Real-Time Course: Reconstruction of cellular diversity and lineage trajectory based on somatic mutational patterns detected from low-pass single-cell transcriptome data

Satoshi Oota (✉ oota@riken.jp)
RIKEN
Kuniya Abe
RIKEN
Hideo Yokota
Kazuho Ikeo
National Institute of Genetics

Article

Keywords:

Posted Date: April 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1558242/v1

Abstract

Single-cell RNA-seq (scRNA-seq) analysis can describe the states of individual cells but cannot clarify direct relationships between cells. Therefore, we developed a method to trace the progenitor cells of cell differentiation pathways through the cell lineage using somatic mutations detected in low-pass scRNA sequences. Using two human placental tissue datasets, we detected 2,172 and 627 somatic mutation candidates. We used only somatic mutation data to classify placental cells and construct a phylogenetic tree; the results were consistent with those of cell classification by expression profiling and estimation of the differentiation process by pseudotime course analysis. Somatic-mutation-based analysis can determine the temporal order of individual cells in the cell lineage, which is not possible with pseudotime analysis. Therefore, somatic mutation signatures in scRNA-seq data are informative for constructing cell lineages, and in combination with expression data obtained simultaneously, they are extremely useful for delineating the processes of human development and pathogenesis.

Introduction

Somatic mutations accumulate during development and aging, resulting in cellular mosaicism among somatic tissues. These mutations are associated with various diseases, including dementia, cardiovascular diseases, and cancers[1,2], as well as the aging process itself[3]. Somatic mutations play a role in normal development, for example in the acquisition of the diversity of neuronal cell populations in the central nervous system[4]. According to RNA sequence data analyses, a clonal expansion of somatic mutations occurs across normal tissues[5].

During early human development, 2.8 (95% confidence interval, 2.4–3.3) somatic mutations occur per cell per replication event, which is slightly higher than the rate of germline mutations[6]. However, in the later life, somatic mutation rates range from to mutations per base pair per cell division across various cell types[7]. Somatic mutation rates can reach ten times higher than those for the germline[8]. As a result, even monozygotic twins can be genetically diverse owing to somatic mutations[9]. Somatic mutations are thus ubiquitous events in various organisms, including humans[10]; they can be regarded as counterparts of germline mutations[11], whose dynamics are subject to the evolutionary process[12]. Similar to germline cells, each somatic cell carries an ‘evolutionary record’ as “as scars (mutations) that accrue” that accrue in the genome with time. Thus, it is theoretically possible to reconstruct the history of a cell using a retrospective cell tracing manner[13,14], i.e., somatic mutations bear temporal information of the cell lineage.

Pioneered by Charles Whitman[15] in the 19th century for the developmental biology of invertebrates, lineage tracing is now an essential tool to study stem cells in adult mammalian tissues. Somatic mutations have been extensively studied in the context of cancer evolution[16]. Researchers have devised mathematical models to elucidate somatic mutation dynamics from various biological perspectives[17]. In the case of cancer evolution studies, however, we must address the challanges in detecting *de novo*

mutations; the error rates of next-generation sequencing technologies are considerably high, making the detection of rare variants difficult. To handle the intractability, we usually need ultra-deep sequencing[18] to increase accuracy. Genuine somatic mutations must be distinguished from germline mutations[19]. Issues such as these can cause noise in the method of using 'phylogenetic' signatures to infer ancestral mutations in progeny cells..

Single-cell sequencing (SCS) aids analyses in various fields, including rare cell types, uncultured microbes, and mosaicism of somatic tissues[20]. Since somatic mutations are attributed to genomic alterations, it is a straightforward approach to use somatic genome sequence data for the detection of rare variants against the zygotic reference genome. However, in SCS, genome data have a shortcoming: A single cell has only two copies of each DNA molecule, which can lead to various technical issues such as coverage nonuniformity, allelic dropout events, false-positive errors, and false-negative errors[21].

In contrast, with RNA sequence sources, we can utilize hundreds or thousands of copies of RNA molecules in a cell. It is important to consider false positives due to biological and technical factors, including RNA editing, sequencing errors (e.g., random errors introduced during reverse transcription and PCR), and potential sampling errors[10]. The overwhelming abundance of single-cell transcriptome data has the potential to compensate for the disadvantages of RNA sequencing data[22]. However, the currently available frameworks rely on the nonlinear transformation of high-dimensional data to human-recognizable low-dimensional representations, which often retain a nonlinear nature and can deviate from physical entities such as absolute time.

Here, we propose a new framework to detect cell lineage trajectories using potential somatic mutations detected in single-cell transcriptome data: Real-Time Course Analysis. Unlike pseudotime course analysis using heterogeneously differentiated cells in a single time snapshot[23,24], the approach in the present study uses somatic mutations to trace cellular lineages back to progeny cells. Thus, our framework can handle the direction of physical time in terms of the expected number of mutations. We focused on gross phylogenetic signatures in single-cell transcriptome data. For example, 60 somatic cells have approximately 10^{96} possible lineages[25]; hence it is virtually impossible to identify only one true cell lineage with an exhaustive search. Our approach is to narrow the window that covers the true lineage using the gross phylogenetic signatures potentially retained in the RNA sequence data. Today, there are several dimension reduction methods to categorize cell types and perform pseudotime course analysis based on machine learning approaches, such as t-distributed stochastic neighbor embedding (t-SNE)[26] and uniform manifold approximation and projection for dimension reduction (UMAP)[27]. However, these methods sometimes lack both biologically relevant interpretations and reproducibility of global data structure[28].

The aims of this study were: (1) reconstructing putative lineage trees of normal placental tissue cells (**Supplementary figure 1**) using low-coverage single-cell somatic variants obtained from single-cell RNA-seq data, in which we utilize the reference genome sequence to compensate for missing alignments with

a particular threshold; (2) developing a proof of concept (PoC) study for a new framework to infer the temporal cellular trajectory using the genetic mosaicism of single-cell populations estimated from the compensated alignments; and (3) comparing and integrating the somatic variant-based cellular trajectory and the corresponding gene expression profiles, dubbed Real-Time Course analysis.

Results

Detection of somatic variants in the single-cell transcriptome data

We mapped 3,088,286,401 bp of Batch1 sequence data onto the reference human genome (GRCh38)[29]. The average coverage percentage was 0.685% [standard deviation (SD) = 0.231]. For Batch2 sequence data, the average coverage percentage was 0.477% (SD: 0.243).

The apparent variants (deviations from the reference genome in the mapped transcripts) contained germline and somatic mutations. As no information was provided for haplotype phasing in the data set of Pavličev et al.[30], we employed a simple method to detect somatic variants in multiallelic sites in a cell population. We categorized possible mutational patterns across the cells, assuming that at most one somatic mutation occurred in a lineage[31] (**Fig. 1**). Our strategy followed two steps: (1) screening all the sites in the transcripts that were different from the corresponding reference genome sites and (2) comparing the screened data across cells and selection of multiallelic sites.

In a biallelic site, we could not determine the lineage in which the variation occurred or whether the site was heterozygous (**Fig. 1**). However, when we looked into a multiallelic site, the only explanation would be that at least one mutation occurred at some point in the cell lineage (**Fig. 1**c and **Fig. 1**e). Therefore, we searched for multiallelic sites.

