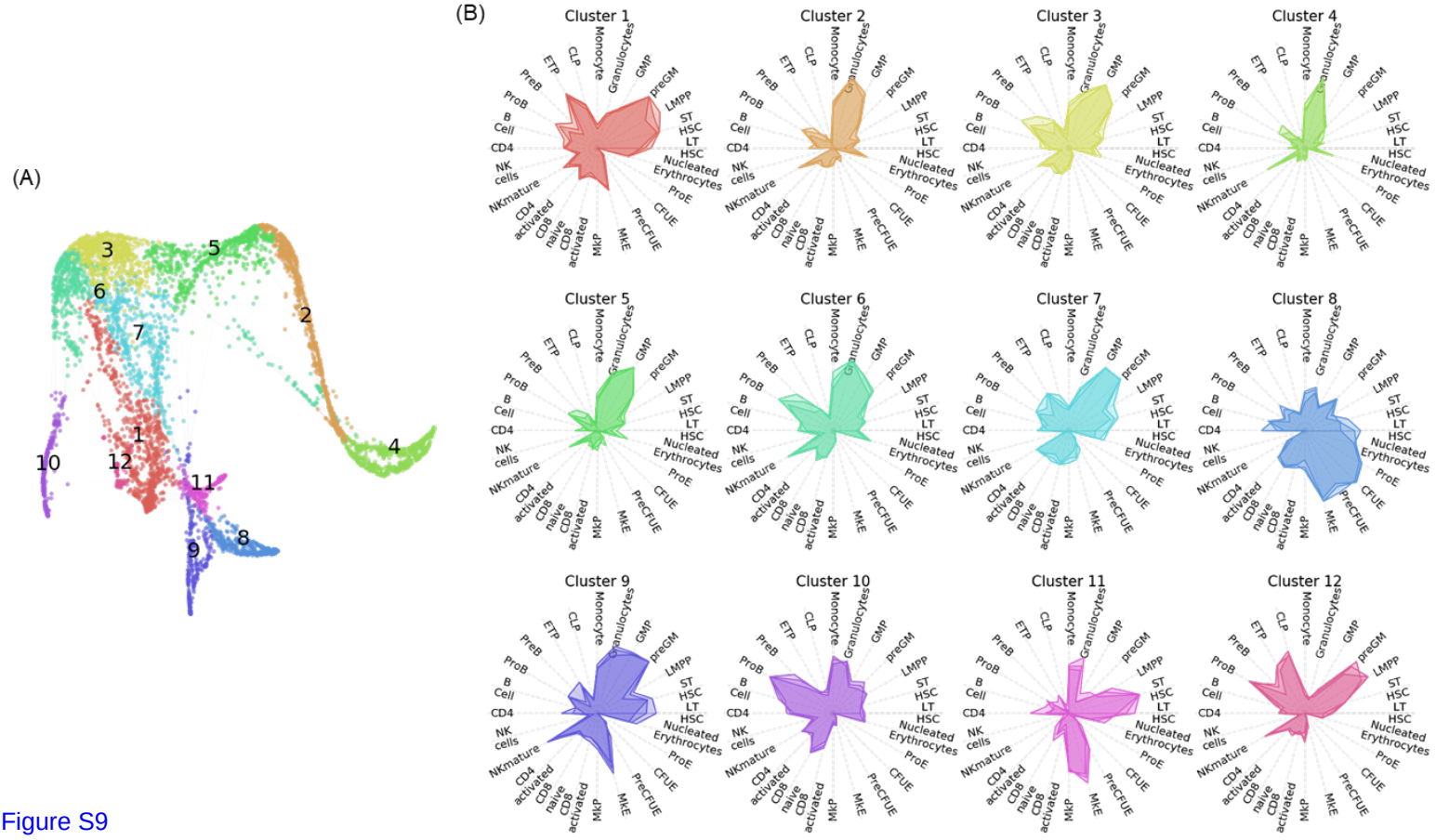
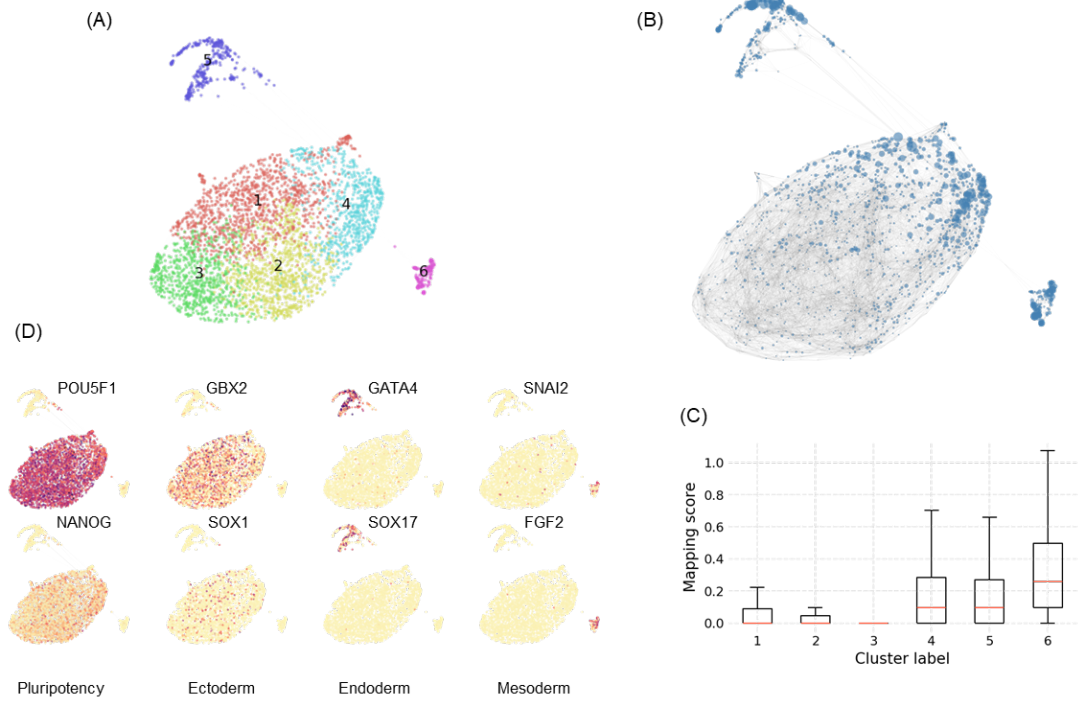

(A)
1
2
3
4
5
6
7
8
9
10
11
12
(B)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Monocyte
Granulocytes
GMP
preGM
LMPP
ST HSC
LT HSC
Nucleated Erythrocytes
ProE
CFUE
PreCFUE
MkE
MkP
CD8 activated
CD8 naive
CD4 activated
NKmature
NK cells
CD4
B Cell
ProB
PreB
ETP
CLP

Figure S9

Figure 6


(A)
5
1
4
3
2
6
(B)
(C)
Mapping score
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
Cluster label
(D)
POU5F1
GBX2
GATA4
SNAI2
NANOG
SOX1
SOX17
FGF2
Pluripotency
Ectoderm
Endoderm
Mesoderm