Cell lineage inference from SNP and scRNA-Seq data

Jun Ding [1], Chieh Lin[2] and Ziv Bar-Joseph[1,2,*]

[1]Computational Biology Department, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA and
[2]Machine Learning Department, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

ABSTRACT

Several recent studies focus on the inference of developmental and response trajectories from single cell RNA-Seq (scRNA-Seq) data. A number of computational methods, often referred to as pseudo-time ordering, have been developed for this task. Recently, CRISPR has also been used to reconstruct lineage trees by inserting random mutations. However, both approaches suffer from drawbacks that limit their use. Here, we develop a method to detect significant, cell type specific, sequence mutations from scRNA-Seq data. We show that only a few mutations are enough for reconstructing good branching models. Integrating these mutations with expression data further improves the accuracy of the reconstructed models. As we show, the majority of mutations we identify are likely RNA editing events indicating that such information can be used to distinguish cell types.

INTRODUCTION

Several recent methods have been developed to infer psuedo-time and branching trajectories from time series single-cell RNA-seq (scRNA-Seq) data (1–10). These methods rely on the assumption that cells that are in a similar state (developmental time, fate etc.) are also close in expression space. Based on these assumptions psuedo-time methods construct models based on minimum spanning trees (MST) (2,9), clustering (4) or other graphical models (1,5,11) to connect cells that share the same state and identify branching events that lead to different cell fates. Such methods have been successfully applied to study several developmental and response processes including lung (7,10), neuron (7,10), myeloid (3,5), heart (12) and liver development (13), various treatment responses (14), aging (15) and more. While scRNA-Seq expression information is useful, it is also very noisy. Further, some recent studies indicate that a small subset of the genes, which are sometimes expressed at very low levels and so do not sig-

nificantly impact overall expression similarities, can have a large impact on changing cell states (16). Indeed, in several cases the relationships identified by the pseudo-time ordering methods do not accurately capture known biological trajectories (10).

In addition to methods that rely on expression data, several genetic based lineage tracing methods have been developed over the last two decades, though these have not been combined with scRNA-Seq analysis (17). Recently, a number of methods that combine scRNA-Seq with Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technology for lineage tracing were developed (18,19). These methods are based on the insertion of random mutations during cell division to a pre-determined RNA. Once RNA is sequenced the set of random mutations can be traced backwards to construct a phylogenetic tree which can then be used to assign cells to branches and fates. Such CRISPR based lineage tracing methods have been recently applied to study zebrafish development (18) and to study the lineages in mouse embryonic cells (20). Results indicate that for short durations (until the RNA mutations saturate) such method can indeed lead to good results when attempting to infer cell branching.

While CRISPR based methods are useful, they are limited in several ways (17,21). First, it is not clear how such method would be applied to higher organisms in vivo, especially when studying diseases and responses in humans. Second, the method requires genetic interventions which may alter wild-type behavior. Finally, the method to date is limited to short durations (given the length of the sequenced region) and so may not be appropriate for all studies.

An alternative to using CRISPR is to rely on de novo mutations. These have several advantages since they do not require any engineering, are not restricted in time and can be used for all species. The major challenge for using such an approach is the fact that such mutations are rare. For example, the mutation rate is $\sim 1.1\mathrm{e}{-8}$ per site per generation (22) in human, which is ~ 35 mutations genome-wide per generation and so it is unlikely that many of them would be encoded in the coding regions that are profiled by scRNA-Seq. However, *de novo* mutations are only one reason why RNAs can differ between cells. Another reason is RNA edit-

*To whom correspondence should be addressed. Tel: +1 412 268 8595; Email: zivbj@cs.cmu.edu

ing, which is a molecular process through which some cells make discrete changes to specific nucleotide sequences with an RNA molecular after it has been generated by RNA polymerase and was previously reported to involve in the cell differentiation process (23). Gommans *et al.* reported that RNA editing is a highly regulated process in higher organisms with editing levels specifically changing during development. They further claim that editing is often cell and tissue specific (24). Gagnizde *et al.* reported that A-to-I patterns reveal specific editing signatures distinguishing major cell types in the human brain, among which neuron and astrocytes constitute the most edited cell types (25). Combined, *de-novo* mutations and RNA editing can provide additional information that is not captured by the expression profiles themselves to aid in reconstructing the branching trajectories. To enable the use of such sequence information when reconstructing dynamic differentiation models from scRNA-Seq data we developed a new method for Trajectory inference Based on SNP information (TBSP) that identifies such mutations (we refer to them as SNPs though several are likely due to RNA editing as we discuss below). Once significant SNPs have been identified we use them to reconstruct a phylogenetic tree for the cells profiled. We show that the tree agrees quite well with known cell states for these cells even though we did not use the expression levels themselves to construct it. Next, we extend a previous method we developed to reconstruct dynamic models of cell differentiation so that it can utilize both expression and SNP data. As we show, the reconstructed models that utilize the SNP data further improve upon models generated by only using the expression level data indicating that SNPs provide information that is not captured by the expression levels themselves. We also discuss the biological meaning of the SNPs and argue that many of them are likely RNA editing events rather than *de-novo* mutations.

MATERIALS AND METHODS

TSBP starts by filtering potential SNPs in the scRNA-Seq data. Next cells in the input data are clustered using all significant SNPs. Starting from the initial clusters, we iterate (in an EM-like algorithm) between cell assignments to clusters and SNP selection until convergence or the maximal iterations. The final selected SNPs are used by TBSP to calculate distances between cells (based on the Hamming distance). Neighbor-joining is then used to construct trajectories for the cells based on the calculated distance matrix. Finally, TBSP combines SNP and expression data to provide a more comprehensive view of the developmental or progression trajectories. Figure 1 presents an overview of TBSP.