We detected sites that were different from the reference genome during the first screening (pairwise comparison with the reference genome). A total of 1,965,629 sites in Batch1 and 830,905 sites in Batch2 were detected during the first screening. Examples of apparent variant sites are shown in **Supplementary figure *2***.

Among these, 89 multiallelic sites were found in 54 Batch1 cells. Moreover, 2,083 sites were multiallelic in 43.2 single cells on average (80% of the 54 single cells) in Batch1. Similarly, in Batch2, 53 sites were multiallelic in all 33 single cells. Moreover, 574 sites were multiallelic in 26.4 single cells on average (80% of the 33 single cells). However, it remains unclear in which lineage the mutation occurred. The state of the reference genome can be used as an ancestral state if the observed nucleotides share the reference site. However, multiallelic sites were always heterozygous under our framework, and the reference genome data were not haplotype-phased. Here, we selected the ‘minor allele’ as a derived variant, assuming that minor alleles are newly derived nucleotides in a cell population. We used 2083 variants sites from Batch1 and 574 variants from Batch2 for subsequent analysis.

Annotation of variants

Next, we examined the locations of these variants in the genic sequences using SnpEff software[45]. **Supplementary tables 1** and **2** show the annotations of the estimated variants. Note that the results include alternative annotations, such as nested intronic genes. Some variant counts overlapped between the categories. The number of variants is different from the results based on our criteria; for example, while we detected 2,083 and 574 variants from Batch1 and Batch2 data with the 80% criterion, respectively, SnpEff software estimated 1,903 and 1,398 variants from Batch1 and Batch2 data with the default parameter set, respectively. SnpEff results showed that there were 550 and 216 missense variants and 199 and 135 synonymous variants in Batch1 and Batch2 data, respectively.

dN /dS ratio in coding regions

We employed an evolutionary framework to assess the quality of the detected somatic variants. Genuine somatic mutations are expected to be subjected to purifying selection in normal tissues[32]. Thus, we calculated the dN/dS ratios from Batch1 and Batch2 sequence data. The overall dN/dS ratios of Batch1 and Batch2 are 0.865 and 0.556, respectively. These results suggest that the somatic variants in the two samples were subject to purifying selection, supporting our expectations.

Phylogenetic analysis of placental single-cell data

Using the variant data described above, we reconstructed the phylogenetic trees of single cells from two different individual placental tissue samples (**Figs 2** and **3**). As we used the parsimony method[33] implemented in MEGA X[34], branch lengths represent the expected numbers of mutations that occurred on the branches. Each leaf (operational taxonomic unit[35]) represented a cell sampled from each tissue.

A model of the differentiation process from villous cytotrophoblasts (vCTB) to extravillous trophoblasts (EVT) has been constructed by Knofler et al.[36], in which EVT are differentiated from vCTB via cell column trophoblasts (CCT) (**Supplementary figure *1***). Pavličev et al. used PCA analysis and the expression patterns of marker genes to classify the single-cell sample into five clusters (cell types): intravillous cytotrophoblast (CYT1, CYT2, and CYT3), EVT, and the maternal decidual cell (DC). Using the single-cell classification by Pavličev et al., we mapped cells onto the lineage tree constructed by our variant analysis. As shown in **Fig. 2**, most (eight out of nine) of the CYT1 cells are on the "trunk of the tree [**Fig. 2** (c)]" and considered as progenitor cells of CYT2, CYT3, and EVT. These cells can be ordered on trees on the basis of their mutation patterns. Most differentiated cells, that is, CYT2, CYT3, and EVT, emerged after the appearance of SRR4371572 CYT1 cells [**Fig. 2** (d)]. The differentiation of CYT2 and CYT3 suggests that intravillous cytotrophoblasts are a heterogeneous group. The expression profiles of

these cells suggest that differentiation proceeds by dynamic gene expression in a different order rather than uniform progression toward a terminal expression profile[30].

Meanwhile, we found a small number of discrepant results regarding the above patterns, as shown in (a) and (b) in **Fig. 2**. **Supplementary figure 3** shows the relationships between the percentage of coverage and cells that were classified in a discrepant manner. We found that discrepant cells had relatively low coverage values.

A similar trend was observed in Batch2 data, demonstrating that most of the CYT1 progenitor cells were mapped before the branching point (**Fig. 3**). These data suggest that our phylogenetic analysis can classify cells based solely on variant information, and the results show good agreement with cell classification based on expression profiling. More importantly, it is possible to align the progenitor CYT1 cells in a temporal order, which is not possible with cell clustering based on expression profiles, including pseudotime course analysis[37-39].

Pavličev et al. reported that their samples contained maternal DCs. However, the results of the present study suggest that the putative DC cells assumed by Pavličev et al. appear to be differentiated from fetal CYT or stem cells of CYT (**Figs 2** and **3**). We reanalyzed single-cell expression data using t-SNE and detected three clusters. Both putative DCs (claimed by Pavličev et al.) belong to t-SNE Cluster 1 (**Fig. 4**). Therefore, our data, both expression and variant data, suggest that these two cells are similar to CYT 1 cells and not DC cells. In fact, the expression of some DC-characteristic genes, such as *CLEC4C*, *THBD*, *CD1C*, *CD80*, *IL10*, and *IL12B*, was not found in these cells[30].

Pseudotime course analysis

Using our variant data analysis method, it is theoretically possible to perform "real-time" course analysis, and the cells can be aligned in temporal order (**Figs 2** and **3**). Here, we compared the results of real-time analysis with those of pseudotime course analysis[37-39].

Fig. 5 shows the results of pseudotime course analysis of SRP090944 data Batch1, and the cells mapped onto the pseudotime contours are marked by letters or numbers representing cell types determined by Pavličev et al.[30], which is more consistent with the real-time course analysis. Pseudotimes of the cells are represented by contours created by a function of *Mathematica*[40]. The results were generally consistent with the results of our real-time course analysis except for the direction of the pseudotimes. Sixteen cells had pseudotimes greater than 2.5, of which ten correspond to CYT 1 progenitor cells. Cells differentiated from CYT1 cells, i.e., CYT2 and CYT3, are detected in regions with lower pseudotime values, suggesting again that these cells are temporally distant from CYT 1 cells in the cell lineage. Although EVT cells are located at close proximity to CYT1 cells on the pseudotime scale, our

real-time analysis indicates that these cells are not closely related to CYT1 cells, which is consistent with previous biological knowledge. Therefore, our real-time analysis can yield similar but more direct results compared to the well-established pseudotime course analysis.

Discussion

This study proposes a new framework to infer cellular lineage trees using somatic variants detected in low coverage single-cell RNA sequence data. We focused on the gross phylogenetic signature of data from the single-cell rather than on each mutation. We showed that it was possible to reconstruct a cellular lineage tree solely based on variant data, and the lineage thus obtained was consistent with the established biological knowledge.

The significance of this study is in inferring the direct relationships between single cells using genetic scars in a cell, that is, somatic mutations. However, conventional analyses provide only indirect relationships based on similarities in expression profiles. In addition, because somatic mutations accumulate over time, our real-time course should be able to trace cellular lineages back to progeny cells, meaning that our approach has the potential to allow us to infer the cellular time course, for example, non-observable past events. In fact, our real-time course provided results more consistent than those of pseudotime-course analysis. The major difference is that the real-time course based on mutation data can infer linear and direct relationships between cells in the lineage in terms of the number of mutations. In contrast, pseudotime course analysis only can infer non-linear and indirect relationships. To compensate for the latter shortcoming, additional studies are warranted to leverage splicing kinetics, i.e., RNA velocity[41]. We also consider that the real-time course innately integrates the cellular trajectories with the corresponding gene expression profiles while the two datasets are separately handled. This means that their relationships can be used to provide insights into the biological system of interest.

