nature biotechnology

Wishbone identifies bifurcating developmental trajectories from single-cell data

Manu Setty[1,6], Michelle D Tadmor[1,6], Shlomit Reich-Zeliger[2], Omer Angel[3], Tomer Meir Salame[4], Pooja Kathail[1], Kristy Choi[1], Sean Bendall[5], Nir Friedman[2] & Dana Pe'er[1]

Recent single-cell analysis technologies offer an unprecedented opportunity to elucidate developmental pathways. Here we present Wishbone, an algorithm for positioning single cells along bifurcating developmental trajectories with high resolution. Wishbone uses multi-dimensional single-cell data, such as mass cytometry or RNA-Seq data, as input and orders cells according to their developmental progression, and it pinpoints bifurcation points by labeling each cell as pre-bifurcation or as one of two post-bifurcation cell fates. Using 30-channel mass cytometry data, we show that Wishbone accurately recovers the known stages of T-cell development in the mouse thymus, including the bifurcation point. We also apply the algorithm to mouse myeloid differentiation and demonstrate its generalization to additional lineages. A comparison of Wishbone to diffusion maps, SCUBA and Monocle shows that it outperforms these methods both in the accuracy of ordering cells and in the correct identification of branch points.

Multi-cellular organisms develop from a single cell that undergoes many stages of proliferation and differentiation, resulting in a vast array of progenitor and terminal cell types. Although many of the key stages and cell populations in these processes have been characterized using fluorescence-activated cell sorting and genetic perturbations, much of development remains uncharted. Emerging high-throughput technologies such as single-cell RNA-Seq[1] and mass cytometry[2] can measure a large number of parameters simultaneously in single cells and interrogate an entire tissue without perturbation. As many tissues maintain homeostasis through continuous and asynchronous development, this presents an opportunity to measure cells at almost all stages of maturity at high resolution. The challenge is to devise computational algorithms capable of exploiting this resolution to order cells based on their maturity and to identify the branch points that give rise to the full complement of functionally distinct cells.

Recently, several reports have demonstrated approaches to order single cells based on their maturity[3,4]. However, these approaches assume non-branching trajectories and thus are poorly suited to model multiple cell fates. Key challenges to constructing branching trajectories are ordering cells on the basis of their developmental maturity, identifying the branch point, and associating the cells with their respective branches. Methods such as SCUBA[5] can identify branches in data, along with pseudo-temporal ordering of cells, but with considerable loss in temporal resolution and accuracy.

Here we present Wishbone, a trajectory-detection algorithm for bifurcating systems. We use mass cytometry data measuring T-cell development in mouse thymus, where lymphoid progenitors differentiate to either $CD8^+$ cytotoxic or $CD4^+$ helper T cells, to demonstrate the accuracy and robustness of Wishbone. The Wishbone algorithm recovers the known stages in T-cell development with high accuracy and developmental resolution. We order double negative (DN) 1–4, double positive (DP), $CD4^+$ and $CD8^+$ cells from a single snapshot along a unified bifurcating trajectory. We show that Wishbone *de novo* recovers the known stages in T-cell development with increased accuracy and resolution compared with competing methods. The resulting trajectory and branches match the prevailing model of T-cell differentiation with the full complement of cell types.

We determine that a substantial part of heterogeneity in expression of developmental markers is explained by developmental maturity, rather than stochasticity in expression. Additionally, we apply Wishbone to early and late human myeloid differentiation data generated using mass cytometry[2] and mouse myeloid differentiation data generated using single-cell RNA-Seq[6]. Wishbone successfully identifies maturation and branch points in myeloid development *de novo*, demonstrating its broad applicability to systems with bifurcating trajectories across diverse single-cell technologies.

RESULT

Learning bifurcating developmental trajectories

To infer a branching trajectory directly from data, Wishbone makes the following assumptions about the data: (i) differentiation is a continuous process, (ii) a snapshot of primary tissue represents the entire differentiation process, and (iii) the trajectory of a cell bifurcates to one of two fates. Differentiation is punctuated by the rise and fall of phenotypic markers, and thus standard distance metrics such as Euclidean metrics do not adequately capture the difference in

[1]Department of Biological Sciences, Department of Systems Biology, Columbia University, New York, New York, USA. [2]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. [3]Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada. [4]Biological Services Unit, Weizmann Institute of Science, Rehovot, Israel. [5]Department of Pathology, Stanford University, Stanford, California, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to D.P. (dpeer@biology.columbia.edu).

Received 19 November 2015; accepted 12 April 2016; published online 2 May 2016; doi:10.1038/nbt.3569

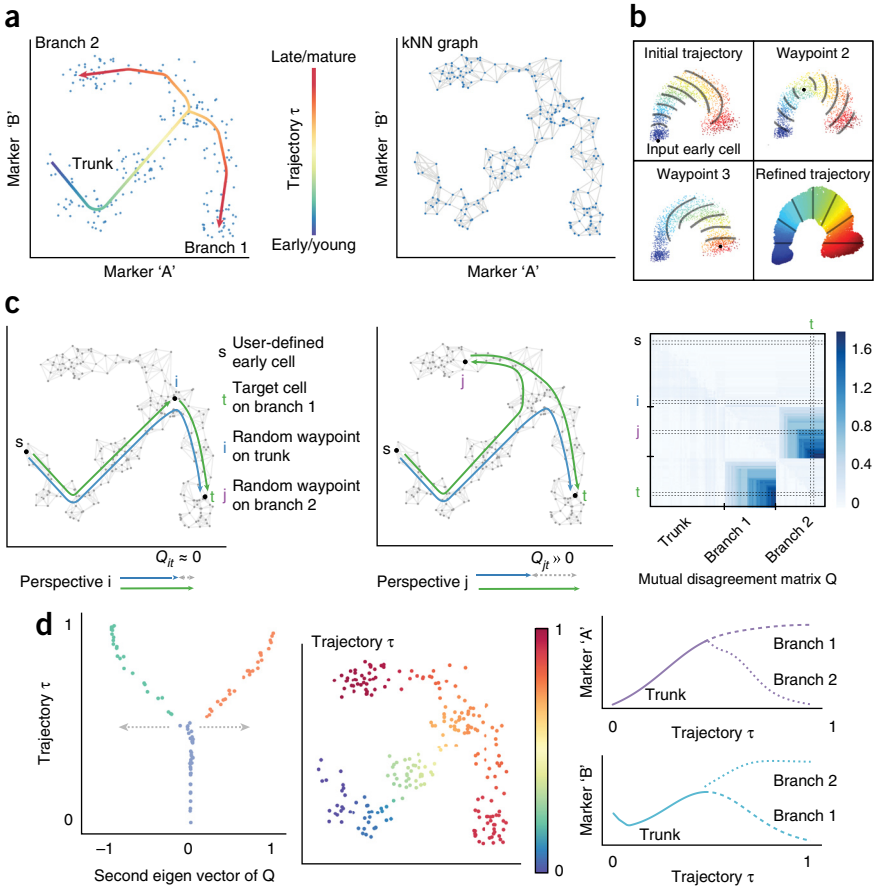

Figure 1 Alignment of cells along bifurcating trajectories. (**a**) Wishbone aims to achieve high-resolution ordering and branching of cells along bifurcating developmental trajectories. The data are represented as a *k*-nearest-neighbor graph where each cell is a node and edges connect each cell to its most phenotypically similar cells. kNN graph (right); data are simulated. (**b**) Wishbone uses a set of cells called "waypoints" to guide the ordering of cells. An initial ordering is derived using the shortest-path distances from the input early cell (top left panel). The distances from waypoints are aligned to the initial ordering to derive waypoint perspectives and the refined trajectory is determined as a weighted average of these perspectives (bottom right panel). The contour lines illustrate bands of cells that are at a similar distance from the corresponding waypoint. (**c**) Waypoints are also used for branch point identification and branch associations. The difference between the shortest path of waypoint *t* from early cell and a path that goes through another waypoint *i* is ≈ 0 if *i* and *t* are on the same trajectory (left) and ≫ 0 if they are on different branches (middle panel). These disagreements accumulate in the presence of a true branch to create a mutual disagreement matrix Q: observed are two sets of waypoints that agree within the set and disagree between sets (right). (**d**) The second Eigen vector of the Q matrix provides a summary of the disagreements with values ≈ 0 for waypoints on the trunk, > 0 for waypoints on one branch, and < 0 for waypoints on the other branch. The branch point and branch associations are used to further refine the trajectory. The resulting trajectory and branches are used to study marker dynamics along differentiation.

maturity between two cells (**Fig. 1a**). Similarly to our previous non-branching trajectory detection algorithm, Wanderlust[3], we use nearest-neighbor graphs to capture developmental distance and identify an initial ordering of cells using shortest paths (Online Methods). Each node in the graph represents a cell, and edges connect each cell to its most similar cells based on expression (**Fig. 1a**). Distances between cells can be computed using shortest paths, that is, a series of short steps through the neighbors in the graph, where each step between closely related cells is likely to represent similarity in degree of maturity.

Wishbone uses shortest paths from an input 'early cell' to build an initial ordering of cells, which is subsequently refined using a selected set of cells, called waypoints. Finally, the inconsistencies in distances between waypoints are used to identify the branch point and branch associations for all cells. The quality of the nearest-neighbor graph is critical for accurate ordering, and the major source of noise is the presence of "short-circuits" —spurious edges between cells that are farther apart in maturity[3]. Notably, a single short-circuit is sufficient to route all shortest paths between developmentally distant cells leading to incorrect ordering. Short-circuits are particularly prevalent in branching data sets, as cells following the bifurcation point might not be sufficiently distinct in their phenotypic characteristics (**Supplementary Fig. 1**). Wishbone overcomes these short-circuits by reconstructing the graph in projected space of reduced dimensions generated using diffusion maps[7] (Online Methods). Diffusion maps consider all possible paths between any pair of cells to dramatically reduce short-circuits. Wishbone uses the top diffusion components to construct the graph, capturing the major geometric structures in the data, while removing small fluctuations likely resulting from measurement noise.

The algorithm uses a select set of cells, called waypoints, to act as guides at different regions of the graph. Waypoints are randomly sampled cells, selected to represent regions along the entire trajectory and its branches (Online Methods). Each waypoint contributes a perspective, based on its computed distance to all other cells (**Fig. 1b**). The placement of a cell in the trajectory is determined by averaging the perspectives of all waypoints, with closer waypoints getting a higher weighting. Thus closer, more reliable waypoints predominantly determine a cell's position, while retaining a degree of influence of the distal waypoints to derive a consistent global structure (**Fig. 1b**, bottom right panel).

Waypoints are also the key to identifying branch points. If two waypoints *i* and *t* are along the same trajectory, the difference between the shortest path from the early cell to *t* and a path that goes through *i* is close to zero (**Fig. 1c**, left panel). On the other hand, if the two waypoints are on different branches, this difference is substantially greater than zero (**Fig. 1c**, middle panel). In the presence of a true branch, the disagreements between waypoints of the two branches accumulate to create two sets of waypoints that agree within each set and disagree between sets. These disagreements create a structured matrix (**Fig. 1c**, right panel): waypoints on the trunk have low disagreements with all waypoints, waypoints on one branch agree with other waypoints on the same branch and have high disagreements with all waypoints on the different branch (Online Methods). This structure can be identified with clustering approaches.