Detecting SNPs from single-cell RNA-seq data

We mapped all the scRNA-Seq reads to the reference genome using HISAT2 (26) with the default parameters. We next used the GATK variant-calling pipeline (27,28) to call all potential SNPs for each of the cells. The obtained SNPs were filtered by the VariantFiltration function included in the GATK pipeline with the recommended parameters (including QD < 2.0 which is cutoff recommended for obtaining significant variants). We further filter SNPs found in
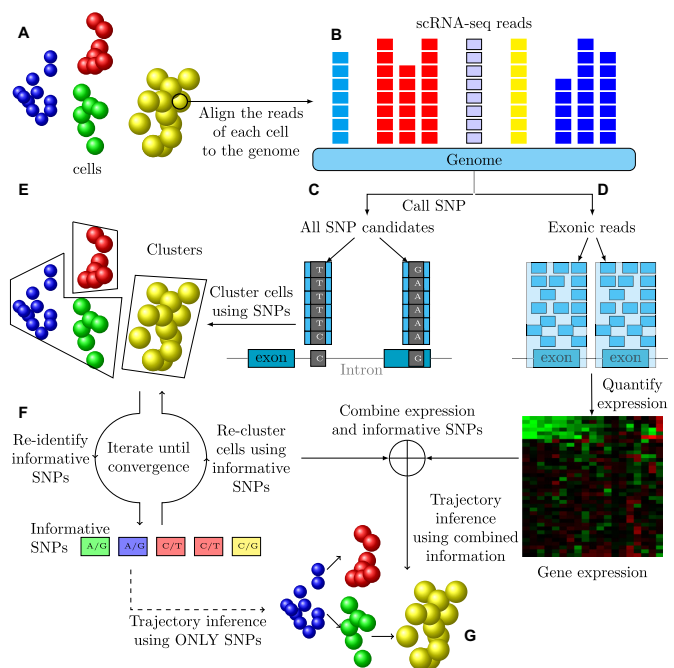

Figure 1. TBSP Method Overview. (**A**) Cells used in the study. (**B**) Reads are mapped to the reference genome. (**C** and **D**) Reads are used to determine expression levels and to identify SNPs. (**E**) Cells are clustered based on identified SNPs. (**F**) Iterating between selecting a subset of key SNPs and clustering using selected SNPs. Once a set of key SNPs is established, it is combined with expression values to determine the branching model. (**G**) Final predicted trajectories (using SNPs and/or expression information).

<10% of the cells (rare SNPs) or >80% of the cells (Universal SNPs). Rare SNPs are most likely false positives. On the other hand, Universal SNPs (which we term *baseline SNPs*) are uninformative and likely represent differences between the cell line or animal used for the experiment and the reference. We also tried other cutoffs such as 20% for Rare SNPs, and found the results to be very similar (Supporting Methods and Supporting Figure S1)

In addition to baseline SNPs that appear in a large fraction of the cells, the method can also identify baseline SNPs in a fraction of the cells. This would happen if the gene in which this SNP resides is only expressed in a subset of the cells. Such SNPs are redundant with gene expression data and so do not provide any additional information. To remove these, we only use SNPs identified in regions where we have multiple aligned reads in most cells (>8 reads on average aligned in >80% of the cells). Since we find that most significant SNPs are only identified in a small fraction of cells (much smaller than 80%) such requirement means that identified SNPs represent real differences between the cells.

Identifying informative SNPs for trajectory inference

We first build a cell-SNP matrix (M) for all cells (denoted by C) and all SNPs after the initial filtering (denoted by X), where $M(i, j)$ is a binary value, which tells whether SNP j is detected in cell i. As the initial set of SNPs X could possibly contain many false positive or non-informative SNPs, our objective here is to find the best subset of SNPs P from X,

which can best distinguish cells of different sub-types.

$$M(i, j) = \begin{cases} 1, & \text{if } X[j] \text{ detected in } C[i] \\ 0, & \text{else} \end{cases} \quad (1)$$

We initialize the cell clusters (denoted by N) using K-means (29) on rows (cells) of the cell-SNP matrix $M(i, j)$ where $i \in C, j \in X$. The number of clusters is determined using Silhouette score (30). To find the most discriminative SNPs among the full set, for each cluster I in the N, we search for the best SNP set $P(I)$.

To find such SNP subset $P = \bigcup_{I \in N} P(I)$, we use an EM-like algorithm that tries to infer the best SNPs for splitting the data into k groups: for each sub-population (I) (cluster) in the data, we identify the set of SNPs that best separates the cells in I from all other cells ($C - I$), where C represents all the cells in the data.

$$P(I) = \text{Argmax}_{x \in X} \frac{f(x, I)}{|I|} - \frac{f(x, C - I)}{|C - I|} \quad (2)$$

$$f(x, \text{cells}) = \sum_{i \in \text{cells}} M(i, x) \quad (3)$$

$$P = \bigcup_{I \in N} P(I) \quad (4)$$

where N denotes all the sub-population (clusters) in the data and $P(I)$ denotes the signature SNPs for $I \in N$. Since this is a challenging combinatorial problem we use a greedy algorithm to find a local optimal solution. See Supporting Methods for complete details.

The initial P matrix might contain both false positive and non-informative SNPs. To refine the matrix and the clustering, we iterate using the identified SNP set. In each iteration, we re-cluster the cells and use the new clusters to re-identify SNPs. This is repeated until convergence or when the maximal iterations are reached. When the algorithm converges, we are left with a selected set of SNPs (P). Similar to all clustering methods this approach can be sensitive to the initial clustering result and so we include additional SNPs if they improve the Silhouette score. See Supporting methods for details.

Inferring the trajectory using identified SNPs

Using the selected SNPs we utilize a distance-based neighbor-joining algorithm (31) to construct an initial trajectory. Each cell is represented using a binary SNP vector $V_i = [M(i, j)|\forall j \in P]$ where $M(i, j)$ was defined in equation (1). The distance between two cells (a, b) is calculated using hamming distance $d(a, b) = \sum_{j=1}^{|P|} |V_a(j) - V_b(j)|$. The distance between two clusters (I_i, I_j) is calculated as the average distance between every pair of cells in the two clusters $d(I_i, I_j) = \frac{\sum_{a \in I_i, b \in I_j} d(a,b)}{|I_i| * |I_j|}$. These cluster distances are used as the input to the Neighbor-joining algorithm to build the cell trajectory.