A potential issue of our approach is that the real-time course relies on the phylogenetic signature retained in low-pass single-cell sequence data. The phylogenetic signature can be obscured by random noise due to sequencing errors as well as variant dropouts due to low coverage. In fact, as shown in **Supplementary figure 3**, the discrepant cells had relatively low coverage values (indicated by red bars). To assess the quality of the detected somatic variants, we employed an evolutionary framework in terms of apparent selective pressure. Because somatic mutations have a negative impact on cell survival in the cell population, it is rational to assume that genuine somatic mutations are subjected to purifying selection in normal tissues[32]. According to the overall dN/dS values, somatic mutations were subject to purifying selection, suggesting that the majority of the detected somatic mutations were not likely to be false positives.

It is important to identify how frequently somatic mutations occur in normal tissues. According to Tomasetti et al.[7], the in vivo tissue-specific somatic mutation probability per base per cell division is $7.6 \times 10^{-10} \pm 1.1 \times 10^{-10}$ (SE) in normal lymphocytes. Therefore, stochastic somatic alterations

occur at an average rate of three mutations per cell division in the human genome, except for repetitive regions[42,43]. This abundance of somatic mutations makes retrospective cell lineage inference feasible, even in normal tissues.

In our single-cell analysis, we need to consider three types of 'variants': (1) polymorphism of the population (mutations between individuals): i.e., germline mutations; (2) de novo mutations within the individual: i.e., somatic mutations; and (3) sequencing errors. In bulk sequencing data, we need to employ a statistical approach to handle somatic mutations; typically, we use variant allele frequency (VAF) to infer whether a variant comes from somatic cells or is inherited from parents[44]. However, the distribution of RNA-VAF and DNA-VAF can differ owing to allelic imbalance (AI)[45]. Single-cell analysis allowed us to exclude germline mutations by simply comparing variants in a cell population.

It remains challenging to identify the lineage in which a somatic mutation occurs (**Fig. 1**). We selected 'minor alleles' as potential somatic mutations, assuming that the transcriptome allele frequencies reflected the genomic allele frequencies in many genomic regions. This assumption is based on an empirical expectation that the allele frequencies of the genome and transcriptome reads are correlated[46].

We used homologous sites of the reference genome to fill in the missing sites in the inter-cell alignments, assuming that genomic regions that were not covered by transcripts were identical to the reference genome sequences. We used an 80% threshold as a pragmatic solution.

In our method, homozygous sites cannot be employed to detect variants because multiallelic sites are required, in which variant sites are consequently chromosome-heterozygous in those progenies. (**Fig. 1**c and **Fig. 1**e). Thus, we missed all potential variants in the homologous sites. This limitation can also be mitigated by using biallelic sites derived from homologous sites (**Fig. 1**g) with statistical haplotype phasing[47]. Nevertheless, we believe that our primary objective of providing a PoC for cell lineage inference using transcriptome data has been achieved. Further analyses using other types of mutations will be conducted separately in the future.

Ideally, single-cell genome data would be more informative for inferring a cell lineage tree. We employed static and rich information to detect somatic mutations. However, there are technical and practical issues with single-cell genome analysis: We need to deal with only two copies of DNA molecules in a single cell, and the cost for sequencing the entire genome of thousands of cells is huge compared to that for RNA sequencing. Conversely, single-cell transcriptome analysis is currently a well-established technology, and a massive amount of single-cell transcriptome data has been (and will be) accumulated in sequence databases. Genotype (mosaicism) information has not been utilized in single-cell gene expression analyses. We showed that it is possible to extract significant genotypic information from RNA sequencing data without incurring additional costs. With rigorous computational approaches (e.g., haplotype phasing), the current technique can be improved.

Collectively, while real-time course analysis of single-cell transcriptome data has limitations in terms of zygosity and coverage at apparent variant sites, it has great potential to delineate cellular trajectories during development and to interpret high-dimensional gene expression data. Furthermore, the biological significance of somatic mutations provides insight into a new perspective of 'evolution' within an individual. This approach should be highly effective, especially in human biology, in which experimental manipulation is strictly restricted.

In conclusion, we propose a new framework for the estimation of cellular trajectories in complex tissues designated as real-time courses. We showed that it is possible to reconstruct a phylogenetic tree of somatic cells in placental tissues using only low-pass single-cell RNA sequence data. This tree is consistent with the known cell lineage of the placenta in four cell types: CYT 1, CYT 2, CYT 3, and EVT. The findings of this study suggest that robust phylogenetic signals exist in RNA sequencing data, and these signals can be used for the construction of cellular networks operating in human development, as well as in human disease pathogenesis. Our research will contribute to single-cell transcriptome analyses in alternative perspectives, for example, to analyze mosaicism of single cell genomes as a phylogenetic signature to infer cellular trajectories. The real-time course is based on mutation data and can infer linear and direct relationships between cells in the lineage in terms of the number of mutations. In contrast, pseudotime course analysis only can infer non-linear and undirected relationships. To compensate for the latter shortcoming, we need additional experiments to leverage splicing kinetics, i.e., RNA velocity[41]. Furthermore, the real-time course can theoretically infer the dynamics of ancestral (progeny) cells that are impossible to observe directly. In contrast, pseudotime course analysis has no capability to reconstruct the past.

We also suggest that the real-time course innately integrates the cellular trajectories with the corresponding gene expression profiles while the two data sets are separately handled. This means that we can use their relationships to increase our insight into the biological system of interest.

References

1. Kennedy, S. R., Loeb, L. A. & Herr, A. J. Somatic mutations in aging, cancer and neurodegeneration. *Mech. Ageing Dev.* **133**, 118-126 (2012).

2. Morley, A. A. The somatic mutation theory of ageing. *Mutat. Res.* **338**, 19-23 (1995).

3. Kelly, D. P. Cell biology: Ageing theories unified. *Nature* **470**, 342–343 (2011).

4. Abeliovich, A. et al. On somatic recombination in the central nervous system of transgenic mice. *Science* **257**, 404-410 (1992).

5. Yizhak, K. et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, eaaw0726 (2019).

6. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714-718 (2017).

7. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl Acad. Sci. U. S. A.* **110**, 1999–2004 (2013).

8. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. U. S. A.* **107**, 961–968 (2010).

9. Jonsson, H. et al. Differences between germline genomes of monozygotic twins. *Nat. Genet.* **53**, 27-34 (2021).

10. García-Nieto, P. E., Morrison, A. J. & Fraser, H. B. The somatic mutation landscape of the human body. *Genome Biol.* **20**, 298 (2019).

11. Milholland, B. et al. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**, 15183 (2017).

12. Rozhok, A. I. & DeGregori, J. Toward an evolutionary model of cancer: Considering the mechanisms that govern the fate of somatic mutations. *Proc. Natl Acad. Sci. U. S. A.* **112**, 8914-8921 (2015).

13. Woodworth, M. B., Girskis, K. M. & Walsh, C. A. Building a lineage from single cells: Genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).