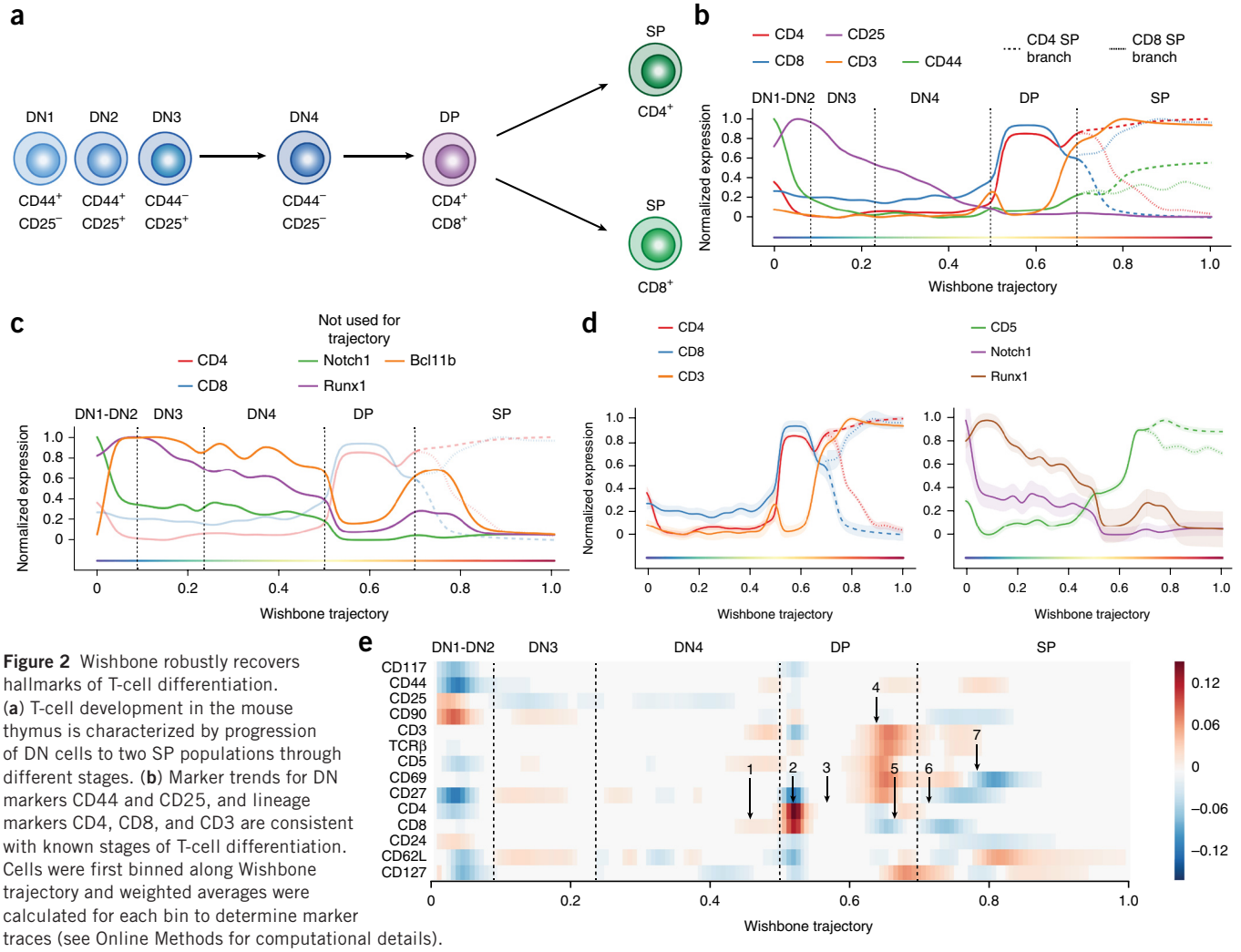

Figure 2 Wishbone robustly recovers hallmarks of T-cell differentiation. (**a**) T-cell development in the mouse thymus is characterized by progression of DN cells to two SP populations through different stages. (**b**) Marker trends for DN markers CD44 and CD25, and lineage markers CD4, CD8, and CD3 are consistent with known stages of T-cell differentiation. Cells were first binned along Wishbone trajectory and weighted averages were calculated for each bin to determine marker traces (see Online Methods for computational details). Following bifurcation, markers with different expression patterns in the two SP populations are shown in a dashed line for CD4 lineage and a dotted line for the CD8 lineage. (**c**) Bcl11b, Runx1, and Notch1 were not used for learning but the dynamics of these markers are consistent with their roles in specific developmental stages. (**d**) The variance of markers along the trajectory is tight, further highlighting the robustness of Wishbone results. (**e**) Derivative plot, showing the changes in expression of markers in successive bins, is used to time key events along the trajectory: (1) $CD8^+CD4^-$ intermediate single positive stage in DN to DP transition, (2) upregulation of CD4 and CD8 establishing DP cells, (3) stable expression of lineage markers during DP, (4) downregulation of both CD4 and CD8 accompanied by coordinated upregulation of CD3, TCRβ, CD5, CD69, and CD27 during positive selection, (5) specific downregulation of CD8 alongside upregulation of CD4 indicating intermediate thymocytes, (6) lineage commitment to two SP population and finally (7) successful completion of negative selection identified by downregulation of CD69 and upregulation of CD62L, indicating successful maturation. The branch with the highest expression is shown for markers with different expression patterns in the two SP branches.

Specifically, from spectral clustering techniques, the second Eigen vector of this matrix summarizes all the disagreements for a given waypoint and provides a quantitative measure of branch association for the waypoints (**Fig. 1d**, left panel, Online Methods). The extent of deviation from zero is a function of the maturity of the cell creating a Wishbone-like structure and giving the algorithm its name (**Fig. 1d**, left panel). Wishbone recovers the ordering of cells along their developmental trajectory, finds the branch point, and assigns cells following this point to one of the two branches (**Fig. 1d**).

Analysis of mouse thymus mass cytometry data set

During T-cell development in the mouse thymus, $CD4^+$ helper T cells and $CD8^+$ cytotoxic T cells bifurcate from lymphoid progenitors (**Fig. 2a**)[8,9]. We applied Wishbone to mass cytometry data from mouse thymus, with surface markers and transcription factors chosen based on their broad functionality in T-cell development (**Supplementary Table 1** and Online Methods). We collected data for five independent thymuses from Black6 mice; an average of 230k cells were collected per sample.

Wishbone was independently applied to each thymus, using only the surface markers for computing cell similarities (**Supplementary Table 1**) and defining the DN cell population as the starting point[8]. Marker trends along the resulting trajectory are depicted in **Figure 2b** (Online Methods). Wishbone accurately recovered the known stages in T-cell development (**Fig. 2b** and **Supplementary Fig. 2**), including the bifurcation into two single positive lineages ($CD4^+$ and $CD8^+$). Specifically, the trajectory begins at the DN stage ($CD4^-CD8^-$), transitions to the DP stage ($CD4^+CD8^+$) and finally branches to the two single positive (SP) populations[8]. We note that Wishbone correctly ordered the DN populations: DN2 ($CD44^+CD25^+$), DN3 ($CD44^-CD25^+$), and DN4 ($CD44^-CD25^-$), even though these cells are rare and constitute <1% of the cells in the thymus. DN1 ($CD44^+CD25^-$) cells are extremely rare,

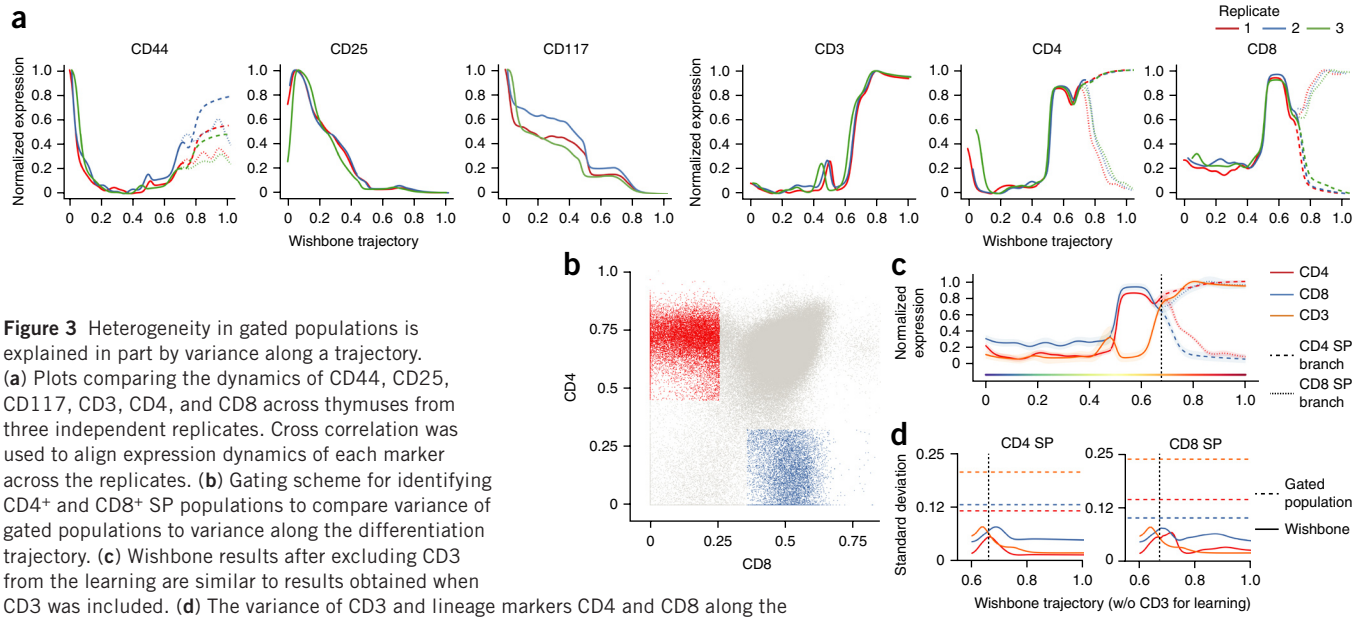

Figure 3 Heterogeneity in gated populations is explained in part by variance along a trajectory. (**a**) Plots comparing the dynamics of CD44, CD25, CD117, CD3, CD4, and CD8 across thymuses from three independent replicates. Cross correlation was used to align expression dynamics of each marker across the replicates. (**b**) Gating scheme for identifying $CD4^+$ and $CD8^+$ SP populations to compare variance of gated populations to variance along the differentiation trajectory. (**c**) Wishbone results after excluding CD3 from the learning are similar to results obtained when CD3 was included. (**d**) The variance of CD3 and lineage markers CD4 and CD8 along the trajectory (solid line) are substantially lower than the population variance (dotted line) in both branches, indicating that heterogeneity in gated populations is a result of comparing cells at different stages along their developmental maturity.

and we do observe a signature resembling these cells at the beginning of the trajectory (**Fig. 2b**). To further test Wishbone's accuracy, we evaluated the expression trends of markers not used while learning the trajectory: transcription factors Runx1 and Bcl11b, and signal molecule Notch1 (**Fig. 2c**). The abundance of all these markers is consistent with their known roles and timing in DN stages of T-cell development (**Supplementary Note 1**).

Additional evidence of Wishbone's accuracy is the tightness of marker variation over the course of the trajectory (**Fig. 2d**). Not only do the median marker levels follow expected trends, but almost every single cell is correctly placed in the trajectory, as indicated by low variance of markers across most of the trajectory. The variance is low for markers irrespective of whether the marker was used for learning the trajectory (**Fig. 2d** and **Supplementary Fig. 3**), reinforcing the robustness of Wishbone results.

Previous studies characterizing thymic development have largely relied on genetic perturbations and subsequent cell sorting that invariably eliminate specific developmental compartments. With 30 channels simultaneously measured, we could place DN, DP, $CD4^+$, and $CD8^+$ cells from a single thymus along a unified bifurcating trajectory and precisely order the course of multiple events along the trajectory measured directly from thymic tissue in an unbiased manner. We used derivative analysis to time key events along the trajectory (**Fig. 2e**) in a single frame of reference and found that Wishbone recovers a precise temporal ordering and branching of cells along with high resolution and accuracy using cells collected from a complex primary tissue (**Supplementary Note 1** and **Fig. 2b**,**e**).