Integrating the identified SNPs with the expression-based trajectory inference

The SNPs we selected provide information that is complementary to the profiled scRNA-Seq expression data. We have thus next integrated SNP information with our previously developed method for reconstructing dynamic regulatory networks from scRNA-Seq data, scdiff (10,32). As the single-cell RNA-seq data is very noisy, it's not accurate to estimate the gene expression in each cluster (state) based on only the direct single-cell RNA-seq measurement. An additional source of information is needed to overcome the noisy nature of the scRNA-Seq data. In scdiff, we use the TF-gene regulatory networks as extra information, which impacts the state transition of the underlying Kalman Filter model. scdiff starts with building the initial tree-structured trajectories by clustering the cells and connecting the cell clusters. Next, scdiff iteratively refine the initial trajectories by integrating the extra TF-gene regulatory networks. It re-estimates the gene expression for each state (cluster) of the trajectory tree using Kalman Filter, which utilizes both the direct scRNA-Seq observation (emission model) and also the TF-gene regulatory information (transition model). With the re-estimated expression for each cluster, all the cells will be re-assigned, and the trajectories will be re-inferred based on the new cell assignments. Such process will be iterated until convergence or maximal iterations. The converged trajectories will be the final predictions together with predicted TFs, which are critical to the state transition in the Kalman Filter.

Here, We integrated the SNP information into scdiff. First, we build up the initial trajectories in the same way as scdiff. Next, we use not only the expression information (and the TF-gene regulatory information) as in scdiff but also the SNP information to refine the initial trajectories. We use the clusters from the initial trajectories to identify informative SNPs (though we do not iterate, just use the greedy heuristic on a fixed set of clusters) list(P) to help re-assign cells to states. This new assignment combines the expression profile and SNP for each cell c_i as follows:

$$\text{Assign}(c_i) = \text{argmax}_s P(c_i, s) \quad (5)$$

$$= \text{argmax}_s P(s) P(c_i|s) \quad (6)$$

$$= \text{argmax}_s P(s) * ps(c_i|s) * pe(c_i|s) \quad (7)$$

$$= \text{argmax}_s log P(s) + \log ps(c_i|s) + \log pe(c_i|s) \quad (8)$$

where $P(c_i, s)$ represents the probability of c_i in cluster s and $pe(c_i|s)$ denotes the conditional probability of c_i in cluster s based on expression, which is calculated in the same way as in scdiff (10). See supporting methods for how we compute $ps(c_i|s)$, the conditional probability of cell c_i in cluster s based on SNP information. The initial trajectories will be refined by the aforementioned cell assignments and the initial clusters will be also updated. Next, we re-identify the informative SNPs using the same method described above on the updated clusters. We iterate the above process until convergence or maximal iterations. The converged trajectories will be the final predictions. Here, we demonstrated the integration of the SNP information using the scdiff method.

However, such SNP information can also be integrated into other existing single-cell expression based methods. TBSP provides a Cell-SNP matrix, which describes the signature SNP vector for each of the cells in the dataset and thus can be used to refine the cell assignments/trajectories in the existing expression-based methods.

Post analyses of the predicted SNPs

We used PAVIS (33) to annotate the genomic position (Exon, Intron, 3'UTR and so forth) of the predicted SNPs. If the predicted SNP is located in the promoter (+5 kb), gene body or downstream (1 kb) of a gene, such gene will be regarded as the SNP target. We use PANTHER (34) GO enrichment tool to analyze the GO terms associated with the targets genes of the predicted SNPs.

Software availability

The TBSP software is freely available on GitHub at https://github.com/phoenixding/tbsp with a detailed user manual and a test set.

RESULTS

Differentiation trajectories can be inferred based on SNP information

We first tested the reconstruction of temporal and spatial trajectories using only the SNP information. For this we used several different scRNA-Seq datasets including the 'Neuron' scRNA-Seq expression data which studies neuron reprogramming (35), the 'Liver' scRNA-Seq data which studies human liver bud development from pluripotency (13) in 2D culture and 3D liver buds (LB) and a 'Lung' scRNA-Seq expression dataset which profiles distal lung epithelium differentiation (36). The Neuron dataset has four time points and a total of 252 cells. The Liver dataset has 765 cells, which falls into four stages ($iPSC \rightarrow DE \rightarrow HE$, $IH \rightarrow MH$, LB and others). The Lung dataset has three time points and a total of 152 cells. For all these datasets the original papers provide some information (based on known markers) about the expected trajectories or organizations, and these can be used to test the accuracy of the SNP based analysis.

For the mouse 'Neuron' data (35), the SNP-based trajectories inferred by TBSP are consistent with current knowledge (Figure 2). SNP data was informative enough to correctly cluster the cells (Supporting Table S1). Next, we looked at the trajectory inferred from these SNPs. As can be seen in Figure 2A, the model correctly starts with Cluster 1 (Mouse Embryonic Fibroblasts-MEF) and then continues to d2_intermediate (Cluster 2), d2_induced, d5_intermediate and d5_failedReprog d5_earlyiN and Neuron. This ordering is very similar to the one presented in (35) based on the expression of marker genes. Cluster 0 (d2_induced dominated) seems to be misplaced in the trajectory (earlier than d2_intermediate). However, the trajectories inferred using all SNPs, including those potentially redundant ones, display the correct assignment in which the d2_induced cells are assigned to follow d2_intermediate cells (Supporting Figure S2).
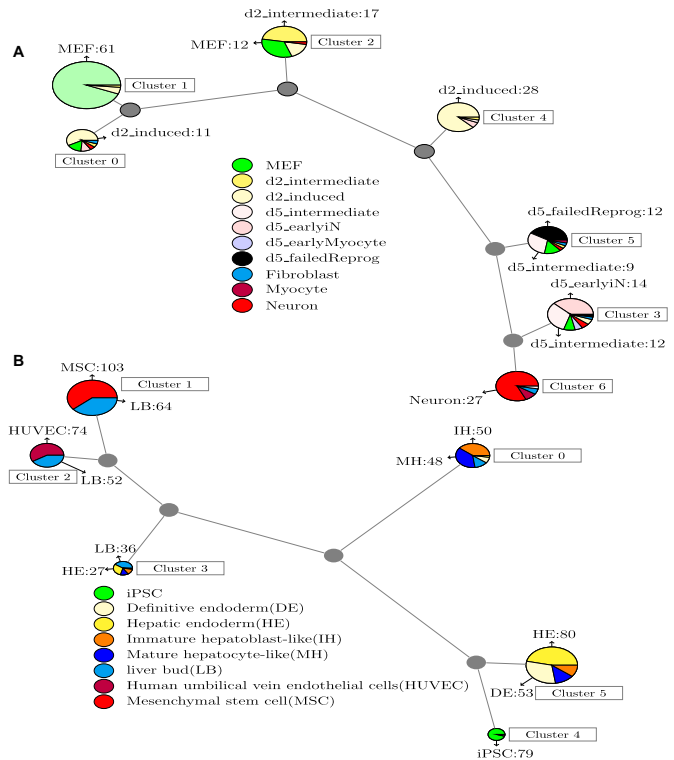