14. Oota, S. Somatic mutations - Evolution within the individual. *Methods* **176**, 91-98 (2020).

15. Davenport, C. B. The personality, heredity and work of Charles Otis Whitman, 1843-1910. *Am. Nat.* **51**, 5-30 (1917).

16. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell* **168**, 613-628 (2017).

17. Altrock, P. M., Liu, L. L. & Michor, F. The mathematics of cancer: Integrating quantitative models. *Nat. Rev. Cancer* **15**, 730-745 (2015).

18. Rheinbay, E. et al. Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55-60 (2017).

19. Sun, J. X. et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLOS Comput. Biol.* **14**, e1005965 (2018).

20. Method of the year 2013. *Nat. Methods* **11**, 1 (2014).

21. Wang, Y. & Navin, N. E. Advances and applications of single-cell sequencing technologies. *Mol. Cell* **58**, 598-609 (2015).

22. Tam, P. P. L. & Ho, J. W. K. Cellular diversity and lineage trajectory: Insights from mouse single cell transcriptomes. *Development* **147**, dev179788 (2020).

23. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117-e117 (2016).

24. Campbell, K. R. & Yau, C. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. *Nat. Commun.* **9**, 2442-2442 (2018).

25. Felsenstein, J. The number of evolutionary trees. *Syst. Biol.* **27**, 27-33 (1978).

26. Hinton, G. & Roweis, S. Stochastic neighbor embedding. *Adv. Neural Inf. Process. Syst.* **15**, 833--840 (2003).

27. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction, *ArXiv*. *1802.03426 (2020)*.

28. Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.* **39**, 156-157 (2021).

29. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *bioRxiv*, 072116 (2016).

30. Pavličev, M. et al. Single-cell transcriptomics of the human placenta: Inferring the cell communication network of the maternal-fetal interface. *Genome Res.* **27**, 349-361 (2017).

31. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893-903 (1969).

32. Yadav, V. K., DeGregori, J. & De, S. The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. *Nucleic Acids Res.* **44**, 2075-2084 (2016).

33. Fitch, W. M. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* **20**, 406-416 (1971).

34. Tamura, K., Stecher, G. & Kumar, S.. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* version 11 **38**, 3022-3027 (2021).

35. Long, C. A., Sokal, R. R. & Sneath, P. H. A. Principles of numerical taxonomy. *J. Mammal.***46**, 111-112 (1965).

36. Knöfler, M. et al. Human placenta and trophoblast development: Key molecular mechanisms and model systems. *Cell. Mol. Life Sci.* **76**, 3479-3496 (2019).

37. 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).

38. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979-982 (2017).

39. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).

40. Wolfram Research & I. 12.2 version (Wolfram Research, Inc. (Champaign, IL, 2020)).

41. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494-498 (2018).

42. Ortega, M. A. et al. Using single-cell multiple omics approaches to resolve tumor heterogeneity. *Clin. Transl. Med.* **6**, 46 (2017).

43. Araten, D. J. et al. A quantitative measurement of the human somatic mutation rate. *Cancer Res.* **65**, 8111-8117 (2005).

44. Dou, Y., Gold, H. D., Luquette, L. J. & Park, P. J. Detecting somatic mutations in normal cells. *Trends Genet.* **34**, 545-557 (2018).

45. Rhee, J. K., Lee, S., Park, W. Y., Kim, Y. H. & Kim, T. M. Allelic imbalance of somatic mutations in cancer genomes and transcriptomes. *Sci. Rep.* **7**, 1653 (2017).

46. Ju, Y. S. et al. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat. Genet.* **43**, 745-752 (2011).

47. Browning, S. R. & Browning, B. L. Haplotype phasing: Existing methods and new developments. *Nat. Rev. Genet.* **12**, 703-714 (2011).

Methods

Single-cell RNA-seq. In this study, we used two public transcriptome datasets obtained from normal placental tissues[30]: SRP090944 Batch1 (54 cells) and Batch2 (33 cells). Pavličev et al.[30] analyzed the placenta data in the context of the cell communication network between two semiallogenic individuals, the mother and the fetus. The authors inferred the cell-cell interactome according to the gene expression of receptor-ligand pairs across cell types and found cell-type-specific expression of G-protein-coupled receptors, suggesting that ligand-receptor profiles could be a reliable tool for cell type identification. The data are registered in the DDBJ Sequence Read Archive as SRR4371525 (GSM2339239)–SRS1732319

(SRX2225328) (**Supplementary tables 3** and **4**). We assumed that Batch1 and Batch 2 shared no *de novo* somatic mutations.

Pavličev et al. categorized the single-cell data into five clusters (cell types) according to the principal comportment analysis (PCA) on the gene expression profile and 300 marker genes: cytotrophoblast (CYT) 1, CYT2, and CYT3, extravillous trophoblast (EVT), and the putative maternal decidual cell (DC). **Supplementary figure 1** shows a model system that integrates the role of Notch/Wnt signaling in trophoblast stemness and extravillous trophoblast differentiation[36]

Mapping of sequence data. We intended to construct multiple alignment data that represent single cells by aligning the homologous regions of their transcriptomes. To this end, read sequences (segmented sequences of base pairs corresponding to part of a single DNA fragment, a set of which is expected to cover partial coding regions) were assigned designated coordinates of the reference genome. The entire data analysis pipeline is shown in **Supplementary figure 4**. We mapped single-cell transcriptome sequence data to the human genome (GRCh38)[29] using the Burrows-Wheeler Aligner[48] after removing adaptor sequences with trimmomatic[49]. We identified variants by comparing the aligned read sequences with the reference genome using Samtools[50]. In the present study, we used only single-nucleotide variants, excluding all detected indel events. If we encountered incomplete sites across cells by more than 50%, we excluded the corresponding sites. We annotated the detected variant sites using the SnpEff software[51], which can predict the effects of genetic variants on genes and transcripts.

Phylogenetic analysis of single cells. We concatenated all acquired variant sites to generate sequence alignments. We reconstructed cellular lineage trees using the maximum parsimony method[33] implemented in MEGA X[34] with default parameters. The results were output in the Newick tree format[52] for the following process: dN/dS ratio computations with the maximum likelihood method and comparative analysis between the mutation patterns and the gene expression profile.

dN/dS ratios in the coding regions. We assembled codon sequences into which the detected variants fell and created codon alignments with exonic variants. We computed the overall dN/dS ratios using the Codeml package[53].

Gene expression reanalysis. For linear dimensionality reduction, we conducted principal component analysis (PCA) on the expression patterns using R (version 3.6.2)[54], and subsequently applied t-SNE[26]and UMAP[27] to the data for nonlinear dimensionality reduction. The Louvain method was used for the clustering[55].

The pseudotime course analysis. We conducted a pseudotime course analysis using monocle3 (version 1.0.0)[37-39] on R (version 4.1.2)[54]. Pseudotime course analysis estimates the temporal dimension of static single-cell transcriptome data and infers individual gene expression dynamics along with cellular changes. We reduced the data dimension to 26 using PCA. To visualize the results, we used *Mathematica*[40].