Wishbone results are consistent across replicates and are robust to parameter choices

To test robustness, we applied Wishbone to three independent mouse thymuses and recovered consistent trajectories and branching behavior across all replicates. We used cross correlation to align the expression of individual markers, providing a quantitative measure for consistency across replicates (**Fig. 3a**). We find that the dynamics of marker expression and order of events along the trajectories are consistent across the replicates (**Supplementary Fig. 4**).

We investigated the sensitivity of the trajectory and branching to the various free parameters: number of neighbors k for the graph construction, number of waypoints nW, the sampling of waypoints from the cells, and number of diffusion components used. The ordering of cells and their branch points are remarkably robust to these different parameter choices across replicates (**Supplementary Figs. 5** and **6**, and **Supplementary Note 2**). Wishbone results are also largely robust to exclusion of individual markers used for learning (**Supplementary Fig. 7**). Moreover, the branches identified by Wishbone remained consistent irrespective of whether cells of DN or SP population are used as the input early cell (**Supplementary Fig. 8**).

Maturity controls for marker levels within individual cell types

We observe considerable heterogeneity in canonical surface markers. We hypothesized that at least part of this variation might be a result of developmental maturity, where cells from varied developmental stages are pooled into a single gated population. Using the fine temporal resolution of Wishbone's trajectory, we compared the marker variance, conditioned on the developmental progression of the cells, to that observed in gated populations. To make this comparison, we first identified the two SP populations using the standard gating scheme on the expression of the two lineage markers: CD4 and CD8 (**Fig. 3b**)[8]. We next compared the variance in these gated populations to the variance of the corresponding markers, conditioned on the Wishbone trajectory. In both SP populations, the variance of the lineage markers CD4 and CD8 and the co-receptor CD3, when controlled for maturity along the trajectory, is substantially lower compared to population variance in the gated populations (**Supplementary Fig. 9a**).

As an additional test, we ran Wishbone without using CD3 as one of the markers while learning the trajectory. The identified trajectory and branches are similar to results obtained including CD3 and is accompanied with only a minor increase in variance of CD3 all along the trajectory (**Fig. 3c**). However, the variance of all receptor and coreceptor molecules, CD4, CD8, and CD3, continue to be substantially lower along the trajectory compared to variance in the gated populations (**Fig. 3d**). These results similarly hold when either of CD4 or CD8 are excluded from learning (**Supplementary Fig. 9b**,**c**).

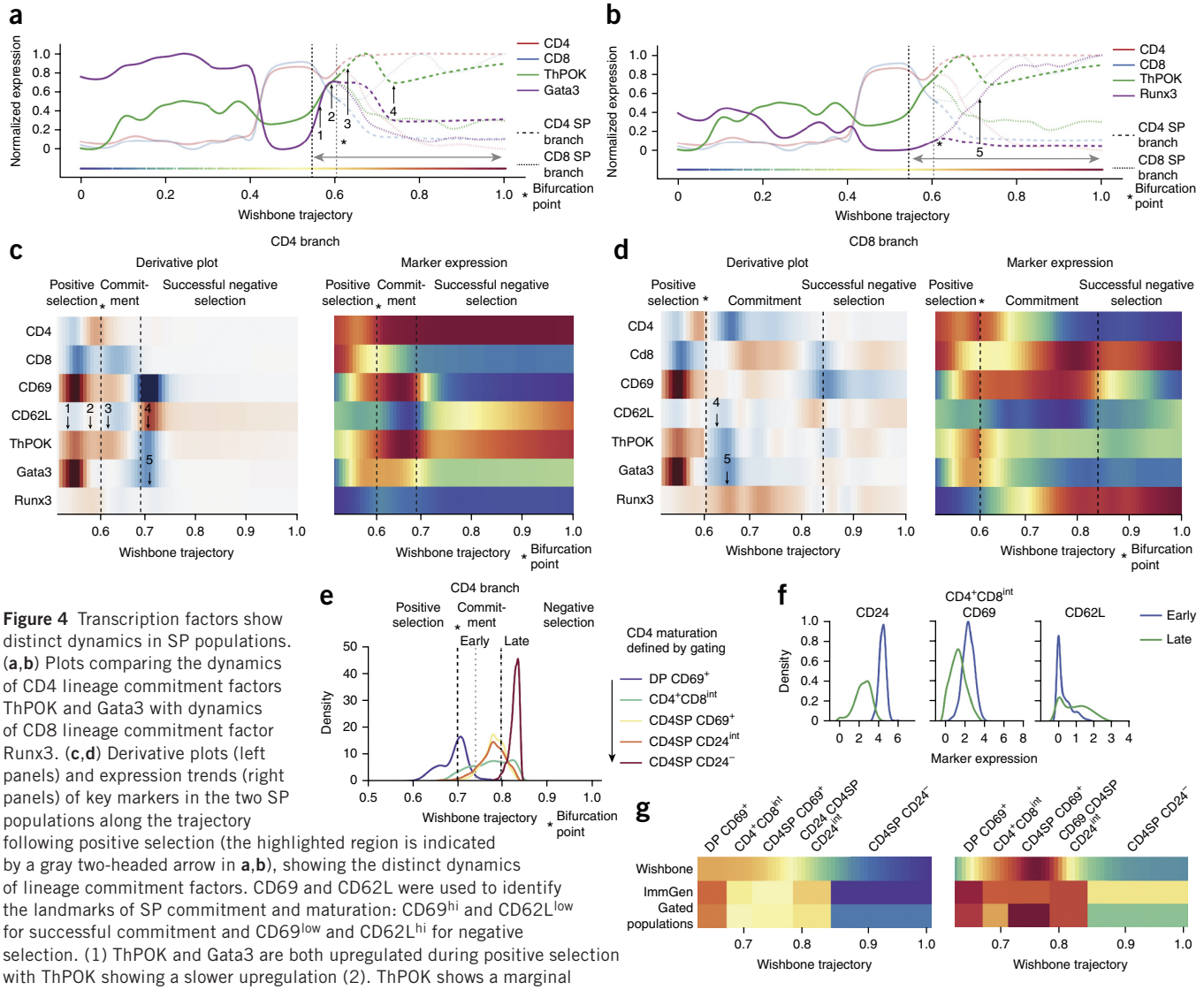

Figure 4 Transcription factors show distinct dynamics in SP populations. (**a**,**b**) Plots comparing the dynamics of CD4 lineage commitment factors ThPOK and Gata3 with dynamics of CD8 lineage commitment factor Runx3. (**c**,**d**) Derivative plots (left panels) and expression trends (right panels) of key markers in the two SP populations along the trajectory following positive selection (the highlighted region is indicated by a gray two-headed arrow in **a**,**b**), showing the distinct dynamics of lineage commitment factors. CD69 and CD62L were used to identify the landmarks of SP commitment and maturation: CD69^{hi} and CD62L^{low} for successful commitment and CD69^{low} and CD62L^{hi} for negative selection. (1) ThPOK and Gata3 are both upregulated during positive selection with ThPOK showing a slower upregulation (2). ThPOK shows a marginal upregulation specifically in the CD4 branch following commitment (3). ThPOK and Gata3 show a marginal downregulation in the CD4 branch during negative selection (**c**(4)). On the other hand, these factors are downregulated in the CD8 branch following commitment (**d**(4)). This downregulation is accompanied with a CD8-specific upregulation of Runx3 (5). (**e**) Cells were gated using the scheme defined in **Supplementary Figure 11** and were expected to be placed in the following order, indicating CD4 maturity: DP CD69^{+}, CD4^{+}CD8^{int}, CD4SP CD69^{+}, CD4SP CD24^{int}, and CD4SP CD24^{-}. Instead cells of the three intermediate gates are placed all along the CD4 Wishbone trajectory. These cells were divided into "early" and "late" populations based on their position in the Wishbone trajectory. (**f**) The "early" cells in the CD4^{+}CD8^{int} gate show significantly higher expression of CD69 and CD24 and lower expression of CD62L compared to "late" cells ($P < 1 \times 10^{-6}$, Kolmogorov-Smirnov test). This indicates that "late" cells are more mature than the "early" cells. (**g**) mRNA expression of CD69 and CD24 in ImmGen-sorted populations are correlated with mean expression in the gated populations, demonstrating that the discrepancy between Wishbone and gating is not data set-specific.

Collectively, these results suggest that a substantial part of the heterogeneity of marker expression in gated populations are a result of comparing cells at different stages along their developmental maturity, rather than stochasticity in marker expression.

Transcription factor dynamics along SP trajectories

Next we set out to explore the dynamics of key transcription factors along the two SP trajectories, using a revised panel including the transcription factors ThPOK, Gata3, and Runx3. ThPOK and Gata3 have been shown to be critical for the CD4 SP population[10] and Runx3 has been demonstrated to be key for CD8 SP commitment[10]. The dynamics of these factors along the trajectory is shown in **Figure 4a**,**b**.

To place the dynamics of these transcription factors in context, we used CD69 and CD62L to identify landmarks of maturation such as lineage commitment and successful negative selection (**Fig. 4c**,**d**). Our results suggest that these factors follow distinct dynamic patterns in achieving commitment. ThPOK and Gata3 are upregulated during positive selection but Runx3 appears to be upregulated only following the detected branch point (**Fig. 4c**,**d**, **Supplementary Note 3** and **Supplementary Fig. 10**). Gata3 has been shown to regulate ThPOK expression[11] and might offer an explanation as to why ThPOK trails Gata3 in expression changes (**Fig. 4c**). The different dynamics can potentially be indicative of distinct regulatory mechanisms through which these factors achieve lineage commitment. Further experiments are necessary to elucidate these mechanisms.

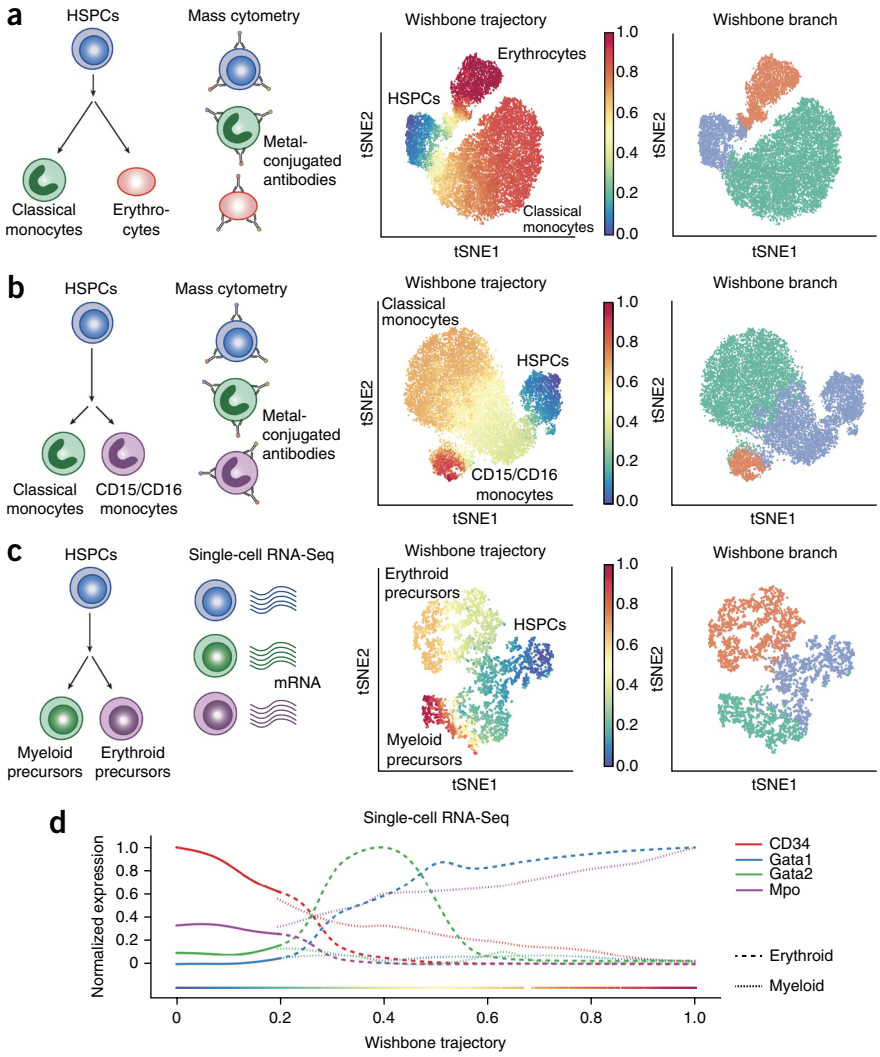

Figure 5 Generalization of Wishbone to branches in human and mouse myeloid development spanning mass cytometry and single-cell RNA-Seq. (**a**) Wishbone was applied to an early step in human myeloid development to track the differentiation of classical monocytes ($CD14^+CD11b^+CD11c^+$) and erythrocytes ($CD235ab^+$) from hematopoietic stem and progenitor cells (HSPCs). (See also **Supplementary Fig. 14**). tSNE maps showing each cell colored by the trajectory (left panel) and the branch associations (right panel). Wishbone accurately orders the cells with HSPCs at the start and the differentiated cells toward the end. The inferred branch associations are also consistent with the annotated cell types (**Supplementary Fig. 14**). (**b**) Same as in **a**, for tracking differentiation of classical monocytes and $CD15^+$ monocytes, a late step in human myeloid development. (**c**,**d**) Wishbone was applied to single-cell RNA-Seq data from the hematopoietic precursors from the mouse and accurately recovered the trajectory and branches to track differentiation of myeloid and erythroid precursors from HSPCs.