Figure 2. Predicted models using SNP information. To use the SNP data for inferring trajectories we relied on a well-established phylogenetic method: neighbor-joining. We interpret the resulting branching models as cell trajectories. In the figure, gray circles represent internal nodes in the phylogenetic tree that are not assigned any cells. The length of the lines in the constructed lineage graph represents the phylogenetic distance between nodes (clusters). The circle size represents the number of cells within the cluster. (**A**) Predicted model for the Neuron data (35). The model correctly starts with Cluster 1 (Mouse Embryonic Fibroblasts-MEF) and then continues to d2_intermediate (Cluster 2), d2_induced (Cluster 4), d5_intermediate and d5_failedReprog (Cluster 5), d5_earlyiN (Cluster 3) and Neuron (Cluster 6). This trajectory is very similar to the one presented in the original paper. (**B**) Predicted model for the Liver data (13). Similar to the mouse neuron data, cells are clustered well using only SNP information. As for the trajectory analysis, the original study (13) reported a bifurcation in 2D and 3D trajectories. In the 2D culture, the iPSC cells differentiate to mature hepatocyte-like (MH) cells, which are different from the liver bud (LB) and mesenchymal stem cell (MSC)-LB cells in their 3D differentiation counterparts. This is also the branching determined based on the SNP information.

For the human Liver differentiation data (13), the SNP-based trajectories also agree with marker based reconstruction (Figure 2B). First, as with the mouse neuron data, cells are clustered well using only SNP information (Supporting Table S1). As for the trajectory analysis, the original study (13) reported a bifurcation in 2D and 3D trajectories. In the 2D culture, the iPSC cells differentiate to mature hepatocyte-like (MH) cells, which are different from the liver bud (LB) and mesenchymal stem cell (MSC)-LB cells in their 3D differentiation counterparts. This is also the branching determined based on the SNP information by TBSP. SNP based models mix IH and MH cells, which is inconsistent with prior knowledge that indicates that IH cells are progenitors of MH cells. However, such a mixture of MH and IH cells also be observed in the Monocle results (Supporting Figure S3), which indicates that the cell

states of the IH and MH may be quite similar. Based on both TBSP and Monocle models, HE and a small fraction of early IH cells seem to serve as progenitors for the branching.

For the mouse 'Lung' differentiation data (36), the SNP-based predicted trajectories are also partially supported by the known model (Supplement Supporting Figure S4). The first time point (E14.5) is associated with a number of unique clusters (1, 3 and 4) residing at the beginning of the tree while more mature epithelial cells (mainly Bi-potential Progenitors (BP), Alveolar Type 2 and ciliated cells) are clustered together afterward, and the last to branch are Type 1 cells. On the other hand, the SNP-only model incorrectly assigns the E16.5 time point to a later branching location than its actual position in the process. Still, when the SNP information is combined with expression data, it still improves the accuracy of the reconstructed trajectories as we discuss below and in Supporting Results.

We also compared TBSP with Monocle2 (3). These comparisons demonstrated the advantages of our SNP-based methods as shown in Supporting Figure S3. For example, when analyzing the neuron reprogramming dataset, d2_induced cells are displayed on a separate branch in the Monocle 2.4.0 results while neuron cells are on the top branch. In contrast, the reconstructed model based on SNP data correctly connects them. Please refer to the Supporting Results for the complete details on all 3 datasets.

Unique SNPs are associated with the different clusters

The reconstruction above used 36, 55 and 33 SNPs for the Neuron, Liver and Lung data respectively. As can be seen in Figure 3, several of these are associated with one or only a few of the clusters in each model. For example, in the liver data, SNPs 31-38 are only enriched in Cluster 2 (Human umbilical vein endothelial cell (HUVEC) cells). SNPs 49-52 are only enriched in Cluster 0 (MH,immature hepatoblast-like (IH) cells) and SNPs 15–20 are only enriched in Cluster 4 (iPSC). Clusters of cells that are connected in development (for example, one is right after the other) usually share overlapping SNP patterns. An example of this is liver Cluster 4 (iPSC) and Cluster 5 (definitive endoderm (DE), hepatic endoderm (HE)). See Supporting Results for detailed discussion. We have also plotted the SNP distribution along the known trajectories(Supporting Figure S5). For the neuron data, we see several mutations that are only associated with the initial state. We also see some cell types (for example d2_intermediate and d2_induced or d5_earlyiN and Neuron) that share the same mutations while other types (for example, Fibroblasts) do not. This supports their use in unsupervised model reconstruction. For the liver data, we observe the main difference between MSC, LB cells and IH, MH cells. iPSC cells, also display a subset of unique SNPs. For the lung data, we observe a clear separation between progenitor cells E14 and terminal cells (AT1, AT2, Clara, and ciliated).

Predicted SNPs may represent RNA-editing changes

As noted in Methods, we removed baseline mutations that are only identified because some genes are only expressed in the subset of the cells. All of the mutations we identified overlap genes that are expressed in the majority of cells. Given the relatively low genomic mutation rate and the small number of cell divisions in the data we studied (less than 10), it is unlikely that most of the SNPs we identified represent de novo mutations. We have thus looked for alternative explanations for the significant SNPs we found. One such possibility is RNA editing, which is a molecular process through which some cells make discrete changes to specific nucleotide sequences with an RNA molecular after it has been generated by RNA polymerase and was previously reported to involve in the cell differentiation process (23).

The predicted SNPs are dominated by A/G (A → G, G → A) and C/T (U) (C → T, T → C) substitutions (Figure 4 A). Note that these substitutions are quite similar. The direction of the A/G substitution (A to G or G to A) depends on the reference and the A/G and C/T are essentially the same substitution at different strands. In the neuron data, A/G and C/T substitutions account for 34.4% and 51.7% respectively(combined total of 86.1%). For the Liver data, A/G accounts for 39.2% and C/T accounts for 41.2%, (80.4%) and in the lung data, A/G and C/T substitutions account for 81.4% of all predicted SNPs. A/G and C/T substitution dominance were also shown for RNA-editing (37).