Comparative analysis of mutation patterns and gene expression profiles. We mapped the clustered cells to the cellular lineage trees, reconstructed based on cellular genotypes. To this end, we developed a *Mathematica*[40] code, *AssignCluster2Cell*, using two packages: Phylogenetics for *Mathematica*[56] and Phylogenetics[57]. *AssignCluster2Cell* reads a Newick[52] formatted tree file and a cluster table of a gene expression profile, and maps cluster ID to the tree topology, evaluating degrees of monophyleticity (DoM) at each node in terms of the proportion of the clusters in its subtree (**Fig. 6**).

We also mapped the clustered cells to the tree to evaluate the DoMs. The pie charts in **Fig. 6** show the DoM at each node; for example, if the pie chart at a node has only one slice, the subtree is completely monophyletic; if the pie chart at a node has two slices, the subtree is polyphyletic (subtree B in **Fig. 6**). However, because cell type 2 is dominant in subtree B, we can infer that cell type 1 is derived from cell type 2. In this manner, we determined the relevance of cell types based on gene expression profiles and cell lineages.

Theoretically, the root of a cellular lineage tree represents a zygotic cell and the root represents the zygotic genome of the fertilized egg. However, in the present study, the root of an estimated tree represented a progenitor cell of the observed cells. The zygotic cell should reside somewhere between the root and the reference genome.

References

48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009). 10.1093/bioinformatics/btp324, Pubmed:19451168.

49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014). 10.1093/bioinformatics/btu170, Pubmed:24695404.

50. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009). 10.1093/bioinformatics/btp352, Pubmed:19505943.

51. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012). 10.4161/fly.19695, Pubmed:22728672.

52. Felsenstein, J. & PHYLIP. *Phylogeny Inference Package*. 3.2 version. *Cladistics* **5**, Vols. 164-166, (1989).

53. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007). 10.1093/molbev/msm088, Pubmed:17483113.

54. R Core Team ((Vienna, Austria, 2016)).

55. Blondel, V. D., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theor. Exp.* **2008**, 10008 (2008). 10.1088/1742-5468/2008/10/P10008.

56. Polly, P. D. (Indiana University Bloomington. *Indiana*).

57. Zachar, I. GitHub repository. in *ed686781c686980dcd686982ff111438bb686935a686986c686983a686951f686982b686984 (GitHub, 2017)*, Vol. 2021. 686980.

58. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767-1771 (2010). 10.1093/nar/gkp1137, Pubmed:20015970.

59. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics***27**, 2156-2158 (2011). 10.1093/bioinformatics/btr330, Pubmed:21653522.

Declarations

Data Availability Statement

The data that support the findings of this study are divided into two sets, SRP090944 Batch1 (54 cells) and Batch2 (33 cells). The both data sets are available in the DDBJ Sequence Read Archive (https://www.ddbj.nig.ac.jp/dra/index-e.html) as SRR4371525 (GSM2339239)–SRS1732319 (SRX2225328).

Figures

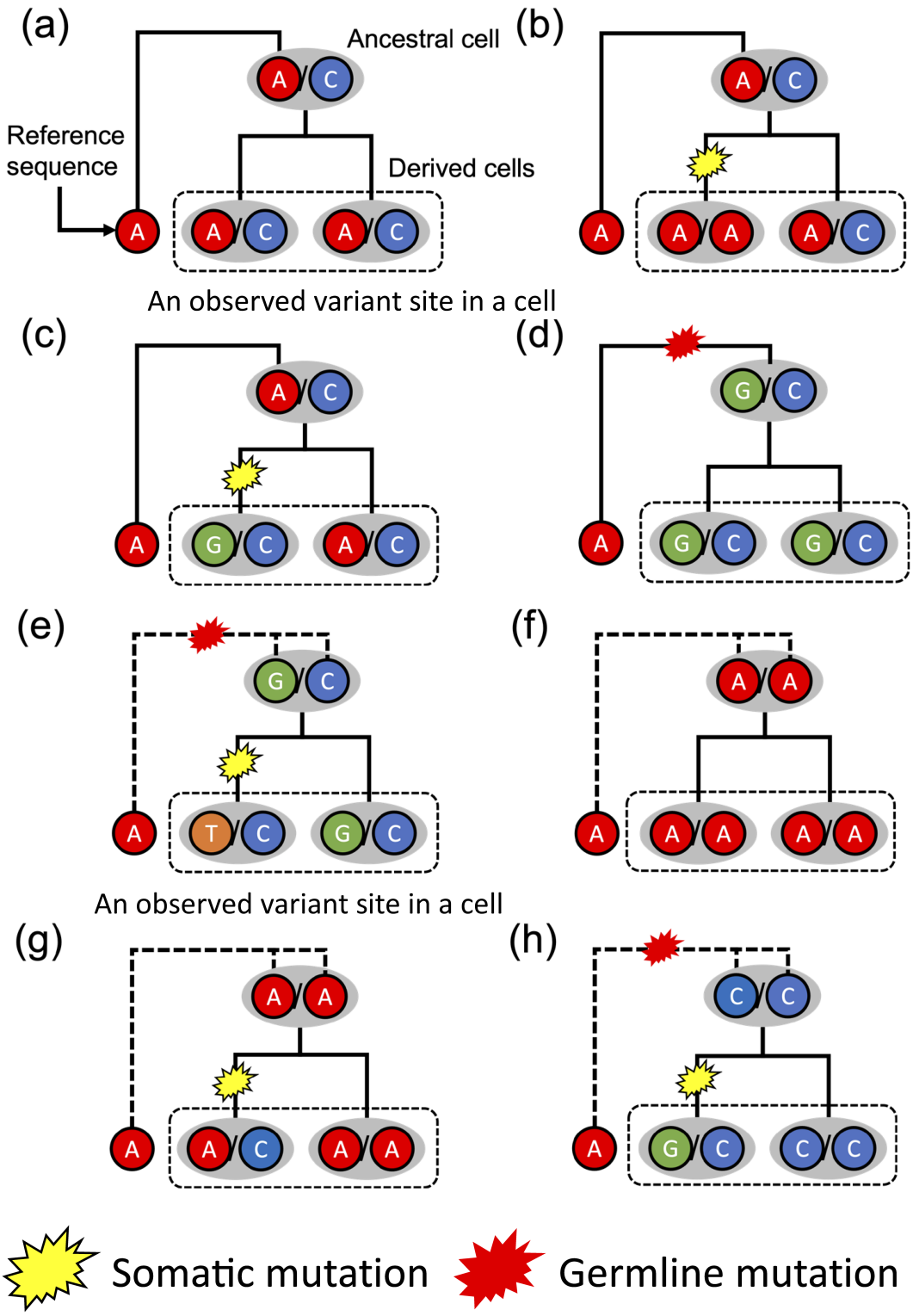

Figure 1

Detection of somatic mutations via multiallelic variants. (a) An apparent variant site derived from an ancestral heterozygous site without germline and somatic mutations. (b) An apparent variant site derived from an ancestral heterozygous site with a somatic mutation. However, we cannot distinguish this configuration from (a). (c) A variant site derived from an ancestral heterozygous site with a somatic mutation. We can distinguish this configuration from (a) and (b) owing to the third kind of nucleotide G.