We compared marker dynamics along the Wishbone trajectory to the dynamics derived from gating of developing SP cells[12–14] (**Supplementary Fig. 11**, Online Methods) and compared the ordering of cells within each population along the Wishbone trajectory.

We observe that cells within most gated populations are spread out along the trajectory (**Fig. 4e** and **Supplementary Fig. 12b**), particularly cells of the $CD4^+CD8^{int}$ (int for intermediate) population, where the lineage decision is thought to occur[15]. To address this discrepancy between Wishbone and gating, we divided the $CD4^+CD8^{int}$ cells into "early" and "late" groups based on their positions in the Wishbone trajectory (**Fig. 4e**, Online Methods) and compared the expression of known maturation markers CD69, CD24, and CD62L in the two groups. The "early" cells show significantly higher expression of CD69 and CD24, and lower expression of CD62L compared to "late" cells (**Fig. 4f**; $P < 1 \times 10^{-6}$, Kolmogorov-Smirnov test) demonstrating that the cells in "early" and "late" are immature and mature, respectively. Similar results for additional gated populations in the CD4 and CD8 branch (**Supplementary Fig. 12**) indicate that the conventional gating scheme leads to inclusion of cells at different stages of maturation in each gate. We conclude that Wishbone provides more reliable estimates of cell maturation and hence marker dynamics along SP maturation.

To understand the source of this discrepancy, we compared the mean gene expression of markers in Immunological Genome Project (ImmGen)-sorted populations[16] to Wishbone marker dynamics and indeed mean protein expression in our gated populations and mean mRNA expression in ImmGen populations was correlated. Thus, the discrepancy between dynamics observed along the Wishbone trajectory and gating is not a data set-specific observation (**Fig. 4g**). The mixing of developmentally distinct cells in each gate can lead to confounding effects on expression change patterns along maturation. Whereas CD24 showed a continuous decrease along the Wishbone trajectory, the expression was more variable in the gated and ImmGen populations (**Fig. 4g**, left panel). Sustained upregulation of CD69 following positive selection is not observed in gated populations as CD69 itself is used for gating (**Fig. 4g**, right panel). Finally, Gata3 expression changes in gated populations do not show the dynamics observed in Wishbone (**Fig. 4c** (1, 5)) even though Gata3 was not used for gating (**Supplementary Fig. 12g**), further demonstrating the ability of Wishbone to recover marker dynamics at higher resolution.

Application of Wishbone to human myeloid development

We evaluated the performance of Wishbone on two human myeloid development data sets. We used previously published mass cytometry data[2] consisting of markers that are suitable to recover early and late Myeloid bifurcations[17] but not amenable for fine-grained profiling of transitional myeloid populations (**Supplementary Fig. 13**, Online Methods).

Wishbone was able to track the differentiation of monocytes ($CD14^+CD11b^+CD11c^+$) and erythroid cells ($CD235ab^+$) from hematopoietic stem and progenitor cells (HSPCs) (**Fig. 5a** and **Supplementary Fig. 14a**) and accurately recover the monocyte and erythroid branches (**Fig. 5a**). More over, expression of markers along the trajectory was consistent with known literature[18] (**Supplementary Fig. 14c**).

Wishbone accurately recovered the trajectory starting from HSPCs and the branching of the two-monocyte classes, classical

monocytes and CD16 monocytes ($CD16^+CD15^+$) (**Fig. 5b**). This is a harder problem as most of the markers showed identical distributions between the two populations except for the characteristic markers, CD15 and CD16 (**Supplementary Fig. 14d**). The expression of markers along the trajectory was consistent with known literature (**Supplementary Fig. 14e**), with the detected bifurcation point coinciding with significant downregulation of CD38.

Extension of Wishbone to single-cell RNA-Seq data

Single-cell RNA-Seq technologies can profile thousands of single cells and enable genome-wide characterization of developing systems[6,19]. However, such data pose a challenge in that the behaviors of many genes are related not to developmental maturity but to processes such as cell cycle and stress. Thus, the success of trajectory and branch detection relies on removing unrelated factors and retaining those that track with differentiation.

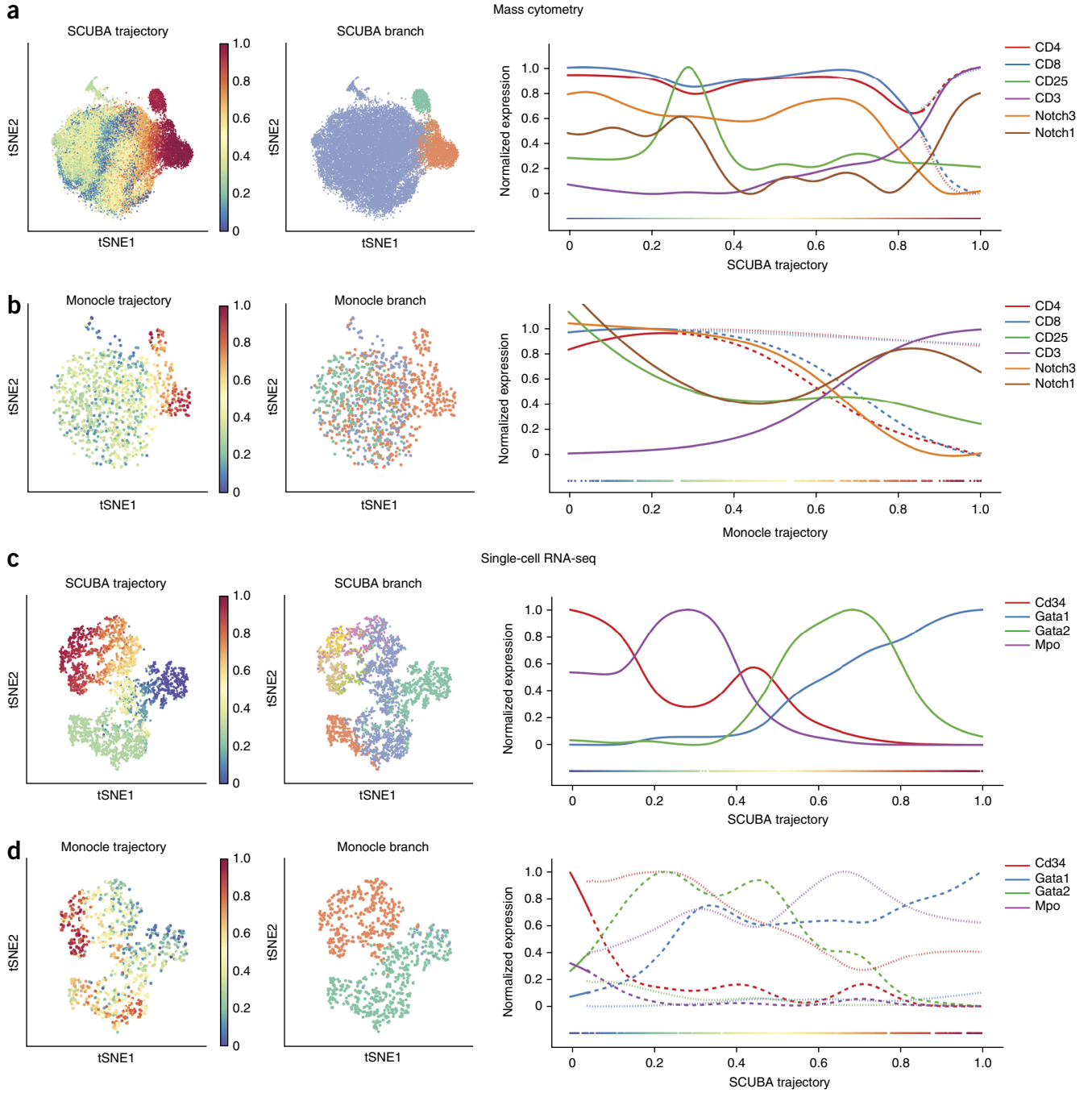

Figure 6 Wishbone outperformed competing methods in both ordering of cells and branch associations. (**a**) tSNE maps showing SCUBA results for a random sample of 20,000 mouse thymic cells (left and middle panels). SCUBA trajectory does not distinguish between the DN and DP stages. Although SCUBA recovers the SP branches, it suffers from a loss of resolution in the SP stage (right panel). (**b**) Plots showing Monocle results for a random sample of 1,000 mouse thymus cells. Monocle fails to correctly order the cells, and the branches do not correspond to the SP populations. (**c**) SCUBA accurately recovers the ordering of mouse myeloid cells and the marker dynamics are largely consistent with known biology. SCUBA, however, results in a large number of incoherent branches. (**d**) Monocle fails to accurately order the myeloid precursors correctly and also fails to detect a coherent HSPC branch.

We used recently published single-cell RNA-Seq data[6] to select cells involved in differentiation of myeloid and erythroid progenitors from HSPCs (**Fig. 5c** and **Supplementary Fig. 15a**). We devised an extension of Wishbone, adapted to single-cell RNA-Seq, that uses diffusion maps to help focus on components related to development and maturation. Diffusion maps capture the major structures and trends in the data, and in the case of mass cytometry, different diffusion components track the differences among constituent cell types (**Supplementary Fig. 6a**). We projected genes down onto each diffusion component, ranked genes based on how well their expression tracked along this component, and then used this ranking to perform gene set enrichment analysis (GSEA)[20] (Online methods). Some diffusion components were enriched for immune-related functions (e.g., defense response, antigen processing, and phagocytosis), whereas other components were enriched for other biological processes (e.g., cell cycle, ribosome biogenesis and metabolic processes) (**Supplementary Fig. 15c**, Online Methods). This provides a natural way to retain the components that are most relevant to the differentiation processes. With similar reasoning, Buettner *et al.*[21] use latent variable models to remove the contribution of cell cycle in single-cell RNA-Seq.

We constructed Wishbone's neighbor graph based on a projection of the data onto only the differentiation-related components, and once this graph was constructed we proceeded with Wishbone as described for mass cytometry. Wishbone accurately recovered the trajectory starting from HSPCs and terminating at the two precursor cell types and the branch associations (**Fig. 5c**). The marker trends showed a consistent decrease of HSPC marker CD34 along the trajectory with an increase in expression of myeloid marker Mpo[22] along the myeloid branch (**Supplementary Fig. 15c**). Consistent with known biology, Gata2 was upregulated before Gata1 along the erythroid lineage[23] (**Fig. 5d**).