In addition to the type of substitution, their locations also match. Predicted SNPs are found mostly in non-coding regions (Figure 4B) especially in 3'UTR regions. For example, in the neuron data, 44.4% of the SNPs are found in the 3'UTR, 19.4% of the SNPs are found in the intronic regions, 2.8% upstream of the gene (5kb), 2.8% are found in the 5'UTR region and 5.6% are found in Exons, 25% are found in the other region including gene downstream(1kb) and intra-genic regions. See Supporting Results for the distribution in other datasets. This also agrees with the fact that most RNA editing sites are located in the non-coding region (37).

We found that predicted SNPs are located near the Alu elements, which is also observed for RNA-editing sites (38). We also looked at the intersection between SNPs identified by TBSP and previously identified RNA-editing sites from the RADAR database (39). For the human data (for which we have many more known sites compared to the mouse) we find that 3 of the 55 SNPs we identified for the liver data are found in RADAR (*P*-value: 2.4e−4). Note that current knowledge of RNA editing sites is still limited. Finally, we used an RNA-editing site prediction tool RED-MEL (40) to score each of the predicted liver SNPs. Twenty one out of 55 predicted SNPs are identified as RNA-editing sites by RED-MEL (*P*-value = 0). See Supporting Results for details.

To further investigate the origin of the SNPs identified, we have analyzed data from a study that jointly performed RNA and DNA seq analysis in single cells (41). We downloaded DNA-seq and RNA-seq for 112 cells from this study and used our method to identify mutations in both. Under the parameters used for other studies we analyzed, we obtained 7 DNA SNPs (all in non-coding regions) and 31 SNPs from RNA-seq. None of the 31 RNA-seq SNPs were found in the DNA-seq results. Therefore, these mutations are only identified at the RNA but not the DNA level supporting their likely assignments as RNA-editing events.
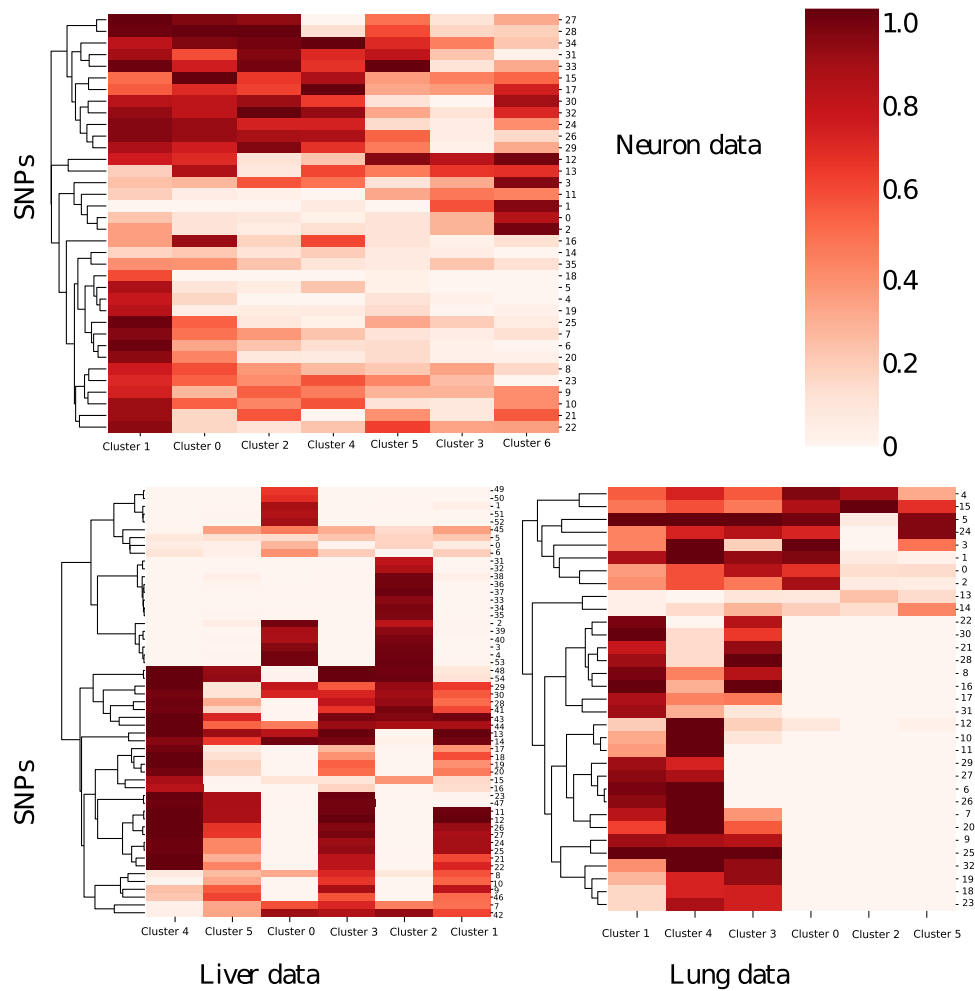
Figure 3. Distribution of predicted SNPs in clusters. All the clusters are ordered based on the trajectory inference. In many cases, we see SNPs at contiguous clusters which can explain their usefulness for reconstructing the trajectories of the different studies. Still, some SNPs (e.g. cluster 2 in the liver data) are very specific and only detected for one cell type.
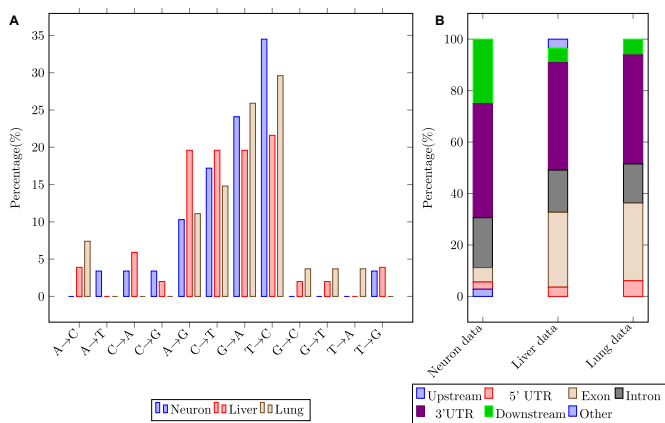

Figure 4. Predicted SNPs may represent RNA-editing changes. (**A**) Predicted SNPs are enriched with A/G(A→G,G → A) or C/T(C → T,T → C). Similar to several other studies that characterize RNA editing sites we find that SNPs detected by TBSP are enriched for specific substitutions. (**B**) The predicted SNPs are enriched in 3′UTR regions which is also where RNA editing sites are enriched in.