(d) An apparent variant site derived from an ancestral heterozygous site with a germline mutation. However, we cannot distinguish this configuration from (a) and (b). (e) A variant site derived from an ancestral heterozygous site with a somatic mutation following a germline mutation. We can distinguish this configuration from (a) and (b) owing to the third kind of nucleotide T. (f) A homogenous site derived from a homozygous site. (g) A variant site derived from an ancestral homologous site with a somatic mutation. We cannot distinguish this configuration from (a) and (b). (h) An apparent variant site derived from an ancestral homologous site with a somatic mutation following a germline mutation. However, we cannot distinguish this configuration from (d). We assume that the reference genome is haplotype-phased for convenience. The solid line represents a haplotype-phased reference site. The dashed line represents the homozygous reference site. Only two representative cells derived from their ancestral cells are shown here.

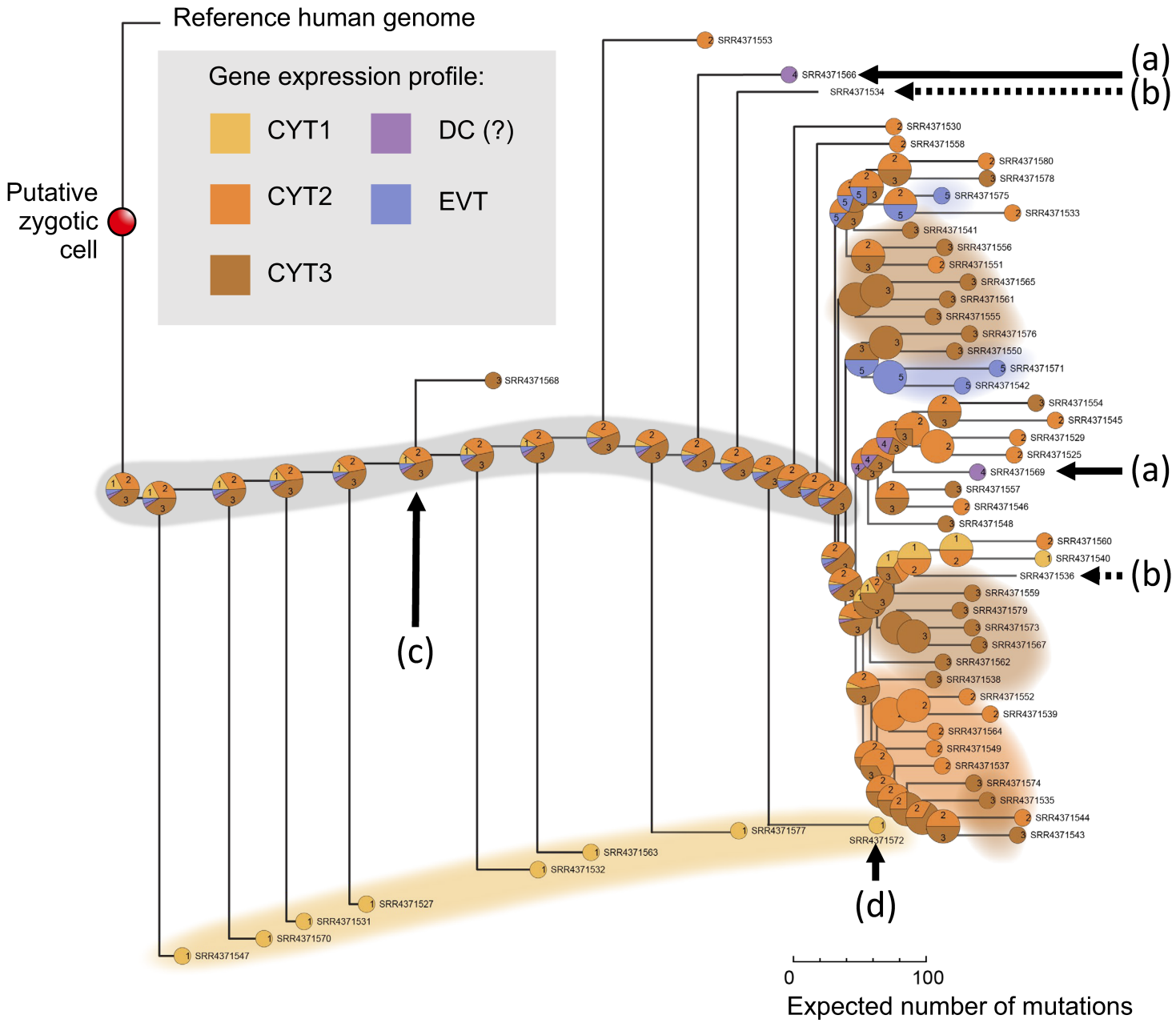

Figure 2

Mapping of the gene expression profile to a cellular trajectory tree (SRP090944 data; Batch1, 54 cells). The gene expression profile (as clustered cells) is mapped to the tree topology: External nodes represent observed cells, and internal nodes represent inferred cells, which are color-coded according to the gene expression patterns with pie charts. (a) Placenta cells claimed as maternal decidual cell (DC) in a previous study[30]; (b) Placenta cells that were not annotated[30]; (c) A putative stem cell self-renewal phase. Scale bar: expected number of mutations. Each number in the pie charts represents a designated cluster according to the gene expression profile.