Wishbone outperformed competing methods in both trajectory and branch identification

We compared the performance of Wishbone to Diffusion maps[7], SCUBA[5], and Monocle[24] (**Fig. 6**). Although we used diffusion maps to build the kNN graph, we tested whether diffusion maps on their own can recapitulate developmental trajectories[25,26]. Note that diffusion maps did not explicitly provide bifurcations, and we could only evaluate their ability to recapitulate an accurate ordering. Diffusion maps correctly recovered the various known stages in T-cell development (**Supplementary Fig. 16b**), especially in the early DN states, but suffered from a considerable lack of resolution in DP and SP populations (**Supplementary Fig. 16a**,**b**). Moreover, whereas diffusion maps recovered the right order in the two myeloid data sets (**Supplementary Fig. 16c**,**e**), in the monocyte data set, diffusion maps ordered precursors after the mature cells (**Supplementary Fig. 16d**). Thus, although diffusion maps substantially reduced the noise in the data (**Supplementary Fig. 1**, Online Methods), the additional steps taken by Wishbone to refine ordering of cells are critical to derive robust, high-resolution trajectories.

Next, we compared Wishbone to SCUBA[5]. SCUBA has a large memory footprint and therefore could only be run by subsampling 20,000 cells from the thymus data set. The SCUBA trajectory of the thymus did not order the stages correctly, and we observed the different DN cells interspersed among the DP cells (**Fig. 6a**). SCUBA did identify the two SP populations as the two branches, but with reduced resolution at the bifurcation point compared to Wishbone (**Fig. 6a**). Moreover, different random sample of cells led to largely inconsistent results (**Supplementary Fig. 17a**,**c**) in both trajectory and branching. SCUBA trajectory in the mass cytometry monocyte and the single-cell RNA-Seq myeloid data sets was consistent with known biology, but yielded a large number of incoherent branches (**Fig. 6c** and **Supplementary Fig. 17d**). Moreover, SCUBA failed to correctly recover the order and branching in the monocyte-erythroid data set (**Supplementary Fig. 17e**).

Finally, we compared Wishbone to Monocle[24], which was specifically developed for application to single-cell RNA-Seq data. Monocle could not be run with more than 1,000 cells, and we therefore subsampled 1,000 cells from each data set. Monocle did not recover the correct ordering in the thymus data with DN and DP cells interspersed (**Fig. 6b**). Although the trajectory does end at the two SP populations, the branching identified by Monocle did not correspond to any specific stages in T-cell development, and both the SP populations were identified to be part of the same branch (**Fig. 6b**). Repeated subsampling of the data resulted in largely inconsistent results with the two SP populations repeatedly assigned to the same branch (**Supplementary Fig. 18a–c**). Monocle also failed to recover the trajectory and branches in the single-cell RNA-Seq myeloid data set with incorrect ordering of cells and lack of detection of a coherent branch (**Fig. 6d**). Monocle did recover ordering in the monocyte data set, but the branching results in all the myeloid data sets did not correspond to the correct mature cell populations (**Supplementary Fig. 18d**,**e**).

Thus, Wishbone outperformed competing methods in fine ordering of cells, identification of branch point and branch associations, and consistent robustness across replicates.

DISCUSSION

We have developed an algorithm that enables accurate and high-resolution ordering of cells along branched developmental trajectories (**Supplementary Fig. 19**). We first demonstrated Wishbone on T-cell development in the mouse thymus, using the throughput of mass cytometry to collect ≥200,000 cells per sample. Wishbone constructed a bifurcating trajectory starting from DN stages through maturation of the two SP lineages, providing an order and timing of events that closely recapitulated previous studies of this system[15]. The high resolution of Wishbone enabled us to identify subtle but key dynamics of lineage markers such as detection of the rare $CD8^+CD4^-$ intermediate SP cells during transition of DN to DP cells and the intermediate $CD4^+/CD8^{low}$ state toward the end of DP.

The selection of a good marker set was key to the resolution we achieved. Marker choice can be guided by a combination of prior knowledge and preliminary screens. However, in the myeloid branches we demonstrated that even with a limited panel that included only a small number of distinguishing myeloid markers, Wishbone correctly ordered cells, identified the bifurcation, and associated cells to the proper branch. Although an explicit ground truth is not necessarily available, both SCUBA and Monocle failed to recover the expression trends and bifurcations that are consistent with known biology in these more challenging data sets. Wishbone only required a few canonical markers to properly identify bifurcation, and achieved increasingly finer resolution in transitional populations, as additional markers were included.

Single-cell RNA-Seq is an attractive alternative to mass cytometry as its unbiased, genome-wide nature provides measurements for thousands of genes and circumvents the need for *a priori* selection of a limited marker set. However, transcriptional changes unrelated to development can confound the analysis, and even data for developmentally related genes has substantial noise, including drop-out effects[27]. We used diffusion maps to consolidate the key biological trends and removed unrelated biological processes. We demonstrated that Wishbone substantially outperformed methods developed specifically for single-cell RNA-Seq data[5].

Even with the increasing throughput of single-cell RNA-Seq, current data sets include thousands of cells, as compared to hundreds of thousands available in mass cytometry. As transitional populations have been shown to be as rare as 1/10,000 cells[3], the throughput of mass cytometry is better suited to achieve finer temporal resolution. In our view, the two technologies are complementary. For example, single-cell RNA-Seq can be used for unbiased marker selection in less-studied developmental systems, and finer temporal resolution can then be achieved with mass cytometry using the identified panel.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Mouse thymus mass cytometry data have been deposited to Cytobank under accession 52942.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank A. Bloemendal, Z. Good, N. Hacohen, S. Krishnaswamy, J. Levine and A.J. Carr for their helpful comments. M.D.T. is supported by an NSF graduate fellowship. This work was supported by NSF MCB-1149728, NIH DP1-HD084071, NIH R01CA164729 to D.P. D.P. holds a Packard Fellowship for Science and Engineering. This work was also supported by David and Fela Shapell Family Foundation INCPM Fund, the WIS staff scientists grant from the Nissim Center, for the Development of Scientific Resources, and ISF 1184/15 to N.F.

AUTHOR CONTRIBUTIONS

S.B. and D.P. conceived the study. M.S., M.D.T., O.A., and D.P. designed and developed Wishbone. M.S. and D.P. performed statistical analysis and comparison of Wishbone. S.R.-Z., T.M.S., and N.F. performed all bench experiments and data acquisition. M.S., S.R.-Z., N.F., and D.P. performed the biological analysis and interpretation. M.D.T., M.S., P.K., and K.C. programmed the software tools. M.S., S.R.-Z., N.F., and D.P. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
2. Bendall, S.C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
3. Bendall, S.C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
4. Shin, J. *et al.* Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* **17**, 360–372 (2015).
5. Marco, E. *et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. USA* **111**, E5643–E5650 (2014).
6. Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
7. Coifman, R.R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. USA* **102**, 7426–7431 (2005).
8. Koch, U. & Radtke, F. Mechanisms of T cell development and transformation. *Annu. Rev. Cell Dev. Biol.* **27**, 539–562 (2011).
9. Yui, M.A. & Rothenberg, E.V. Developmental gene networks: a triathlon on the course to T cell identity. *Nat. Rev. Immunol.* **14**, 529–545 (2014).
10. Egawa, T. Regulation of CD4 and CD8 coreceptor expression and CD4 versus CD8 lineage decisions. *Adv. Immunol.* **125**, 1–40 (2015).
11. Wang, L. *et al.* Distinct functions for the transcription factors GATA-3 and ThPOK during intrathymic differentiation of CD4(+) T cells. *Nat. Immunol.* **9**, 1122–1130 (2008).
12. Love, P.E. & Bhandoola, A. Signal integration and crosstalk during thymocyte migration and emigration. *Nat. Rev. Immunol.* **11**, 469–477 (2011).
13. Mingueneau, M. *et al.* The transcriptional landscape of αβ T cell differentiation. *Nat. Immunol.* **14**, 619–632 (2013).
14. Yamashita, I., Nagata, T., Tada, T. & Nakayama, T. CD69 cell surface expression identifies developing thymocytes which audition for T cell antigen receptor-mediated positive selection. *Int. Immunol.* **5**, 1139–1150 (1993).
15. Singer, A., Adoro, S. & Park, J.H. Lineage fate and intense debate: myths, models and mechanisms of CD4- versus CD8-lineage choice. *Nat. Rev. Immunol.* **8**, 788–801 (2008).
16. Heng, T.S. & Painter, M.W. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9**, 1091–1094 (2008).
17. Rosenbauer, F. & Tenen, D.G. Transcription factors in myeloid development: balancing differentiation with transformation. *Nat. Rev. Immunol.* **7**, 105–117 (2007).
18. Doulatov, S. *et al.* Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat. Immunol.* **11**, 585–593 (2010).
19. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
20. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
21. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-Sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
22. Pinkus, G.S. & Pinkus, J.L. Myeloperoxidase: a specific marker for myeloid cells in paraffin sections. *Mod. Pathol.* **4**, 733–741 (1991).
23. Kaneko, H., Shimizu, R. & Yamamoto, M. GATA factor switching during erythroid differentiation. *Curr. Opin. Hematol.* **17**, 163–168 (2010).
24. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
25. Haghverdi, L., Buettner, F. & Theis, F.J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
26. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33**, 269–276 (2015).
27. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).

ONLINE METHODS

Mouse thymus data and mass cytometry. Female C57BL/6 mice were obtained from Harlan Laboratories. All mice were housed at the Weizmann Institute in compliance with national and international regulations. Thymocytes were isolated from the thymus of 6-week-old C57Bl mice. Cells were stained with metal-conjugated antibodies according to manufacturer's protocol (**Supplementary Table 1**). Briefly, around 200k cells were stained with cell-ID TM Cisplatin (Fluidigm) (5 min RT). Next cells were stained with surface antibodies (30 min RT), and fixed with 1.6% PFA (10 min RT). After permeabilization with 100% ice-cold Methanol (15 min, 4 °C), the cells were stained with intracellular antibodies (30 min, RT). Finally, the cells were labeled with Iridium DNA intercalator for DNA content and analyzed by CyTOF mass cytometry using CyTOF2. Data were normalized using bead normalized with bead standards[1].

We collected data for five independent thymuses from Black6 mice using two different marker panels. The first predominantly contains cell surface markers and the second combines the most informative of these surface markers with known regulators of lineage commitment (**Supplementary Table 1**).

Data preprocessing and choice of parameters for Wishbone. Mass cytometry data channels were first arcsinh transformed with a cofactor of 5 (ref. 2). Cell doublets, barcodes, dead cells, and debris were removed from the data using the gating scheme shown in **Supplementary Figure 20**. Next, the cells were clustered using Phenograph[28].

For the mouse thymus data set, the clusters corresponding to myeloid cells (expression of CD11b, CD11c), B cells (CD19), NK cells (CD161), regulatory T cells (CD25) and TCRγδ cells were filtered out from the analysis (**Supplementary Fig. 21**). The remaining clusters correspond to the DN, DP, and SP populations. Finally, for each thymus a start cell was sampled from the DN population, and the same start cell was used for all the analyses of that thymus. The results presented in the paper were generated using number of nearest neighbors $k = 15$ and number of waypoints $nW = 250$.

Human bone marrow mass cytometry data were downloaded from ref. 2. Doublets, cell debris and dead cells were removed as described above. Phenograph was used to identify the clusters of cells and all the lymphoid clusters were removed (**Supplementary Fig. 13**) and clusters for generating the data sets used in **Figures 5** and **6** were identified by expression of characteristic markers (**Supplementary Fig. 14a**,**d**).