Please see Supporting Table S2 for the list of SNPs identified.

GO terms associated with the predicted SNPs

We also looked at the function of genes for which we identified SNPs in each of the datasets (Methods). In the neuron data, we found 26 such genes associated with the 36 predicted SNPs. The most significant GO terms associated with these 26 genes are 'Regulation of protein depolymerization' (P-value = 1.32e−4, FDR = 1) and "Regulation of protein complex disassembly" (P-value = 1.98e−4, FDR = 1), which are consistent with the potential protein degradation related functions of RNA-editing previously reported by study (42). In the Liver data, we found 42 genes associated with the 55 predicted SNPs. These 42 genes are enriched with 'protein targeting (P-value = 4.51e−9, $FDR = 7.05e−5$)' and 'contranslational protein targeting to membrane (P-value = 2.02e−08, FDR = 1.06e−4)', which is also supported by (42). For the Lung data, we found 24 genes associated with 33 predicted SNPs. The top GO

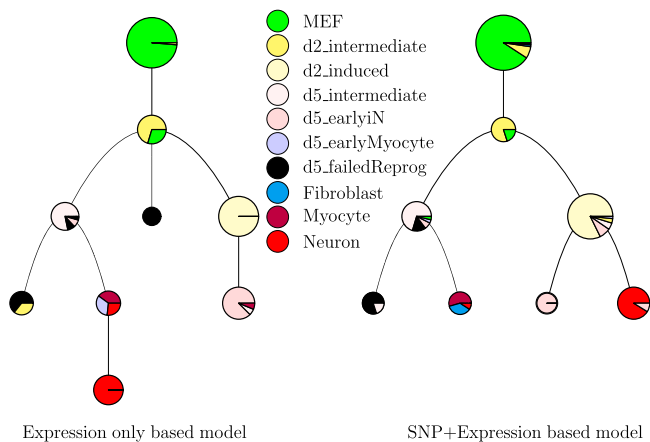Expression only based model   SNP+Expression based model

Figure 5. Combining expression data with SNP information to improve the reconstruction of branching models. Expression only and SNP added trajectory inference for the mouse Neuron data. (**Left**) A model reconstructed using only expression information for the neuron development data. The lowest cluster (red) is a descendant of the cluster with a mixture of Fibroblast, Myocyte and Neuron cells. In contrast, when using the SNPs. (**Right**) the neuron dominant cluster is descending from the d2_induced and d5_earlyiN dominant states. Also the d5_earlyiN cells are relatively closer to the Neuron cells, which is more consistent with the trajectory reported in (35).

terms for these 24 target genes are 'protein-containing complex assembly (*P*-value = 1.94e−4, FDR = 0.376)' and 'protein-containing complex subunit organization (*P*-value = 5.37e−5, FDR = 0.139)'.

SNP information further improved expression based pseudo time reconstructing methods

We next used the SNP data to improve the reconstruction of expression based trajectory inference methods. For this, we extended scdiff, a method we have previously developed to reconstruct trajectories from scRNA-Seq data (10). Briefly, scdiff is a probabilistic method that integrates expression data with TF-gene interaction data to learn a branching model and assign cells to states. We used the SNP data to further improve cell assignment and state inference (Methods). Results of the combined model are shown in Figure 5. As can be seen, for the Neuron data, the SNP based model leads to trajectories that are more consistent with prior knowledge (35). In the expression only model, the neuron (red) cluster is a descendant of the cluster with a mixture of Fibroblast, Myocyte and Neuron cells. In contrast, when using the SNPs the neuron dominant cluster is descending from the d2_induced, and d5_earlyiN dominant states. See Supporting Figures S6,S7 and Supporting Results for the additional analyses on other datasets.

To further study if SNP information is complimentary to expression data we combined the two and used the combined dataset as input for another trajectory inference method, Monocle (version 2.4.0) (3). For each cell we concatenated the SNP vector for that cell to the expression values to create a new input (Supporting Methods). We next ran Monocle on the SNP plus expression input to reconstruct trajectories. Results indicate that SNP data improves

the trajectories reconstructed by Monocle. The major difference between the expression only and expression plus SNP trajectories is the position of d2_induced cells. In the expression plus SNP model, d2_induced cells are predicted to be progenitors (including for Neuron cells), consistent with their known role. In contrast, when only using expression data as input for Monocle, d2_induced cells are located at a separate branch without any descendants. See Supporting Results and Supporting Figures S8, S9 for the complete details.

TBSP scalability

While all three datasets discussed above studied dynamic processes, they have only profiled hundreds of cells each (though with relatively high overage). This number is quite low compared to more recent studies that usually profile thousands of cells. To test the scalability of TBSP, we have also used it to analyze a larger dataset which profiled close to 4000 Hematopoietic stem/progenitor cells (HSPCs) from mice bone marrow (43). While the data was collected over two days in order to study cell differentiation, all HSPCs were pooled before profiling and so unlike the longitudinal datasets discussed above, no time information is available for cells in this dataset. Still, TBSP was able to identify several significant SNPs for this data and these were used to cluster the cells and derive trajectories. Results are presented in Supporting Figure S10. Even though it does not use the expression data itself, TBSP derived trajectories agree well with the observation in the original paper of near-continuous differentiation process, from Haematopoietic stem cell/multipotent progenitor to Haematopoietic progenitor cell. Both the differentiation direction and the continuous process are captured by the SNP-based model, in which the trajectory starts from Cluster 0 (dominated by Haematopoietic stem cell), ends at Cluster 1 (which is heavily dominated by Haematopoietic progenitor cells) and the percentage of Haematopoietic progenitor cells in the cluster is increasing along the trajectories. See Supporting Results for complete details. We also include information on runtime and computing resources for different number of cells (we simulated up to 30 000 cells with 1.5M reads on average for each cell). Please refer to the Supporting Figure S11 for the runtime and memory requirements.

DISCUSSION

Existing methods for the analysis of time series scRNA-Seq data mostly utilize the expression levels for each cell. Such methods reconstruct developmental or response trajectories based on similarities between their expression (often in reduced dimensional space). While these methods have been successfully applied, it is also clear that in many cases that cannot accurately capture the dynamic process that they are modeling due to noise, dropouts and the impact of low expressed genes.