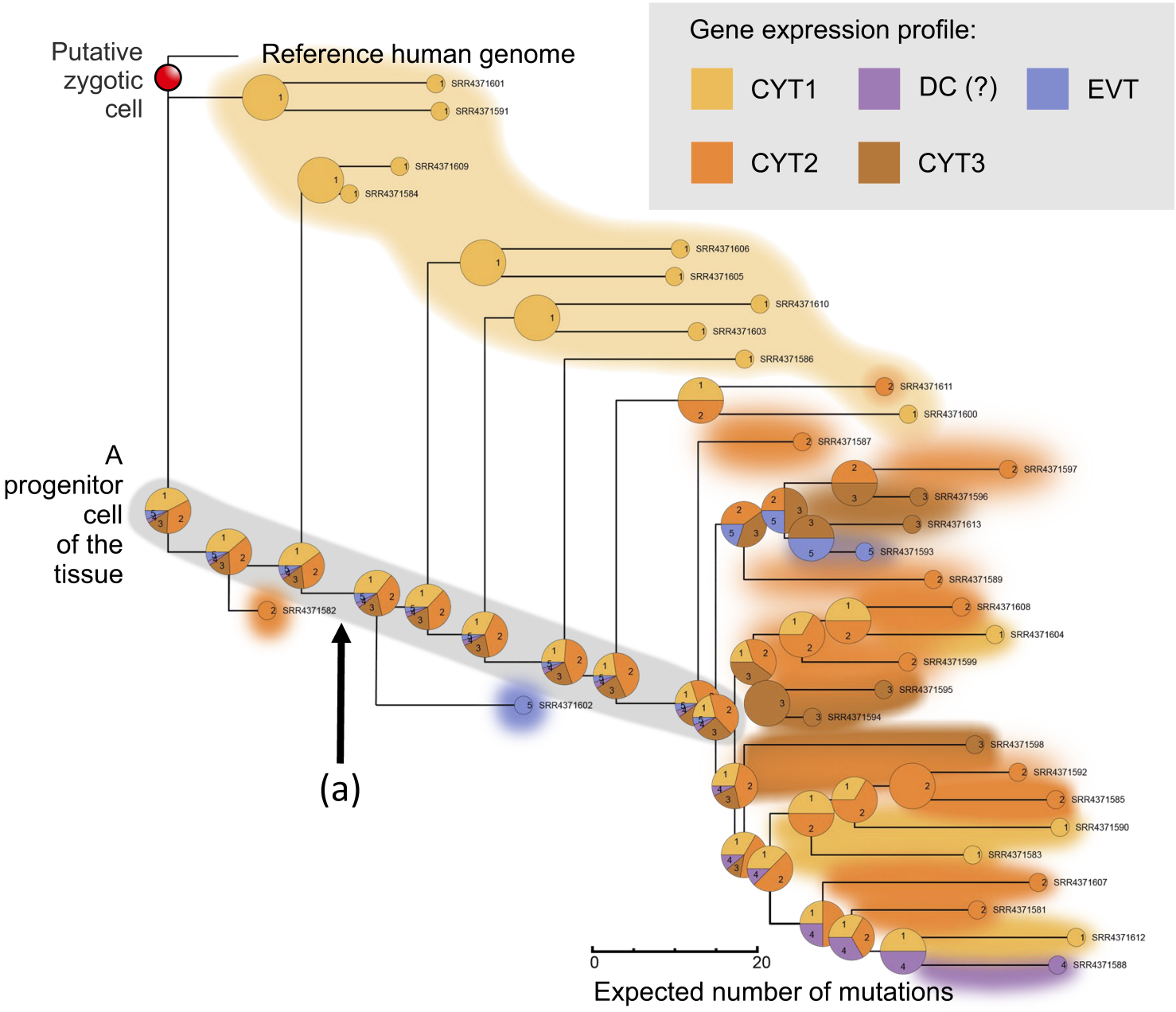

Figure 3

Mapping of the gene expression profile to a putative lineage topology (SRP090944 data; Batch2, 33 cells). (a) A putative stem cell self-renewal phase. The other notations are the same as in Figure 3.

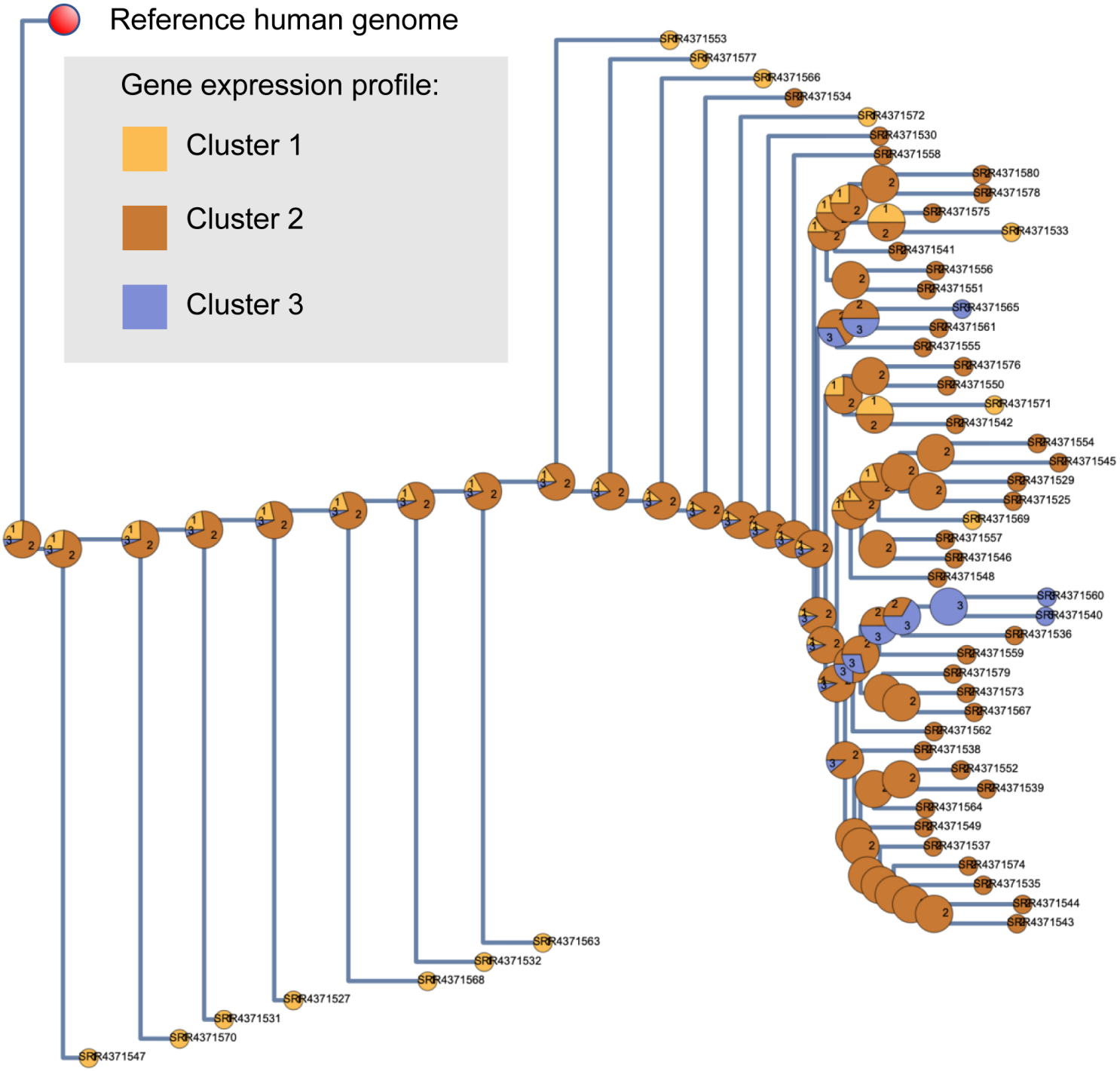

Figure 4

Mapping of the reanalyzed gene expression profile by t-SNE to the cellular trajectory tree (SRP090944 data; Batch1, 54 cells).