Overview of the Wishbone algorithm. *Introduction.* Differentiation is a complex process involving multiple cell fate decision points. This process can be seen as hierarchical tree with the multipotent stem and progenitor cells at the root and the mature differentiated cell types at the bottom with various precursor cells as intermediate cell types[15,17,29]. Emerging high- throughput technologies such as single cell RNA-Seq[1,19,30] and mass cytometry[2] are enabling generation of data with unprecedented resolution and require computational algorithms capable of exploiting this resolution. Wishbone uses multi-dimensional single-cell data to align cells along bifurcating trajectories. Wishbone was developed to study systems where the developmental trajectory bifurcates to one of two cell fates (**Fig. 1a**).

There are two key challenges involved in studying trajectories with branches: (i) ordering of cells within the trunk and in each of the branches, (ii) identification of branch point and assignment of cells to either the trunk or one of the branches. Previous studies attempting to study differentiation have largely relied on sorted populations. While these have led to important advances, the dynamics of marker behavior along the maturation trajectory cannot be characterized without an accurate, high resolution ordering of cells, capable of characterizing the order and timing of key molecular events during development. The second challenge is to assign the cells to their respective branches. Given a right set of markers, it is relatively straightforward to classify the mature cells into the correct branches. However, there are many uncharacterized bifurcations where such markers are not well defined. Moreover, a precise identification of branch point is central to achieve a high resolution into bifurcating trajectories to understand the series of events leading up to and following cell fate decisions. Furthermore, cells can be thought of as being in a state of flux at the branch point necessitating a soft assignment of branches.

Wishbone addresses these challenges by taking a graph-based approach to measure distances between cells, similar to the approach used by Wanderlust[3], our previously published algorithm for detecting non-branching trajectories. Wishbone first constructs a nearest-neighbor graph of cells and estimates distances between them using the shortest path algorithm. The greedy nature of shortest path algorithms makes them susceptible to short-circuits i.e., connections between developmentally distal cells. Wishbone overcomes this problem by use of diffusion maps[7], a dimensionality reduction technique, to reduce noise and eliminate short-circuits. The initial ordering of cells as determined by shortest path distances from an input early cell are increasingly prone to noise as distance increases. Wishbone uses a series of cells called waypoints, sampled all along the entire trajectory to locally refine the ordering of cells and overcome this noise. Finally, the disagreements between waypoint's and the early cell's distances to other cells are used to detect the branch point and branch assignments. The ordering and branch assignments are iteratively repeated until convergence. These randomly sampled waypoints provide a sparse approximation for the entire data set. Randomly sampled subsets have previously been used to achieve more computationally efficient dimensionality reduction[31]. A key distinction between Wishbone's waypoints and other such sparse approximation schemes is that the waypoints are themselves the driving force underlying the algorithm. Thus, Wishbone recovers an accurate high-resolution ordering and branching of cells in bifurcating trajectories.

Wishbone makes the following assumptions about the data: (i) the maturation process along differentiation is continuous, and (ii) the snapshot of primary tissue at any given point is representative of the entire differentiation process with various intermediate populations represented, and (iii) the developmental trajectory bifurcates to one of only two cell fates.

Nearest-neighbor graph and shortest paths. Differentiation is characterized by a series of increases and decreases in expression of specific markers[3]. Furthermore, the rise and fall of the markers involved in development create nonlinear relations between the markers and their relation to maturity (**Fig. 1**)[3]. Therefore, distance metrics such as Euclidean distance fail to accurately capture the similarity between cells that are at distinct stages of development. As previously demonstrated[3], nearest-neighbor graphs are a powerful alternative to capture developmental distances. Here, each cell is a node and is connected to its nearest neighbors, that is, the cells that are most similar in the phenotypic profiles. The underlying assumption being that for very short distances, marker similarity represents a similar developmental maturity. The edge weights are set to the similarity between the connected cells.

Given the graph, a path can be defined from one cell to another through a series of short steps represented by edges, since each of these edges represent a more confident developmental proximity. While there are many possible paths through the graph between any given pair of cells, an efficient choice is to take a path such that the total weight of edges is minimized. This minimal sum of weights is also referred to as the shortest path distance between two cells and can be used as a distance metric between cells[3]. Thus, the shortest path distances to all cells from the viewpoint of a cell early in development, denoted s, can be used as a starting point to build the trajectory.

Short-circuits and diffusion maps. Key to the success of the algorithm is construction of a good graph, where edges in the graph connect cells that are indeed close in their developmental progression. One of the problems affecting the construction of a good graph is presence of short-circuits: spurious edges between cells that are farther apart in development but are identified as neighbors due to measurement noise. A single short-circuit is sufficient to route all the shortest paths between distant cells through this anomalous edge resulting in incorrect ordering of cells. Since short-circuits are relatively rare in non-branching trajectories, Wanderlust proposed the use of ensemble of graphs where trajectories were determined by repeatedly sampling a subset of edges from the graph[3].

Short-circuits, however, are considerably more prevalent in branching data sets, particularly close to the bifurcation point, as the markers that characterize the branches might not be sufficiently distinct in this region. Furthermore, depending on the extent of separation of the branch and noise in the characteristic markers, short-circuits might also be present all along the maturation trajectory. An illustrative example is shown in **Supplementary Figure 22a**. The ensemble of graph methods fails to sufficiently remove these short-circuits as it assumes the number of short-circuits to be significantly

doi:10.1038/nbt.3569

fewer in number (**Supplementary Fig. 22b**: each panel was derived by sampling a subset of edges).

Wishbone therefore uses diffusion maps[7] to remove short-circuits in the data and construct a graph that is more faithful to the developmental trajectory. Diffusion maps are a nonlinear dimensionality reduction technique to derive a low-dimensional description of high dimensional data by exploiting local similarities[7]. Rather than rely solely on the shortest paths in the phenotypic marker space, diffusion maps generate a low-dimensional embedding by approximating all possible paths through the graph, avoiding the harmful effect of short-circuits. One can view diffusion maps as a nonlinear version of Principal Component Analysis (PCA). Often data are de-noised by projecting data onto the top principal components, assuming the smaller components represent noise[32]. Similarly, by projecting the data onto the top diffusion components, we capture the major structures in the graph and remove small fluctuations, providing a nonlinear data clean up step.

While diffusion maps often generate a first-order approximation of the developmental trajectory, the resulting resolution is not sufficiently fine as shown in **Figure 6**. Therefore, Wishbone constructs a nearest-neighbor graph in the embedded space to bring together advantages of graph-based methods for trajectory building and de-noising nature of diffusion maps. The graphs constructed in the embedded space tend to be free of most short-circuits (**Supplementary Fig. 22c**) and therefore shortest paths can be used for computing distances between cells.

Graph construction and initial ordering of cells. Formally, given a data set with N cells and M markers, Wishbone starts by transforming the high-dimensional phenotypic data into low-dimensional data using diffusion maps. The embedding is computed by using the diffusion geometry code (http://www.math.duke.edu/~mauro/diffusiongeometries.html) with default parameters. This embedded space is used to construct a k-nearest-neighbor graph (k - NNG), **G**, spanning all the cells. Each cell i is connected to its k nearest cells via Euclidean distance in the embedded space and edges connecting cells to their nearest neighbors are weighted by the Euclidean distance between them.

An early cell s, provided as user input, is then used to compute an initial alignment of cells by computing shortest path distances from s to all cells. The distance from s to any given cell i is calculated using Dijkstra's algorithm:

$$D_{si} = \min_{\mathbf{P}} \sum_{e \in \mathbf{P}} G_{e1,e2} \tag{1}$$

where **P** is a path between s and i, and $G_{e1,e2}$ is the weight of edge e. Note that the graph is undirected and therefore $D_{is} = D_{si}$.

The trajectory or ordering of cells is initialized to the shortest path distance from s i.e., $\tau_i^{(0)} = D_{is}$. This initial ordering encapsulates this early cell's perspective of the other cells' progression, based on their computed shortest-path distance from s (**Fig. 1b**, top right panel).

Waypoints and perspectives. Shortest path distances are robust at short distances but become less reliable with increasing distance from the source cell. **Supplementary Figure 23a** shows the loss in reliability of shortest path distances. The additive nature of noise leads to accumulation of mistakes with distance and becomes the dominant factor with greater distances (**Supplementary Fig. 23a** and **Fig. 1b**).

Wishbone, like Wanderlust, overcomes this issue by sampling a series of cells throughout the trajectory termed "waypoints" to act as guides in ordering the cells[33]. The ordering of cells is then averaged across the waypoints with closer waypoints giving a bigger "vote." This improves the robustness by taking advantage of reliability of shortest path distances over short distances. As described later, waypoints are also used for branch associations and due to the importance of waypoints for both the ordering of the cells and branch identification, a key difference between Wanderlust and Wishbone is how waypoints are selected and weighted.

A random sample of cells can potentially select outliers as waypoints. Wishbone therefore refines the choice of waypoints by using a median filter[33]. For each randomly selected waypoint, its k nearest neighbors are identified and the waypoint is replaced by the cell closest to the median profile generated using these neighbors. This refinement step has been shown to be effective in preventing the outlier cells from being chosen as waypoints for learning trajectories[33].

Next, shortest path distances are computed for each of the waypoints to obtain the distance matrix, $\mathbf{D} \in \mathbb{R}^{nW \times N}$, where nW is the number of waypoints including the early cell s. Individually, the distances from each waypoint are still affected by the same issues of increasing noise with distance from the waypoint (**Supplementary Fig. 23a**). But collectively, each cell is close to a number of waypoints that can reliably estimate the ordering along developmental axis.

Waypoints are introduced to robustly order the cells by computing a weighted average but the distances from different waypoints in **D** are not aligned and therefore are not directly comparable (**Supplementary Fig. 23a**). Thus, Wishbone computes the positioning or ordering of cells from the perspective of each waypoint using the initial trajectory $\tau^{(0)}$ as the reference. The perspective of a cell i with respect to waypoint w is the distance of i from the s from the viewpoint of w and is computed as

$$P_{wi} := \begin{cases} \tau_w^{(0)} + D_{wi} & \text{if } \tau_i^{(0)} > \tau_w^{(0)} \\ \tau_w^{(0)} - D_{wi} & \text{otherwise} \end{cases} \tag{2}$$

This ensures that cells beyond the waypoint in the initial ordering have a higher perspective than cells that lie before the waypoint (**Supplementary Fig. 23b** and **Fig. 1b**). Thus, the unaligned distance matrix **D** is converted to an aligned perspective matrix $\mathbf{P} \in \mathbb{R}^{nW \times N}$ where each entry represents the position of a cell along the trajectory from the viewpoint of the corresponding waypoint yielding nW proposed orderings for each cell. Note that the perspective of the early cell s is the initial ordering $\tau^{(0)}$ itself.

These perspectives can now be used to increase the accuracy of the ordering by computing a weighted average across the proposed orderings. The weighting scheme should increase the vote for closer waypoints to take advantage of reliability of shortest paths over shorter distances. However, it is important to also include a degree of influence from the distal waypoints to derive a consistent global structure. This requirement is satisfied by weights that are inversely proportional to the distance from the waypoint. Thus, the weights are calculated by a Gaussian kernel applied to the distances, as defined by

$$W_{wi} = \exp\left(\frac{-D_{wi}^2}{\sigma}\right) \Big/ \sum_{k=1:N} \exp\left(\frac{-D_{wk}^2}{\sigma}\right) \tag{3}$$

where σ is the standard deviation of distance matrix **D**. The denominator is the summation of inverted distances over all cells and used for normalization. This defines the weight matrix $\mathbf{W} \in \mathbb{R}^{nW \times N}$. The weighted average is then calculated by

$$\tau_i^{(1)} = \sum_{w \in Waypoint\ set} P_{wi} * W_{wi} \tag{4}$$

The vector $\tau^{(1)}$ is the refined trajectory of all cells (**Fig. 1b**, bottom right panel).

Note that the **W** matrix is adjusted to ensure waypoints on one branch have reduced influence on ordering of cells in the other branch by a muting scheme described in the section "Refining the ordering using branch association scores."

Branch point identification. Inconsistencies between waypoints are used to identify a branch point and the branch associations of each cell. Consider a waypoint t and a second waypoint i, with t being further along the trajectory. If i and t lie along the same trajectory (either both lie in the trunk, or i lies on the trunk and t on of one of the branches, or i and t are both on the same branch), the path from s to t will lie roughly along the path used for calculating perspective of t relative to i (**Fig. 1c**, left panel). Therefore, the perspective relative to t will be in agreement with (i.e., be similar to) the perspective of early cell s. Now consider another waypoint j such that j and t lie on different branches. In this case, there will be a disagreement between the perspective of s and t regarding the placement of j. The path from s to t will be substantially shorter than the path to determine the perspective of t with respect to j (**Fig. 1c**, middle panel). That path first involves a traversal to j from s and another traversal back through j's branch and then back out on t's branch to reach t.

Therefore, for any two waypoints, the mutual disagreement between the perspective of one waypoint relative to other and the early cell's perspective provides a quantitative measure of whether the two waypoints lie on same or different branches.

Mutual disagreement between a single pair of waypoints alone does not suggest a branch point, as such a disagreement could be caused by noise accumulated during longer walks. However, when a true branch exists, there will be disagreement between a considerable number of waypoints (those on different branches), which will cue the existence of a branching. In the case of branching, a clear structure emerges, where two groups of branch points A and B all disagree between waypoints across A and B and agree with waypoints within the same branch.

To identify this structure, Wishbone computes disagreements for all pairs of waypoints to construct the matrix $\mathbf{Q} \in \mathbb{R}^{nW \times nW}$, where

$$Q_{ij} = |P_{ij} - \tau_j^{(0)}| \quad (5)$$

In particular, $Q_{ij} \gg 0$ if the two waypoints, i and j, are on different branches and $Q_{ij} \approx 0$ if one or both are on the trunk, or both are on the same branch. In summary, the distance matrix **D** is used to determine a perspective matrix **P**, which in turn is used to both refine the order and calculate the disagreement matrix **Q** used to determine branch associations as described below.

Figure 1c (right panel) shows an example of the **Q** matrix. This matrix captures similarities and differences between waypoints that belong to the same or different branch, respectively. In particular, in the case of a branching trajectory, **Q** is effectively composed of three blocks. The first block consists of the waypoints in the trunk with $Q_{ij} \approx 0$ for trunk waypoint i and any other waypoint j. The remaining two blocks represent the two branches with $Q_{ij} \approx 0$ if i and j are on the same branch and $Q_{ij} \gg 0$ if they are on different branches. A natural way to identify these blocks or clusters is by use of unsupervised clustering methods.

Spectral clustering methods are a family of clustering algorithms designed to work on adjacency matrices representing graphs. Specifically, spectral methods are based on Eigen decomposition of the graph adjacency matrix and connections between the resulting Eigen vectors and properties of the graph structure. The **Q** matrix can itself be seen as an adjacency matrix with the disagreement representing the weight of the edge between waypoints and as such spectral clustering can be used to classify the waypoints into trunk and the two branches. For a real symmetric matrix, the second highest Eigen value, $\mathbf{v}_2$, approximates the optimal graph partition[34]. Specifically, if v_{2i} is the projection of the node i onto the second Eigen vector, the graph partition divides the nodes into two clusters whose elements can be identified by the sign of v_{2i}.

The **Q** matrix is real because all the perspectives and distances are real. It is also symmetric. Consider any two waypoints, i and j and assume i follows j without loss of generality. Then, $Q_{ij} = |P_{ij} - \tau_j^{(0)}| = |\tau_i^{(0)} + D_{ij} - \tau_j^{(0)}|$ and $Q_{ji} = |P_{ji} - \tau_i^{(0)}| = |\tau_j^{(0)} - Dji - \tau_i^{(0)}| = |-\left(-\tau_j^{(0)} + Dji + \tau_i^{(0)}\right)| = |\tau_i^{(0)} + Dij - \tau_j^{(0)}|$, since $Dji = Dij$. Therefore, **Q** matrix is symmetric. Thus, the second highest Eigen values of **Q** can be used to identify branch associations of waypoints. Wishbone uses the *Matlab eigs* function for Eigen value decomposition.

The second Eigen vector, $\mathbf{v}_2$, of **Q** matrix in **Figure 1c**, right panel is shown in **Figure 1d**. If waypoint w is on of the branches then $v_{2w} > 0$ *or* $v_{2w} < 0$ and $v_{2w} \approx 0$ if w is on the trunk, since $Q_{ij} \approx 0$ for all pairs of waypoints on the trunk. Moreover, $|v_{2w}|$ increases as waypoints progress further along the trajectory away from the branch point. This creates a Wishbone-like structure, giving the algorithm its name (**Fig. 1d**).

The v_{2w} values provide a reliable partitioning of waypoints that lie toward the end of developmental trajectory on either branch. However, the values of v_{2w} are noisy during transition from trunk to the two branches and care must be taken to pinpoint the branching point (**Fig. 1d**). Any path from a waypoint in a particular branch to a waypoint in the other branch will first traverse toward the trunk and then away from it. By definition, the point on path with minimum trajectory value $\tau^{(0)}$, represents the point at which the path will be closest be to the early cell. This also represents the point where the path changes direction to enter the other branch, or in other words an estimated branch point (**Supplementary Fig. 24a**). For increased robustness, Wishbone uses multiple paths between branch waypoints to estimate the branch point. Specifically, all the paths between the five furthest waypoints along the trajectory $\tau^{(0)}$, from each side of the $\mathbf{v}_2$ spectrum are determined. The position with shortest distance to s from each path is selected and the median position over all paths is an estimate of the branching point, *bp* (**Supplementary Fig. 24b**). Formally, let **BrA** and **BrB** be the set of five furthest waypoints of the two branches.

$$\begin{aligned} \mathbf{BrA} &= \{i \,|\, highest\ five\ \tau_i^{(0)}\ with\ sign(v_{2i}) > 0\} \\ \mathbf{BrB} &= \{i \,|\, highest\ five\ \tau_i^{(0)}\ with\ sign(v_{2i}) < 0\} \end{aligned} \quad (6)$$

These waypoints are then used to determined the branch point by

$$bp = median\left\{\tau_k^{(0)} : k = min_p\left\{\tau_p^{(0)} : p \in path(i,j)\right\}\right\} i \in \mathbf{BrA}\ and\ j \in \mathbf{BrB} \quad (7)$$

Branch assignment to cells. The v_{2w} values provide a partition of the waypoints that help determine branch assignments to all cells. Cells toward the end of the trajectory would have acquired most of the characteristics of the differentiated cell types and thus are relatively easy to classify. However, cells undergoing fate decision near the branch point do not always have a clear identity. This necessitates the use of a soft score of branch association for cells close to the branch point rather than hard branch assignments.

Wishbone uses the v_{2w} values to estimate a branch association score or *BAS* for each cell. The estimated *BAS* has the following properties: for all cells on the trunk, $BAS \approx 0$, whereas $BAS < 0$ or $BAS > 0$ for cells on the two branches. The deviation from zero is a measure of confidence for branch association. Note that the $\boldsymbol{v_{2w}}$ values already satisfy these properties. Therefore, *BAS* for each cell is determined by a weighted average of the $\boldsymbol{v_{2w}}$ values, with closer waypoints getting higher votes. First, the $\boldsymbol{v_{2w}}$ values are normalized for each branch to account for any trajectory value differences between the two branches:

$$v_{2w\,norm} = sign(v_{2w}) * \frac{abs(v_{2w})}{max(v_{2k}\,|\,sign(v_{2k}) == sign(v_{2w}))} \quad (8)$$

The weight matrix defined in equation (3) and $\mathbf{V_{2wnorm}}$ are then used to calculate *BAS* for each cell i

$$BAS_i = \sum_{w \in W} W_{wi} * v_{2w\,norm} \quad (9)$$

An example of scores for all cells is shown in **Supplementary Figure 25a**. These scores satisfy the properties outlined above and represent a soft association of branches to cells. For any downstream analyses, *BAS* scores and the branch point *bp* can be used to determine branch assignments for all cells. Cells before the branch point are considered part of the trunk and cells beyond the branch point are assigned to one of the two branches based on sign of *BAS*. Formally, branch assignment for cell i is determined as

$$c_i = \begin{cases} 1(Trunk)\,\text{if}, \tau_i^{(0)} \le bp \\ 2\ \text{if}, \tau_i^{(0)} > bp\ and\ BAS_i > 0 \\ 3\ \text{if}, \tau_i^{(0)} > bp\ and\ BAS_i \le 0 \end{cases} \quad (10)$$

The branch assignments for the illustrative data set are shown in **Supplementary Figure 24b**.

Refining the ordering using branch association scores. Following a branching point, waypoints that are part of one branch should not significantly influence the ordering of cells in the other branch. Wishbone achieves this by a cross branch-muting scheme that adjusts the weights defined in section "Waypoints and Perspectives" to ensure waypoints of a branch predominantly influence the ordering of cells in its respective branch. An example of weights before adjustment is shown in **Supplementary Figure 25a**. The left panel shows a waypoint in branch B that can influence the ordering of the mature cells in branch A.

The *BAS* scores described in the previous section can also be used for cross branch muting. Recall that weights that define the influence of

doi:10.1038/nbt.3569

waypoints on ordering cells is determined by a Gaussian kernel on the distance matrix **D** equation (3):

$$W_{wi} = exp\left(\frac{-D_{wi}^2}{\sigma}\right) \Big/ \sum_{k=1:N} exp\left(\frac{-D_{wk}^2}{\sigma}\right)$$

For any cell i, the sign of BAS_i defines branch membership. Therefore, for a given waypoint w, if the sign of BAS_w is not the same as the sign of BAS_i, the weight W_{wi} must be muted to reduce the influence of w in ordering the cell i. The extent of muting is directly proportional to the deviation of BAS_i from 0 and ensures that the influence of waypoints on one branch for ordering cells on the other branch progressively reduces along the developmental trajectory.

For each cell i, the weights of waypoints that belong to a different branch are muted as below

$$Mut_b = max(exp(-|BAS_k|\,|\,sign(BAS_k) \neq sign(BAS_i))), waypoint\ k$$
$$Wmut_{wi} = \begin{cases} W_{wi}, if\ sign(BAS_w) == sign(BAS_i) \\ W_{wi} * max(exp(-|BAS_i|), Mut_b), otherwise \end{cases} \quad (11)$$

This muting scheme exponentially reduces the influence of waypoints that do not belong to same branch. The weights after muting are shown in **Supplementary Figure 25b**. The waypoint in branch B no longer influences the ordering of mature cells in branch A (**Supplementary Fig. 25b**, left panel). The muting does not affect the weights of waypoints outside branches (**Supplementary Fig. 25b**, right panel).

Finally, as mentioned before, the refined trajectory is calculated by the weighted average over all the perspectives using the muted weights as

$$\tau_i^{(1)} = \sum_{w \in Waypoint\ set} P_{wi} * Wmut_{wi} \quad (12)$$

Iterative refinement of trajectory and branching. The waypoints are themselves cells. Therefore, their position often changes following the refinement step. Since all cell positions depend on waypoint positioning, the shift in waypoints might obsolete the newly calculated ordering. Therefore, the refinement step is repeated with the new waypoint positions until the ordering of all cells converges.

As defined earlier, $\tau_i^{(0)} = D_{is}$. At any iteration t, the perspective and **Q** matrix are calculated by

$$P_{iw} = \begin{cases} \tau_i^{(t-1)} + D_{iw}\ \text{if}\ \tau_i^{(t-1)} < \tau_w^{(t-1)} \\ \tau_i^{(t-1)} - D_{iw}\ \text{otherwise} \end{cases} \quad (13)$$
$$Q_{ij} = |P_{ij} - \tau_j^{(t-1)}|$$

The **Q** matrix is then used to determine the branch point bp and the BAS scores. These scores are then used for cross branch muting and a refined ordering or trajectory at iteration t is determined by

$$\tau_i^{(t)} = \sum_{w \in W} P_{iw} * Wmut_{iw} \quad (14)$$

This procedure is repeated until convergence: $corr(\tau^{(t)}, \tau^{(t-1)}) > 0.9999$. Finally, the branch assignments are calculated using the branch point bp and BAS scores. Note that the most time consuming parts of Wishbone are the construction of nearest-neighbor graph **G** and computation of the shortest path distances to all cells from waypoints to build the distance matrix **D**. These are both one-time steps and are not repeated during the iterations. Moreover, besides from graph construction, computation of **D** matrix is the most computationally intensive task and is performed only once. On the other hand, **P** and **Q** matrices, determined at each iteration, are not computationally intensive.

In summary, Wishbone aligns cells along bifurcating developmental trajectories in high resolution with accurate detection of the bifurcation point. The graph-based approach used by Wishbone is central is achieving the high resolution. Diffusion maps help overcome short-circuits, a key hurdle of the graph-based approaches for constructing trajectories. Waypoints and their perspectives not only help alleviate the additive noise of shortest path distances but also provide the basis for identifying branch associations by means of disagreements between waypoint and early cell perspectives. Finally, spectral clustering methods are used on these disagreements to determine branch association scores and refine the ordering. The trajectory detection and branch associations are repeated until convergence. **Supplementary Note 4** shows the pseudocode of the Wishbone algorithm.

Selection of diffusion components. Diffusion maps decompose the data along the major axes of variation and capture the major structures in the data[7]. In mass cytometry, this is reflected by the first few components capturing the differences among constituent cell types provided that the markers were chosen appropriately. The subsequent Eigen vectors typically capture noise and/or outliers. As an example, the top diffusion components for mouse thymus replicate 1 is shown in **Supplementary Figure 6a**. The first component is trivial with same value for all the cells and is associated with Eigen value of 1. Components 2,3 and 4 identify the differences among the constituent DN, DP and two SP cell types. The fifth component and beyond do not explain major structure in the data and encode for outliers and/or noise. Thus, Wishbone was run using components 2,3 and 4.

While this procedure does require manual selection of components, the first two to four non-trivial components typically explain the differences between cell types in data sets with trajectories with two branches (**Supplementary Figs. 6a**, **14b**,**c**, and **15b**), making the selection feasible. Moreover, Wishbone is robust to the inclusion of a number of noisy higher order components (**Supplementary Fig. 6b**), with very similar results achieved when including any number between 3 to 9 of the top components. A degree of automation can be achieved by examining the distribution of Eigen values of the diffusion components, and selecting the Eigen vector with biggest Eigen gap (difference between successive Eigen values) among first few components. The Eigen value distribution of the mouse thymus replicate 1 is shown in **Supplementary Figure 6c** and shows that there is a large Eigen gap between the 4th and 5th Eigen values. This is consistent with the observation of fifth and higher components encoding outliers and noise, there by justifying the use of the components 2, 3 and 4 for learning trajectories.

In single-cell RNA-Seq, diffusion components not only explain the differences between cell types, but also identify the variation along various biological processes like metabolism, cell cycle etc. See section "Application of Wishbone to single-cell RNA-Seq data" for component selection procedure for single-cell RNA-Seq.

Marker expression along trajectory and derivative plots. The trajectory was first divided into 150 equally spaced bins. A Gaussian filter centered at each bin was used to estimate the weighted average expression of individual markers in each bin. The density of cells is non-uniform along the trajectory and binning the trajectory for estimating average expression rather than moving average captures the density differences. The weight matrix $\mathbf{K} \in \mathbb{R}^{150 \times N}$ is determined as follows:

$$K_{bj} = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(\tau_j^{(t)} - b_\tau)^2}{\sigma^2}\right) \quad (15)$$

where b_τ is the mean trajectory value in bin b and σ is the standard deviation of the trajectory. The weighted average expression of a marker in each bin is then calculated as

$$E_b = \sum_{i=1}^{N} K_{bi} * M_i \quad (16)$$

where M_i is the marker expression in cell i. The weighted standard deviation for each bin is also calculated along similar lines.

For bins past the branch point, the weighted expression on a particular branch is determined by muting the weights of cells on the other branch. Markers with a weighted average difference of at least 0.1 in any bin beyond the branch point are plotted with dotted lines representing the expression in the two branches.

After calculating of the weighted averages, the derivatives were calculated as the difference in weighted average in successive bins.

Cross correlation of trajectories. For a given marker, cross correlation of expression along trajectories of different replicates were determined. The trajectory was shifted to maximize the mean of all cross correlations.

Variance analysis. The two SP populations were identified by the gating scheme defined in **Fig. 3b**. Population level s.d. was calculated for each marker in these gated populations. The calculation of s.d. along trajectory is described in "Marker expression along trajectory and derivative plots."

For running Wishbone after exclusion of a particular marker, diffusion maps were first used to determine low dimensional embedding of the phenotypic space without the marker. Wishbone was then run on the embedded space with the same parameters used for the runs with all markers.

Trajectories in gated populations and comparison to ImmGen. Gating of SP cells was performed using the scheme recommended by the Immunological Genome Project (ImmGen)[13]. While ImmGen used Forward and Side Scatter channels to remove non-lymphoid cells, we used mass channels, which measure non-lymphoid cell surface markers and removed these cells using the clustering method Phenograph (see "Data preprocessing and choice of parameters for Wishbone").

Raw mRNA expression data were downloaded from the Immunological Genome Project website (GEO accession number: GSE15907). These data were background corrected using RMA and quantile normalized using the affy R Bioconductor package[35]. The expression of each gene was then scaled to be between 0-1 among the different sorted T-cell populations: T.DP69^{+}, T.4^{+}8^{int}, T.4SP69^{+}, T.4SP24^{int}, T.4SP24^{-}, T.4int8^{+}, T.8SP69^{+}, T.8SP24^{int}, T.8SP24^{-}.

Adaptation of Wishbone to single-cell RNA-Seq data. *Data processing.* The count matrix was downloaded from GEO (GSE72857)[6]. As a first step, cells with less than 250 molecules were discarded. Library size correction was performed by dividing the molecule counts of each cell by the library size[36]. The corrected molecule counts were then multiplied by the median of the library size across cells[36]. To address gene drop-out[27], the data were transformed using PCA to identify "meta-genes." We note that while phenotypic space defined by cells is nonlinear in its nature, the relationships between genes are largely collinear, making PCA appropriate on the gene dimension. These meta-genes were then used to cluster cells using Phenograph[28] and the clusters corresponding to HSPCs, erythroid precursors and myeloid precursors were identified by expression of characteristic genes: HSPCs - Cd34, Erythroid precursors - Gata1, Gata2, Myeloid precursors - Mpo, Csf1r, Irf8 (**Supplementary Fig. 15a**).

Selection of diffusion components. Diffusion maps decompose the data along the major axes of variation and capture the major structures in the data[7]. In mass cytometry data with a relevant marker set, this amounts to accounting for differences in the constituent cell types (**Supplementary Fig. 6a**). In genome-wide data, many of the components reflect additional biological processes such as cell cycle, stress and metabolism that would confound building trajectories. Therefore, we identify the biological processes associated with each diffusion component and keep only those that are related to development and maturation.

To identify the biology associated with each component, we sought to find genes whose expression pattern was correlated with the component. Mean expression in sliding windows of 10 cells along the component was used to determine the correlation between each gene and component. Gene Set Enrichment Analysis (GSEA)[20] was performed using the correlation based ranking to annotate each component. Gene sets from Gene Ontology Biological Process[37] database were used for annotations. Once the diffusion components are annotated, we can take one of two approaches, in sufficiently studied systems we can positively select the relevant components. In less studied systems, we can simply remove confounding components such as cell cycle, ribosomes and metabolism.

In the application here, the top 15 principal components were used for constructing the diffusion maps and the resulting enrichments for the top diffusion components of the single-cell RNA-Seq data set are shown in **Supplementary Figure 15c**. Components 2 and 3 are enriched for ontologies related to immune cell differentiation. Wishbone was run using the components 2 and 3 with a randomly selected cell from the HSPC cluster as the input early cell. We note that both trajectory and branches are robust to the number of principal components used (**Supplementary Fig. 15d**).

Diffusion maps, Monocle, and SCUBA. Diffusion maps provide low dimensional projections of the phenotypic space. Euclidean distance between the points in their low dimensional embedding onto the diffusion map is equivalent to their diffusion distance[7]. Therefore, we used the Euclidean distance from the start cell in the space spanned by the first three non trivial diffusion map Eigen vectors to construct the developmental trajectories in mouse thymus and human myeloid mass cytometry data sets (**Fig. 6** and **Supplementary Fig. 16**). Diffusion components used for Wishbone were also used to estimate the distances from the start cell in the mouse myeloid single-cell RNASeq data set (**Fig. 6**).

Monocle was downloaded from Bioconductor[24]. 1,000 cells were randomly sampled from each of the data sets and Monocle was run with default parameters apart from number of branches, which was set to two and the root cell was set to the start cell used for Wishbone. The results in **Supplementary Figure 18a–c** were obtained by repeatedly sampling 1,000 cells from replicate 1 with the number of branches set to 2. The start cell was set to the same start cell used for Wishbone. Different runs were compared using the procedure described in section "Cross correlation analysis."

SCUBA was downloaded from https://github.com/gcyuan/SCUBA (ref. 5). 20,000 cells were randomly sampled from the mouse thymus and SCUBA was run with default parameters. The MassCytometry_preprocess.m script was used. The expected input, processDataMat Matlab data matrix was created and the ordering and branches were determined using the function EstimatePseudotime. As with Monocle, the results in **Supplementary Figure 17a–c** were obtained by repeatedly sampling 20,000 cells from replicate 1 and the different runs were compared using the procedure in "Cross correlation analysis." SCUBA was run using all cells for the human and mouse myeloid data sets.

The resulting trajectory and branches from both Monocle and SCUBA were visualized using tSNE (*t*-distributed stochastic neighbor embedding) projections of the full data set. Marker expression along the Monocle trajectories was determined using the procedure described in "Marker expression along trajectory and derivative plots." A similar procedure was used for SCUBA with muting turned off since SCUBA resulted in more than two branches.

Data availability and software. The **mouse** thymus mass cytometry data can be downloaded from Cytobank (https://www.cytobank.org/cytobank/experiments/52942). The cleaned data along with the Wishbone results for different replicates are available at http://www.c2b2.columbia.edu/danapeerlab/html/wishbone.html. Myeloid mass cytometry data were downloaded from Cytobank (http://reports.cytobank.org/1/v1). Mouse myeloid single-cell RNA-Seq data were downloaded from GEO (accession number: GSE72857). Wishbone results for all the myeloid data sets are available at http://www.c2b2.columbia.edu/danapeerlab/html/wishbone.html. Wishbone has been integrated into our single analysis suite *cyt* and can be downloaded from (http://www.c2b2.columbia.edu/danapeerlab/html/cyt-download.html) and from **Supplementary Software**. A python package for Wishbone algorithm is available through github (https://github.com/ManuSetty/wishbone).

28. Levine, J.H. *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
29. Waddington, C.H. *An Introduction to Modern Genetics* (George Allen & Unwin, 1939).
30. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
31. de Silva, V. & Tenenbaum, J.B. Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems* **15**, 721–728 (2003).
32. Amir, A.D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
33. Gut, G., Tadmor, M.D., Pe'er, D., Pelkmans, L. & Liberali, P. Trajectories of cell-cycle progression from fixed cell populations. *Nat. Methods* **12**, 951–954 (2015).
34. von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
35. Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).
36. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
37. Huang, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

doi:10.1038/nbt.3569