Here, we presented a complementary approach, TBSP, which utilizes what, until now, was a discarded part of the data: The errors in the sequenced RNAs. While some of these may indeed be just that (errors), others, especially

those that pass stringent filtering criteria and that are identified in multiple cells, are likely to be true differences. As we show, by using the identified SNPs we can construct reasonable trajectories assigning cells to different branches even without using the expression level information. We applied TBSP to four different datasets ranging in size from less than 200 to close to 4000 cells. As we show, in all cases the method was able to obtain good clusters and trajectories when only using SNP data. When combined with expression data to resulting models are even better and improve upon expression only models.

Some of the SNPs we identified may represent *de-novo* mutations inserted during cell division. However, it is unlikely that the majority are indeed such mutations given the small number of expected *de-novo* mutations in coding regions for the data that we studied as we estimated in the introduction. Instead, we argue that many of these likely represent RNA-editing events. Several lines of evidence support this claim. First, the type of mutation we observed, A/G or C/T substitutions, is consistent with RNA-editing sites are also mostly substitutions (44). Second, the locations of these mutations are enriched near the Alu elements, which has been also reported for RNA-editing sites (45). Third, the identified SNPs significantly overlap known RNA-editing sites reported in RADAR database (46). Fourth, many reside in genes that are known to be associated with processes that are regulated by RNA-editing such as protein degradation (42).

While SNP only models provide useful information about cell fate and branching, their performance is limited. Since the method depends on the mutation changes it often requires longer time scales than expression changes as we observed for the lung data. However, when SNP information is combined with gene expression data, the resulting models can improve upon the expression only models. For the lung data, the AT2 dominated cluster is the sibling node of AT1 dominated cluster when SNP information is combined with expression whereas the AT2 dominated cluster is the descendant of the AT1 dominated cluster in the expression only model (36).

While we believe that the integration of SNP and expression information would be useful for many studies, we note that it may not be a viable option in some cases. For example, when sampling rates are very short it is unlikely that many SNPs would be identified by our method TBSP even if large changes in expression occurs. Low coverage would also impact the accuracy of the method. In addition, many unique molecular identifier (UMI) technologies sacrifice the full-length coverage to sequence part of the primer used for cDNA generation (47,48), which would reduce the ability to detect SNP. Still, several existing and new datasets are sequencing full-length cDNA, and these can benefit from the method we presented.

Software implementing TBSP is freely available at GitHub (https://github.com/phoenixding/tbsp). As the number and types of biological processes that are studied using scRNA-Seq data increases, methods that can accurately infer developmental and response trajectories become an important part of the analysis and modeling process. We hope that TBSP, which aims at better uti-

lization of existing data, would aid researchers seeking to analyze such time series scRNA-Seq data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

REFERENCES

1. Bendall,S.C., Davis,K.L., Amir,E.D., Tadmor,M.D., Simonds,E.F., Chen,T.J., Shenfeld,D.K., Nolan,G.P. and Pe'er,D. (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, **157**, 714–725.
2. Trapnell,C., Cacchiarelli,D., Grimsby,J., Pokharel,P., Li,S., Morse,M., Lennon,N.J., Livak,K.J., Mikkelsen,T.S. and Rinn,J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381.
3. Qiu,X., Mao,Q., Tang,Y., Wang,L., Chawla,R., Pliner,H.A. and Trapnell,C. (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*, **14**, 979.
4. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495.
5. Setty,M., Tadmor,M.D., Reich-Zeliger,S., Angel,O., Salame,T.M., Kathail,P., Choi,K., Bendall,S., Friedman,N. and Pe'er,D. (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.*, **34**, 637.
6. Marco,E., Karp,R.L., Guo,G., Robson,P., Hart,A.H., Trippa,L. and Yuan,G.-C. (2014) Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5643–E5650.
7. Welch,J.D., Hartemink,A.J. and Prins,J.F. (2016) Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data. *Genome Biol.*, **17**, 106.
8. Rashid,S., Kotton,D.N. and Bar-Joseph,Z. (2017) Tasic: determining branching models from time series single cell data. *Bioinformatics*, **33**, 2504–2512.
9. Shin,J., Berg,D.A., Zhu,Y., Shin,J.Y., Song,J., Bonaguidi,M.A., Enikolopov,G., Nauen,D.W., Christian,K.M., Ming,G. *et al.* (2015) Single-cell rna-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell*, **17**, 360–372.
10. Ding,J., Aronow,B.J., Kaminski,N., Kitzmiller,J., Whitsett,J.A. and Bar-Joseph,Z. (2018) Reconstructing differentiation networks and their regulation from time series single-cell expression data. *Genome Res.*, **28**, 383–395.
11. Matsumoto,H. and Kiryu,H. (2016) Scoup: a probabilistic model based on the ornstein–uhlenbeck process to analyze single-cell expression data during differentiation. *BMC Bioinformatics*, **17**, 232.
12. Skelly,D.A., Squiers,G.T., McLellan,M.A., Bolisetty,M.T., Robson,P., Rosenthal,N.A. and Pinto,A.R. (2018) Single-cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. *Cell Rep.*, **22**, 600–610.
13. Camp,J.G., Sekine,K., Gerber,T., Loeffler-Wirth,H., Binder,H., Gac,M., Kanton,S., Kageyama,J., Damm,G., Seehofer,D. *et al.* (2017) Multilineage communication regulates human liver bud development from pluripotency. *Nature*, **546**, 533.
14. Wallrapp,A., Riesenfeld,S.J., Burkett,P.R., Raja-Abdulnour,E.E., Nyman,J., Dionne,D., Hofree,M., Cuoco,M.S., Rodman,C., Farouq,D. *et al.* (2017) The neuropeptide nmu amplifies ilc2-driven allergic lung inflammation. *Nature*, **549**, 351.
15. Kowalczyk,M.S., Tirosh,I., Heckl,D., Rao,T.N., Dixit,A., Haas,B.J., Schneider,R.K., Wagers,A.J., Ebert,B.L. and Regev,A. (2015)

Single-cell rna-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.*, **25**, 1860–1872.

16. Sun,N., Yu,X., Li,F., Liu,D., Suo,S., Chen,W., Chen,S., Song,L., Green,C.D., McDermott,J. *et al.* (2017) Inference of differentiation time for single cell transcriptomes using cell population reference data. *Nat. Commun.*, **8**, 1856.

17. Kester,L. and van Oudenaarden,A. (2018) Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell*, **23**, 166–179.

18. McKenna,A., Findlay,G.M., Gagnon,J.A., Horwitz,M.S., Schier,A.F. and Shendure,J. (2016) Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, **353**, aaf7907.

19. Michlits,G., Hubmann,M., Wu,S.-H., Vainorius,G., Budusan,E., Zhuk,S., Burkard,T.R., Novatchkova,M., Aichinger,M., Lu,Y. *et al.* (2017) Crispr-umi: single-cell lineage tracing of pooled crispr–cas9 screens. *Nat. Methods*, **14**, 1191.

20. Frieda,K.L., Linton,J.M., Hormoz,S., Choi,J., Chow,K.-H.K., Singer,Z.S., Budde,M.W., Elowitz,M.B. and Cai,L. (2017) Synthetic recording and in situ readout of lineage information in single cells. *Nature*, **541**, 107.

21. Doudna,J.A. and Charpentier,E. (2014) The new frontier of genome engineering with crispr-cas9. *Science*, **346**, 1258096.

22. Roach,J.C., Glusman,G., Smit,A.F.A., Huff,C.D., Hubley,R., Shannon,P.T., Rowen,L., Pant,K.P., Goodman,N., Bamshad,M. *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, **328**, 636–639.

23. Behm,M. and Öhman,M. (2016) Rna editing: a contributor to neuronal dynamics in the mammalian brain. *Trends Genet.*, **32**, 165–175.

24. Gommans,W.M., Mullen,S.P. and Maas,S. (2009) Rna editing: a driving force for adaptive evolution? *Bioessays*, **31**, 1137–1145.

25. Gagnidze,K., Rayon-Estrada,V., Harroch,S., Bulloch,K. and Papavasiliou,F.N. (2018) A new chapter in genetic medicine: RNA editing and its role in disease pathogenesis. *Trends Mol. Med.*, **24**, 294–303.

26. Kim,D., Langmead,B. and Salzberg,S.L. (2015) Hisat: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357.

27. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res.*, **20**, 1297–1303.

28. Van der Auwera,G.A., Carneiro,M.O., Hartl,C., Poplin,R., Del Angel,G., Levy-Moonshine,A., Jordan,T., Shakir,K., Roazen,D., Thibault,J. *et al.* (2013) From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, **43**, 11–10.

29. MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Univ. of Calif. Press, Oakland, Vol. **1**, pp. 281–297.

30. Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

31. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

32. Friedman,C.E., Nguyen,Q., Lukowski,S.W., Helfer,A., Chiu,H.S., Miklas,J., Levy,S., Suo,S., Han,J.D.J., Osteil,P. *et al.* (2018) Single-cell transcriptomic analysis of cardiac differentiation from human pscs reveals hopx-dependent cardiomyocyte maturation. *Cell Stem cell*, **23**, 586–598.

33. Huang,W., Loganantharaj,R., Schroeder,B., Fargo,D. and Li,L. (2013) Pavis: a tool for p eak a nnotation and vis ualization. *Bioinformatics*, **29**, 3097–3099.

34. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) Panther: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.

35. Treutlein,B., Lee,Q.Y., Camp,J.G., Mall,M., Koh,W., Shariati,S.A.M., Sim,S., Neff,N.F., Skotheim,J.M., Wernig,M. *et al.* (2016) Dissecting direct reprogramming from fibroblast to neuron using single-cell rna-seq. *Nature*, **534**, 391.

36. Treutlein,B., Brownfield,D.G., Wu,A.R., Neff,N.F., Mantalas,G.L., Espinoza,F.H., Desai,T.J., Krasnow,M.A. and Quake,S.R. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, **509**, 371.

37. Zinshteyn,B. and Nishikura,K. (2009) Adenosine-to-inosine rna editing. *Wiley Interdiscipl. Rev.: Syst. Biol. Med.*, **1**, 202–209.

38. Daniel,C., Silberberg,G., Behm,M. and Öhman,M. (2014) Alu elements shape the primate transcriptome by cis-regulation of rna editing. *Genome Biol.*, **15**, R28.

39. Ramaswami,G. and Li,J.B. (2013) Radar: a rigorously annotated database of a-to-i rna editing. *Nucleic Acids Res.*, **42**, D109–D113.

40. Xiong,H., Liu,D., Li,Q., Lei,M., Xu,L., Wu,L., Wang,Z., Ren,S., Li,W., Xia,M. *et al.* (2017) Red-ml: a novel, effective rna editing detection method based on machine learning. *Gigascience*, **6**, 1–8.

41. Macaulay,I.C., Haerty,W., Kumar,P., Li,Y.I., Hu,T.X., Teng,M.J., Goolam,M., Saurat,N., Coupland,P., Shirley,L.M. *et al.* (2015) G&t-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods*, **12**, 519.

42. Bajad,P., Jantsch,M.F., Keegan,L. and O'Connell,M. (2017) A to i editing in disease is not fake news. *RNA Biol.*, **14**, 1223–1231.

43. Nestorowa,S., Hamey,F.K., Sala,B.P., Diamanti,E., Shepherd,M., Laurenti,E., Wilson,N.K., Kent,D.G. and Göttgens,B. (2016) A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, **128**, e20–e31.

44. Eisenberg,E., Adamsky,K., Cohen,L., Amariglio,N., Hirshberg,A., Rechavi,G. and Levanon,E.Y. (2005) Identification of rna editing sites in the snp database. *Nucleic Acids Res.*, **33**, 4612–4617.

45. Kim,D.D.Y., Kim,T.T.Y., Walsh,T., Kobayashi,Y., Matise,T.C., Buyske,S. and Gabriel,A. (2004) Widespread rna editing of embedded alu elements in the human transcriptome. *Genome Res.*, **14**, 1719–1725.

46. Picardi,E., D'Erchia,A.M., Lo Giudice,C. and Pesole,G. (2016) Rediportal: a comprehensive database of a-to-i rna editing events in humans. *Nucleic Acids Res.*, **45**, D750–D757.

47. Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.

48. Ziegenhain,C., Vieth,B., Parekh,S., Reinius,B., Guillaumet-Adkins,A., Smets,M., Leonhardt,H., Heyn,H., Hellmann,I. and Enard,W. (2017) Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell*, **65**, 631–643.