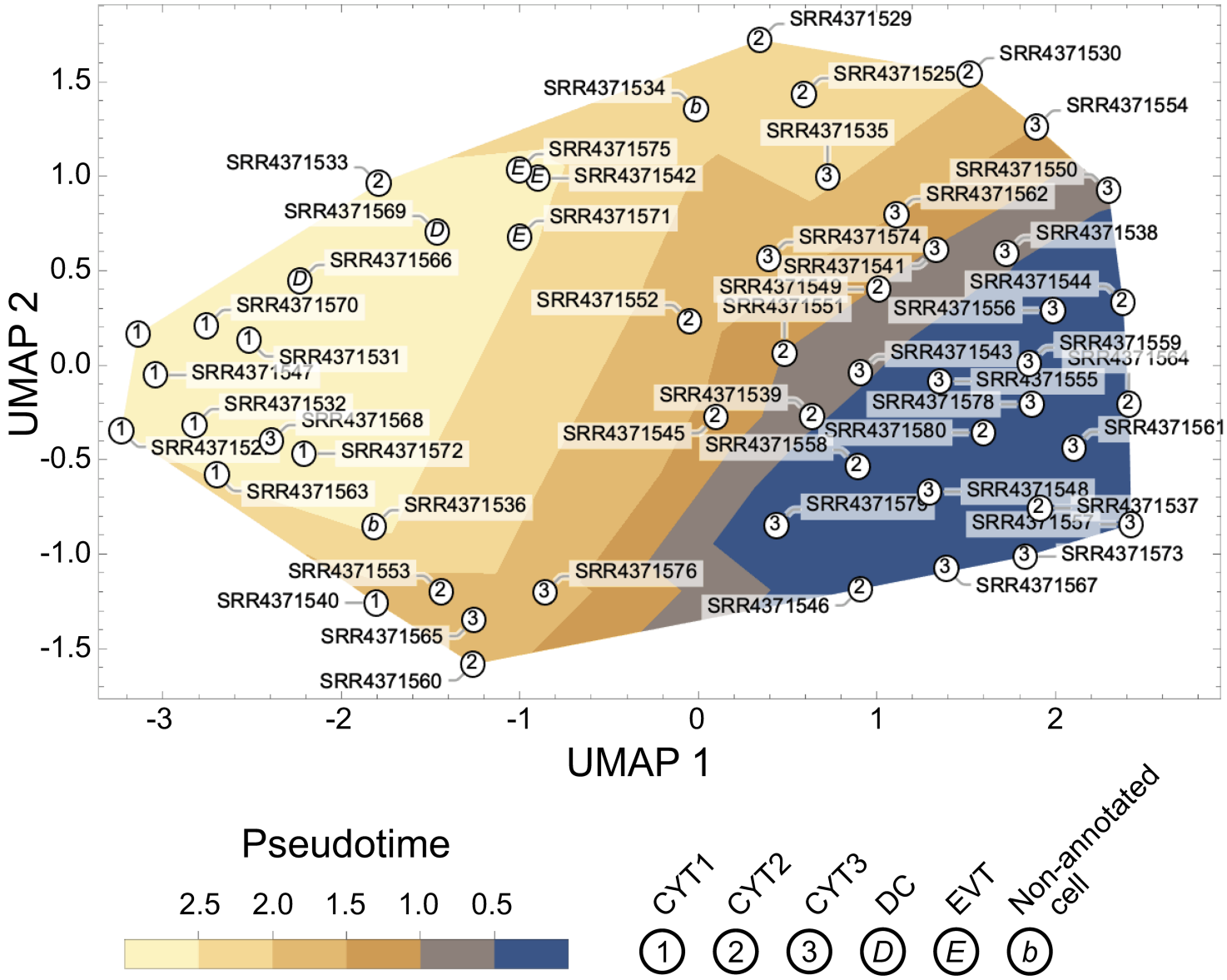

Figure 5

Results of the pseudotime course analysis of SRP090944 data Batch1. Contours were created by function of Mathematica[40] with interpolation order 3. The legend represents pseudotimes with colors. All cells are not labeled in this figure due to Mathematica algorithm. 1: intravillous cytotrophoblast (CYT) 1; 2: CYT 3; 3: CYT3; D: putative decidual cell (DC); E: extravillous cytotrophoblast (EVT); b: non-annotated cell (see Figure 3).

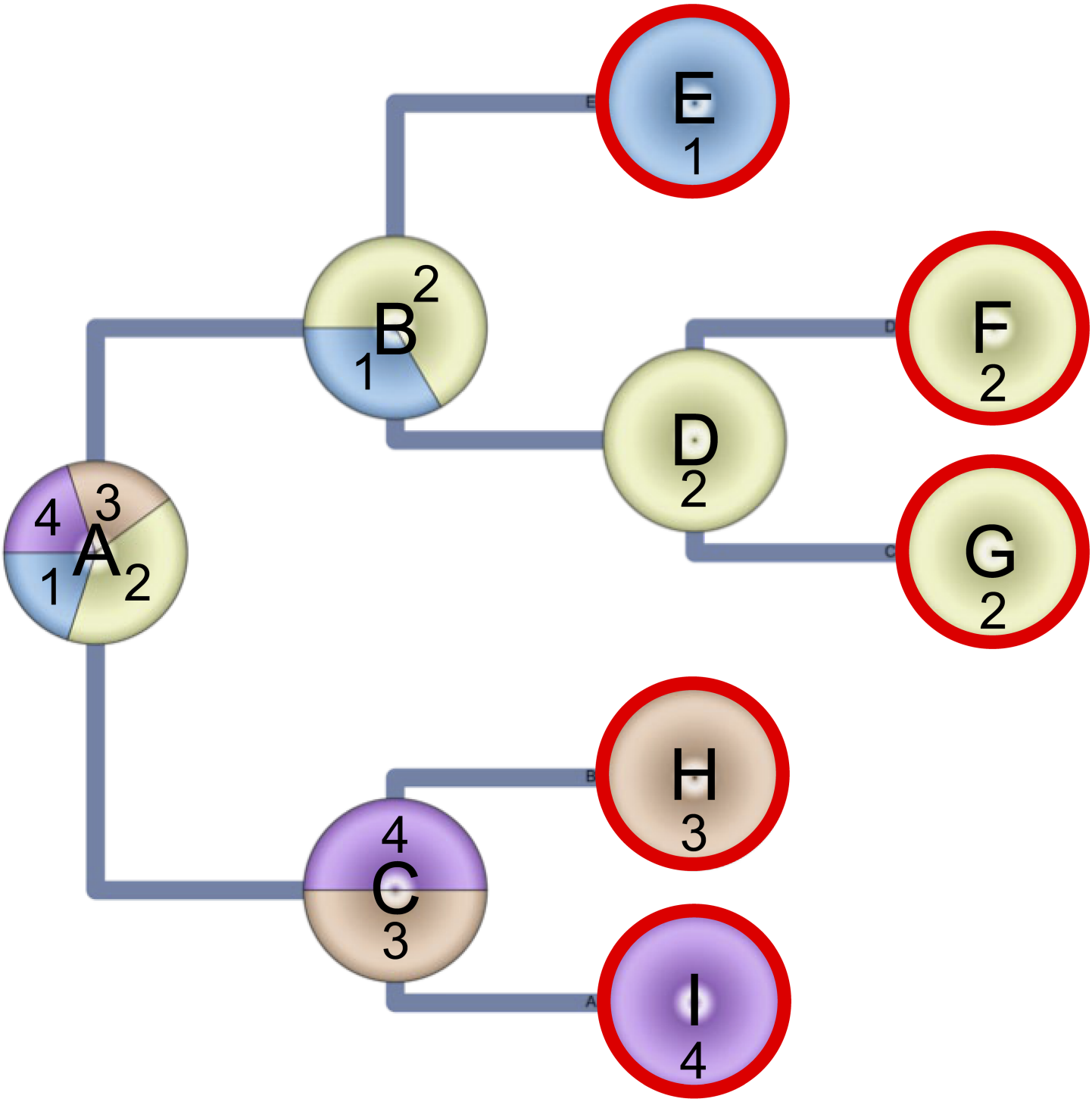

Figure 6

Evaluation of the degrees of monophyleticity (DoM) at each node in terms of the proportion of the clusters in its subtree. The gene expression profile (a cluster table) is mapped to the tree topology. If consistent with each other, we can observe a set of subtrees, each of which is monophyletic in terms of the clusters, otherwise we will observe para- or polyphyletic subtrees in terms of the clustering. We can evaluate the consistency with a size of a monophyletic subtree over the total number of the subtrees. Code *AssignCluster2Cell* calculates the proportion of the number of cells assigned to the clusters, by

which we can evaluate the DoM of being monophyletic (monophyleticity) of a subtree. A: ancestral stem cell; B, C, and D: derived stem cells; E, F, G, I, and I: observed differentiated single cells.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx