PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Accuracy of Answers to Cell Lineage Questions Depends on Single-Cell Genomics Data Quality and Quantity

Adam Spiro, Ehud Shapiro*

Department of Computer Science and Applied Mathematics and Department of Biological Chemistry, Weizmann Institute of Science, Rehovot, Israel

* ehud.shapiro@weizmann.ac.il

OPEN ACCESS

Citation: Spiro A, Shapiro E (2016) Accuracy of Answers to Cell Lineage Questions Depends on Single-Cell Genomics Data Quality and Quantity. PLoS Comput Biol 12(6): e1004983. doi:10.1371/journal.pcbi.1004983

Editor: Roger Dimitri Kouyos, University of Zurich, SWITZERLAND

Received: March 16, 2016

Accepted: May 13, 2016

Published: June 13, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by: The European Union FP7-ERC-AdG (233047); the EU-H2020- ERC-AdG (670535); the DFG (611042); the Israeli Science Foundation (ISF, P14587); the ISF-BROAD (P15439); the NIH (VUMC 38347); and the Kenneth and Sally Leafman Appelbaum Discovery Fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Advances in single-cell (SC) genomics enable commensurate improvements in methods for uncovering lineage relations among individual cells, as determined by phylogenetic analysis of the somatic mutations harbored by each cell. Theoretically, complete and accurate knowledge of the genome of each cell of an individual can produce an extremely accurate cell lineage tree of that individual. However, the reality of SC genomics is that such complete and accurate knowledge would be wanting, in quality and in quantity, for the foreseeable future. In this paper we offer a framework for systematically exploring the feasibility of answering cell lineage questions based on SC somatic mutational analysis, as a function of SC genomics data quality and quantity. We take into consideration the current limitations of SC genomics in terms of mutation data quality, most notably amplification bias and allele dropouts (ADO), as well as cost, which puts practical limits on mutation data quantity obtained from each cell as well as on cell sample density. We do so by generating *in silico* cell lineage trees using a dedicated formal language, eSTG, and show how the ability to answer correctly a cell lineage question depends on the quality and quantity of the SC mutation data. The presented framework can serve as a baseline for the potential of current SC genomics to unravel cell lineage dynamics, as well as the potential contributions of future advancement, both biochemical and computational, for the task.

Author Summary

A human cell lineage tree describes the entire developmental dynamics of a person starting from the zygote and ending with each and every extant cell. Fundamental open problems in biology and medicine are in fact questions about the human cell lineage tree: its structure and its dynamics in development, growth, renewal, aging, and disease. Consequently, a method to know the human cell lineage tree would allow resolving these problems and enable a leapfrog advance in human knowledge and health. Recent advancements in single-cell genomics have the potential to uncover various properties of the human cell lineage tree and thus promote our understanding of various biological phenomena. In this

Competing Interests: The authors have declared that no competing interests exist.

paper we present a computational framework along with specific results, which enable to understand what can be achieved using the limitations of current technologies and predict future capabilities based on future improvements. This approach can serve as a valuable tool for researchers who plan to perform lineage experiments both in designing and optimizing the actual experimental needs and predicting the costs and limitations of the plan. This work can also help researchers focus on developing what is needed for future advancements.

Introduction

Recent advances in SC technologies have generated a unique opportunity to delineate the complex behavior of heterogeneous cell populations and uncover their underlying mechanistic dynamics [1]. The use of SC genomics to reveal cell lineage relationships have been recently demonstrated in various scenarios including diseases such as cancer [2–6] and normal development [7–10]. Lineage analysis of cells sampled from an organism makes use of somatic mutations to discover common history dynamics of the sampled cells. There are several types of somatic mutations that can be used for this task, including Single Nucleotide Variations (SNV) [2, 3, 11–13], Short Tandem Repeats (STR, also called Microsatellites) [6, 8–10, 14–18], Copy Number Variations (CNV) [4, 5, 7], and Transposable Elements (TE) [8] where each type has a different mutational model and different mutation rates. This analysis is mostly effective when analyzing SC since the mixed mutational signal of cell bulks does not allow delineating mutational co-occurrences and cannot distinguish between subpopulations with different mutational patterns. Although published work have shown the great potential of using SC mutational analysis for unraveling cell lineage dynamics, there are still several major limitations, which hamper further generalization of this concept to various biological questions and prevent its use in large scale experiments. These limitations include 1) technical issues related to SC genomics, including the need for DNA amplification that introduces technical noise, 2) lack of high throughput SC isolation techniques, especially if one wants to retain the original 3D structure, or analyze rare cell types that are difficult to isolate, 3) associated costs, such as Whole Genome Amplification (WGA) kits, sequencing costs, and other consumable products (e.g., reagents and microfluidic devices), and 4) lack of computational infrastructure and dedicated algorithms specifically designed for the unique challenges of SC genomics.

The feasibility of using somatic mutations for uncovering cell lineage dynamics is dependent on these issues but also on the specifics of the pursued biological question. Some factors are inherent, such as the mutation rate and number of cell divisions, but others can be overcome by spending more money or by improving biochemical or computational procedures. Using controlled *ex-vivo* experiments is a close approximation to real biological scenarios; however, it can be very costly and laborious. Furthermore, many scenarios cannot be examined due to technical limitations in trying to mimic real biological dynamics (e.g., cell differentiation leading to changes in cellular dynamics), and also various parameter combinations cannot be studied using an *ex-vivo* experiment. A computational alternative is to model and simulate various biological scenarios using a range of parameters and conditions. Not only this approach enables to inspect the strengths and weaknesses of existing methods, it can also enable to predict the impact of future improvements.

Until now, there has not been any systematic examination of how much mutational data is required in order to accurately answer questions related to the structure and dynamics of SC lineage trees. In this work we cover few common biological settings, which capture certain tree properties such as depth (corresponds to number of cell divisions) and clustering relationships,

in order to systematically evaluate the feasibility of answering cell lineage questions using somatic mutations, and predict future capabilities by extending the range of parameters values to represent future enhancements. Using mutational data from an *ex-vivo* experiment we estimated and modeled the properties of the mutational signal quality, afflicted mainly by the random noise and ADO caused during the preprocessing and amplification of SC DNA. We then applied this model onto the signal of simulated lineage trees, generated using a dedicated formal language and simulation tool, based on environment-dependent Stochastic Tree Grammars (eSTG) [19], which is capable of generating both the entire modeled cell lineage tree and the corresponding somatic mutations accumulated through cell divisions. We present the results on a variety of parameters values, including different distance relationships (corresponding to different number of cell divisions) between different cell types, different mutation rates and two types of somatic mutations, including STR [20] and SNV. We also take into consideration current estimated costs of biochemical analysis and for each combination of parameters we calculate the cost-optimized number of cell samples and genomic loci that enable to answer the biological question with high confidence. We map the dependency between the quality and quantity of the SC mutational data and the ability to answer cell lineage questions of specific settings, which can be used as a framework for planning cell lineage experiments and predicting the potential of future enhancements, both biochemical and computational.

Results

We have previously presented a formal language, called eSTG, for describing population dynamics [19] and a corresponding programming and simulation environment, called eSTGt (eSTG tool) [21]. The language captures in broad terms the effect of the changing environment while abstracting away details on interaction among individuals. A prominent feature of the tool is that it can stochastically produce lineage trees, each corresponding to a different stochastic program execution. These lineage trees record the entire execution history of the process, including the dynamics that led to existing as well as to extinct individuals. In this paper we simulated cell lineage trees using eSTGt by specifying and executing eSTG programs. The output of each program's execution is an instance of a stochastic lineage tree, which also includes the corresponding somatic mutations as specified by the eSTG programs. By running multiple executions of the programs we collected sufficient statistics as described below. The program specifications used in this manuscript can be found in S1 File.

As mentioned above there are several types of endogenous somatic mutations, including STR, SNV, CNV and TE. Since CNV and TE have a complex dynamics and are hard to predict and model we decided to focus on STR and SNV, which are the most appropriate candidates for inferring general cell lineages retrospectively. For STR mutations, we used the stepwise mutation model [22], which assigns an equal probability p_{STR} for either an increase or a decrease of one repeat unit during each cell division (see Methods). Current estimations of the STR mutation rate p_{STR} range between 10^{-3}–10^{-5} mutations per locus per cell division depending on various factors such as the STR length, repeat type and the specific cell genotype [20]. The low mutation rate might correspond to short STRs of normal cells whereas the fast mutation rate might correspond to cells harboring Microsatellite Instability, which is common in various types of cancer cells [23]. In order to cover the entire spectrum we chose to simulate three scales of mutation rates, namely, $p_{STR} = 10^{-3}, 10^{-4}, 10^{-5}$. SNV mutations were modeled by randomly mutating each base with probability p_{SNV} following each cell division. The mutation rate p_{SNV} is estimated to be between 10^{-7}–10^{-10} mutations per nucleotide per cell division [24]. Since mutation rate of 10^{-10} was too low to yield any significant signal we present results only for mutation rates $p_{SNV} = 10^{-7}, 10^{-8}, 10^{-9}$.

As we mentioned, SC genomics poses many challenges, since the starting material consists of only one copy of each DNA molecule. DNA isolation and amplification introduce technical noise and methods for measuring and reducing it, both biochemically and computationally, are still under extensive research [1]. We chose to model two types of interferences, namely, ADO and random noise. To this end, we used data from an *ex-vivo* experiment that consisted of clonal expansions from which SCs were sampled and processed. The processing included SC Whole Genome Amplification (WGA) and sequencing of targeted loci. ADO was modeled by taking into consideration both the distribution of samples quality and genomic location, and noise was estimated by comparing the genotype of duplicates, which should be identical (see Methods). After simulating the lineage trees along with their somatic mutations we applied the models of the ADO and noise in order to generate the final mutation table that was used for further analysis. In addition, we also adjusted the parameters of the ADO and noise models in order to predict the performance of future improvements in the processing of SC genomics (see Methods). In the figures below we present results for STR using mutation rate $p_{STR} = 10^{-4}$, which may correspond to highly mutable long STR loci of normal cells, for both current and future predicted signals. Results for the other STR mutation rates (10^{-3}, 10^{-5}) and for SNV (with mutation rates 10^{-7}, 10^{-8}, 10^{-9}) for current and future signal quality are presented in the Supplementary Information.

In order to optimize the cost efficiency of a specific analysis, we used a fixed ratio of 1:1000 between the analysis cost of a single cell and the analysis cost of a single STR locus, thus one can tradeoff between the number of cells and the number of loci analyzed, depending on specific constraints such as sample scarcity or sequencing availability. In the examples below we used fixed costs of 10$ for a single cell analysis and 0.01$ for a single STR locus. These costs are based on rough estimations of current processing (e.g., WGA kits and consumables) and sequencing costs (see Methods) and can of course be adjusted as needed.

Reconstruction of Triplet Subtrees

A triplet tree consists of three leaves sampled from a (full) tree and the subtree they induce on the full tree (Fig 1A). Since there are three possible bifurcation arrangements for the triplet tree, the probability of a random triplet tree reconstruction to correctly reconstruct its topology is 1/3. In order to measure the ability to correctly reconstruct a triplet tree using somatic mutations we simulated such trees with various number of cell divisions along with the corresponding mutational signal, which was distorted with the calibrated ADO and noise. We then measured the percentage of correct reconstructions over 1000 repeated stochastic simulations. Fig 1B shows the percentage of correctly reconstructed triplet trees with various number of cell divisions (X = 2,5,10,20,40, see Fig 1A) as a function of the number of analyzed loci (ranging from 500 to 100,000) using STR mutations with mutation rate 10^{-4}. Fig 1C shows the results that correspond to future signal improvements. It can be seen, for example, that using 5 cell divisions (X = 5) and 25,000 loci the probability of correctly reconstructing a triplet tree is about 50% (compared to 33% for random reconstruction) using the current signal and almost 70% using the predicted future enhancements. Results for the other STR mutation rates and SNV are presented in S2 File.

Identifying Depth Differences

Many lineage questions are in fact questions about the depth relationship between two cell groups. Examples include questions related both to cancer dynamics and normal development or renewal. For example, is relapse after chemotherapy caused by ordinary tumor cells escaping chemotherapy stochastically, or by a separate population of rarely-dividing cancer-initiating

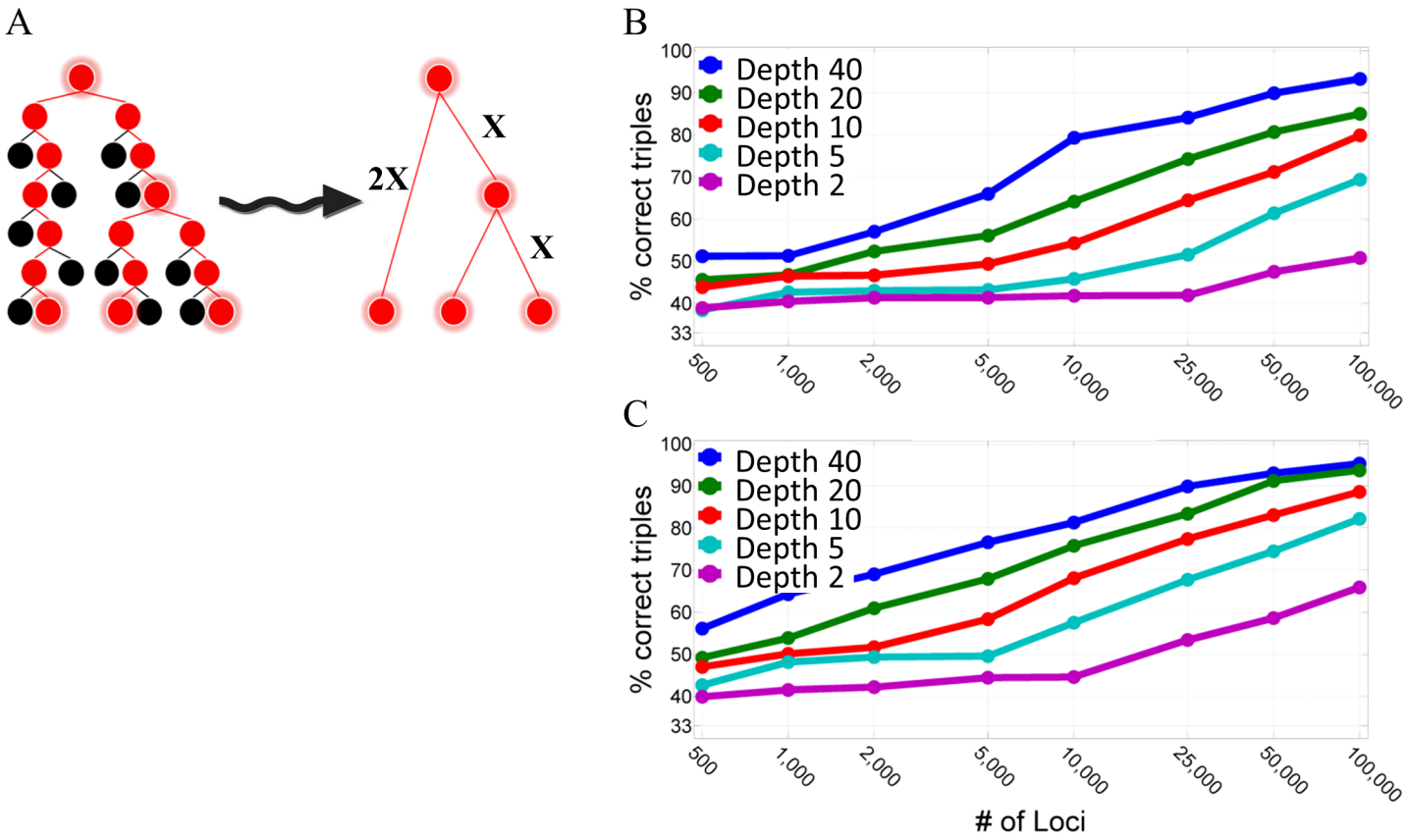

Fig 1. Reconstruction of triplet trees. (A) A triplet tree with the indicated depths (X = number of cell divisions) between the root and the leaves and between the root and the branch. **(B)** Reconstruction accuracy as a function of the number of STR loci and the depth of the leaves (number of cell divisions from the root). Results are shown for STR mutation rate of 10^{-4} mutations per locus per cell division. The graph shows that fewer mutations are needed when there are more cell divisions. The results are averaged over 1000 repeated stochastic simulations. **(C)** Same as (B) but using parameters that represent future enhancements (see Methods).

doi:10.1371/journal.pcbi.1004983.g001

cells that escape chemotherapy due to their slow division rate [6]? If relapse is initiated from slowly dividing cells, these cells would accumulate fewer mutations since they go through fewer cell divisions. By measuring the distance of the cells from the root of the tree (which can be estimated using a combination of unrelated cell bulks) we can compare the depth relationship between different cell groups. Another example question is whether the adult oocyte pool can be renewed during adulthood [10]? Again, by comparing the number of cell divisions between young and adult female, we may know whether oocytes are generated postnatally.

In order to map the feasibility of answering such questions we simulated lineage trees and analyzed two cell groups from different depths in the tree (Fig 2A). For each cell we estimated its relative depth in the tree using its mutational signature and performed a statistical test that compared the relative depth of cells from both groups (see Methods). Fig 2B shows a heatmap that represents the probability of correctly identifying a significant depth difference between the two cell groups, one of depth X and the other of depth X+Y, for X = 40 and Y = 10, as a function of the number of analyzed cells and the number of analyzed genomic loci. It can be seen that in order to obtain a specific success probability one can tradeoff between the number of analyzed samples and the number of analyzed loci (white line in Fig 2B that represents success probability of 95%), however, a minimum cost can be obtained by selecting the combination that corresponds to the minimum of the black line in Fig 2B that shows the corresponding

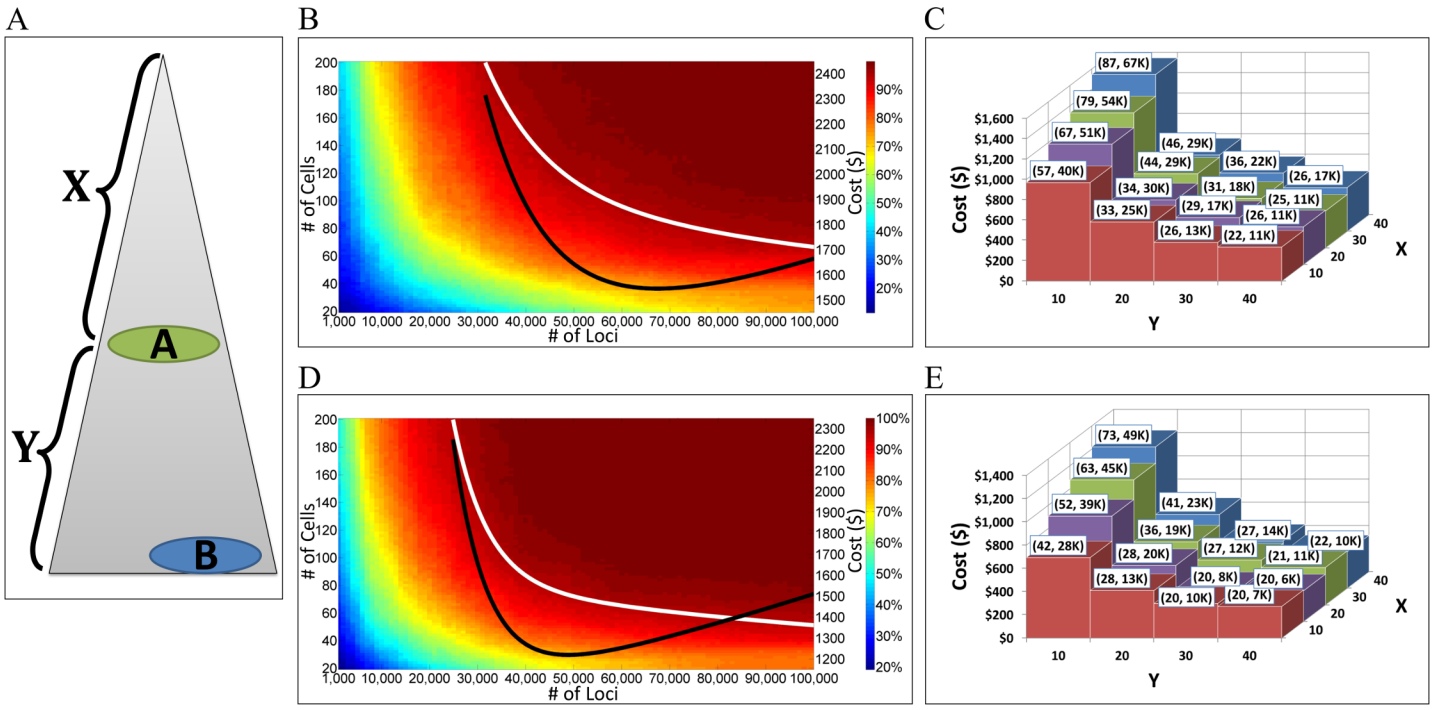

Fig 2. Experiment requirements for identifying that two cell groups have different depth on the cell lineage tree. (A) Cells from two cell groups are sampled from the cell lineage tree. The depth (number of cell divisions since the root) of cells of group A is X and the depth of cells of group B is X+Y. **(B)** The heatmap colors represent the statistical power, i.e., the probability of detecting a depth difference between cells from A and cells from B when such difference does exists, as a function of the number of cells (x-axis) and number of loci (left y-axis) analyzed. The probability of falsely identifying depth difference when it does not exist (type I error) is 5% (see Methods). White line marks the area of power = 95%. Black line indicates the overall analysis cost as shown in the right y-axis–both lines have the same x-axis and every point in the black line represents the cost that corresponds to the combination of the number of loci and number of cells as represented by the white line. In this case, for X = 40 and Y = 10, a minimum cost is obtained using about 65K loci and 90 cells. The same power can be obtained using about 35K loci and 200 cells but the cost increases by about 50%. Results are averaged over 1000 stochastic simulations using STR mutation rate of 10^{-4}. **(C)** Cost optimization for the number of loci and number of cell samples needed for statistical power of 95%, for various values of X and Y. Numbers in parenthesis indicate the number of cell samples and number of required loci respectively. **(D)** Same as (B) but using parameters that represent future enhancements affecting both the quantity and the quality of the signal (see Methods). **(E)** Same as (C) but using parameters that represent future enhancements.

doi:10.1371/journal.pcbi.1004983.g002

analysis cost. Fig 2C shows a summary of the cost-optimized number of samples and number of loci needed for obtaining success probability of 95% using various combinations of X and Y corresponding to various depths of the two cell groups (as depicted in Fig 2A). Fig 2D and 2E show the performance using enhanced parameters that correspond to future enhancements in SC genomics. Results for the other STR mutation rates and SNV are presented in S2 File.

Identifying Independent Subclones

Identifying the clonal relationship between two cell populations arises in many contexts. For example, do progenitor cells commit to a single cell-type or can they produce multiple types as needed [25]? Does geographic separation imply lineage separation or do cells migrate from one area to another [8]? Are the original tumor and its relapse independent clones [6]? The mutational signature of two cell populations can be used to perform clustering analysis in order to examine whether they are separated or intermixed in the lineage tree.

In order to investigate how well can phylogenetic analysis of somatic mutations be used for answering such questions we simulated lineage trees consisting of two subclones, which have a common ancestor of a specific distance (Fig 3A). We then estimated the distance within and between the two cell groups and performed a statistical test to check whether the two cell

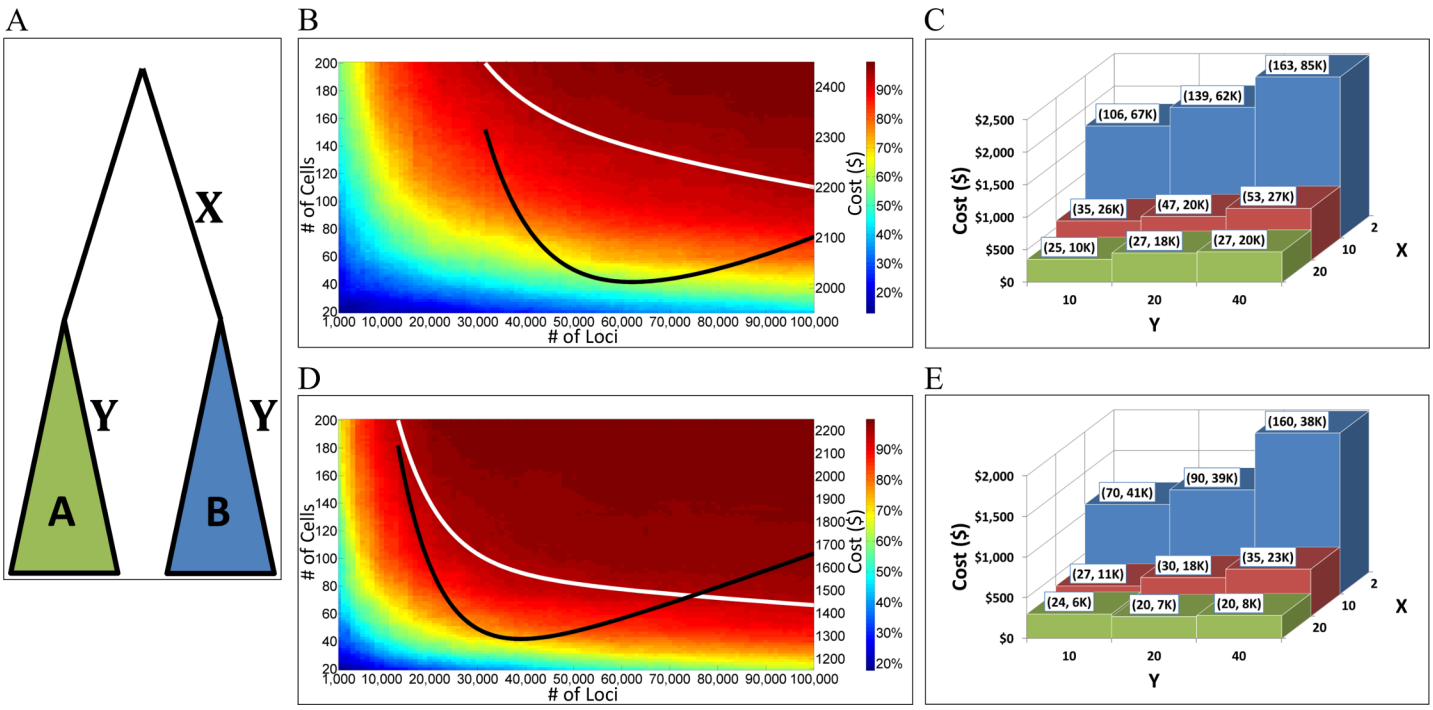

Fig 3. Experiment requirements for identifying that two cell groups are independent subclones. (A) The common root of clones A and B divides X times and two random cells of depth X generate the two clones, which divide Y times such that the depth of the extant cells of A and B is X+Y. **(B)** The heatmap colors represent the statistical power, i.e., the probability of correctly identifying the two clones as independent, as a function of the number of cells (x-axis) and number of loci (left y-axis) analyzed. The probability of falsely identifying that the two clones are separated when in fact they are mixed (type I error) is 5% (see Methods). White line marks the area of power = 95%. Black line indicates the overall analysis cost as shown in the right y-axis–both lines have the same x-axis and every point in the black line represents the cost that corresponds to the combination of the number of loci and number of cells as represented by the white line. In this case, for X = 2 and Y = 20, a minimum cost is obtained using about 60K loci and 140 cells. The same power can be obtained using about 30K loci and 200 cells but the cost increases by about 30%. Results are averaged over 1000 stochastic simulations using STR mutation rate of 10^{-4}. **(C)** Cost optimization for the number of loci and number of cell samples needed for statistical power of 95%, for various values of X and Y. Numbers in parenthesis indicate the number of cell samples and number of required loci respectively. **(D)** Same as (B) but using parameters that represent future enhancements affecting both the quantity and the quality of the signal (see Methods). **(E)** Same as (C) but using parameters that represent future enhancements.

doi:10.1371/journal.pcbi.1004983.g003

groups are separated (see Methods). Fig 3B shows a similar heatmap to Fig 2B but presents the probability of identifying that the two cell groups are independent, using X = 2 and Y = 20 (see Fig 2A). Fig 3C presents the cost-optimized combinations for various values of X and Y. Fig 3D and 3E show the performance using enhanced parameters that correspond to future enhancements. Results for the other STR mutation rates and SNV are presented in S2 File.

Discussion

During normal mitotic cell division DNA is replicated with very high, but not absolute, precision, which leads to the incorporation of somatic mutations. These somatic mutations accumulated since the zygotic stage, endow each cell in our bodies with a genomic signature that is unique with a very high probability [17]. Sequencing cell bulks for somatic mutations may supply a coarse estimation of the cell population distribution but cannot specify the deterministic position in the lineage tree of each cell and uncover population heterogeneity. Advancements in single cell genomics offer a unique opportunity to detect somatic mutations private to each cell and use them to understand the underlying dynamics of cell lineages with high precision. Unfortunately, sequencing accurately the entire genome of each single cell is still prohibitively expensive and technically challenging. In recent years there have been several attempts to use

single cells genomic data in order to uncover various lineage dynamics. These attempts included SC whole genome sequencing [26], exome sequencing [5], and genotyping of targeted loci [6], or combinations of thereof [2]. There is a tradeoff between genomic coverage and sample density and the question of finding their right quantity and balance depends on parameters such as cost, technical constraints and the specifics of the lineage question. In this paper we offer a framework for answering this question by modeling and simulating the entire process of lineage analysis taking into consideration the different aspects of SC genomics analysis, calibrated using real experiments, and possible lineage dynamics. The suggested framework can help researchers in planning and optimizing their lineage experiments and can also point out experimental aspects that should be improved in order to increase the chances for meaningful outcomes. We selected a basic triplet tree structure and two aspects of lineage questions that are widely tackled, namely identifying depth differences and identifying independent clusters, and mapped the feasibility of answering them using a wide variety of parameters, including different mutation types, different mutation rates and various combinations of distances between the cell groups. The results can serve as a guideline for planning a lineage experiment or as a reference point for tailoring a solution for a more specific setting. Future experiments can help in fine-tuning the different modeling aspects, such as ADO, noise and possible lineage scenarios. Furthermore, these aspects can also be updated as new and more advanced biochemical protocols, technological or computational tools are developed.

Methods

Mutational Models

STR mutations were modeled using the single-step model (SSM) [22]. For each STR loci of length x, its length is updated during each cell division using the following function:

$$f_{STR}(x) = \begin{cases} x+1 \ with\ probability \frac{p_{STR}}{2} \\ x-1 \ with\ probability \frac{p_{STR}}{2} \\ x \ otherwise \end{cases}$$

where p_{STR} is the mutation probability. In this paper we used three mutation scales, namely 10^{-3}, 10^{-4}, 10^{-5}, corresponding to possible STR mutations rates. We note that some STRs can display more complex mutational patterns; however, the SSM is the most common model used and constitutes a good approximation.

SNV mutations were modeled by randomly changing each base with probability p_{SNV} during each cell division, where we used three mutations scales, 10^{-7}, 10^{-8}, 10^{-9}.

We note that most chromosomes, except for the X and Y chromosomes in males, have two copies. This may introduce additional complexity to the analysis of SC genomic loci since a mutation can occur in one copy or the other. However, for MS loci this can be overcome by analyzing only sex chromosomes of males [6, 9, 10] or by analyzing loci with heterogeneous alleles that contain MS with different repeat number [27]. As for SNV analysis, the probability of a double mutation is low enough in order to allow a unique identification of random somatic mutations in each locus. Since the *ex-vivo* experimental data that we used in order to model the ADO and the noise of the SC genomic signal included mostly data from the X chromosome, we opted to analyze the simplified single allele scenario in this work. However, we are currently working on computational methods for analyzing biallelic signal, which will allow analyzing signal from autosomes and will also enable to extend the results presented here for more complex scenarios.

ADO Modeling

Since a human cell contains only one copy of a diploid genome there is a big chance that some parts of the DNA will be damaged or lost during the different amplification stages. Because of the stochastic nature of the amplification, one could also expect a relatively large variability in the amplification quality of different samples. In addition, there could be amplification biases where some parts of the genome are better amplified than others, resulting in some loci having a better chance to be detected. In order to simulate the dropout patterning of the experimental data we sought to find a modeling approach that will mimic the real behavior as much as possible. The experimental data evidently show that the allelic dropout is not random but is dependent on both the sample quality and the genomic location. In order to capture the variability of the signal quality in both the individual samples and the different loci we modeled the allelic dropout of single cell DNA samples by assigning distinct dropout probabilities for each sample and for each locus. Given M individual samples and N loci we define the *mutation table* $T = \{t_{ij}: i = 1..M, j = 1..N\}$ such that t_{ij} equals the mutation call of sample i at locus j. In the case of allelic dropout we set $t_{ij} = \emptyset$. We define the *mutation signal table* as $X = \{x_{ij}: i = 1..M, j = 1..N\}$, where

$$x_{ij} = \begin{cases} 0 \; if \; t_{ij} = \emptyset \\ 1 \; otherwise \end{cases}$$

We define $P = (p_i: i = 1..M)$ as the probability of obtaining a signal in each sample and $Q = (q_j: j = 1..N)$ as the probability of obtaining a signal in each locus. The probability of obtaining a signal in sample i and locus j thus equals $p_i q_j$.

In order to estimate these probabilities using the real *ex-vivo* data, we used a Maximum Likelihood (ML) approach. Given the mutation signal table data $X = \{x_{ij}\}$, the log likelihood is:

$$logL(P, Q; X) \propto logP(X|P, Q) = \sum_{i=1}^{M} \sum_{j=1}^{N} log(x_{ij} p_i q_j + (1 - x_{ij})(1 - p_i q_j))$$

The ML estimator of P and Q is thus:

$$\underset{P,Q}{\operatorname{argmax}}(\log(L(P, Q; X)))$$

We approximated the solution using simulated annealing and validated the results by repeating the procedure with various starting points. For the data X we used an *ex-vivo* experiment in which 167 single cells were amplified and analyzed for their genomic signal. For prediction of future enhancement we used the calculated probabilities p,q and increased their relative value by 25%.

Noise Modeling

Noise modeling differs between STR and SNV because STRs are much more prone to errors introduced during the amplification stages. For STR mutations we defined noise as the probability for each locus to randomly shift by one repeat unit compared to its true value. In order to estimate this probability we used the analysis results of duplicate cells from an *ex-vivo* experiment and measured the rate of inconsistency between supposedly identical genomes.

For SNV mutations we set the probability for noise to be 10^{-4} as measured using SC calling results of next-generation sequencing data [28].

For prediction of future enhancement we used the noise probability value divided by 2.

Cost of Analysis

We have divided the analysis cost into two parts, namely, the overhead of analyzing a single cell and the analysis cost per single locus. A detailed cost analysis is not presented in this manuscript, however, an approximation for a complete analysis of a single cell is 30$, from which 10$ are considered to be fixed overhead and 20$ are used for analyzing either 2000 STR loci or 20,000 single bases. We thus approximated the analysis cost of a single STR locus to be 20/2000 = 0.01$ and the analysis cost of a single base (SNV) to be 20/20,000 = 0.0001$.

In order to calculate the cost as presented in Figs 2 and 3 we used the following function:

$$f_{Cost} = CostLoc * x + CostSamp * y$$

where *CostLoc* = 0.01 for STR and 0.0001 for SNV, *CostSamp* = 10, *x* = *# of loci*, *y* = *# of samples* and *x*,*y* are constrained to the white line in Figs 2 and 3 (corresponding to success probability of 95%). Minimal cost is obtained by finding the minimum of f_{Cost}.

Tree Reconstruction Algorithm

For the triplet trees reconstruction we used the Neighbor-Joining (NJ) algorithm [29] with the absolute distance function:

Given a mutation table $T = \{T_i^l; i = 1..M, l = 1..N\}$, with M samples and N loci, where T_i^l is the genotyping of locus l in sample i, the distance between each two samples is:

$$D(i,j) = \frac{1}{N}\sum_{l=1}^{N} |T_i^l - T_j^l|$$

where only loci with signal in both samples are counted. For the three example samples with the following 5 loci genotype:

$$T_1 = (10, \emptyset, \emptyset, 8, 12)$$

$$T_2 = (12, \emptyset, 7, 8, \emptyset)$$

$$T_3 = (10, \emptyset, 7, 8, 11)$$

where $\emptyset$ means that there is no signal in that locus, the distances are:

$$D(1,2) = \frac{1}{2}(|10-11| + |8-8|) = \frac{1}{2}(1+0) = \frac{1}{2}$$

$$D(1,3) = \frac{1}{2}(|10-12| + |8-8|) = \frac{1}{2}(2+0) = 1$$

$$D(2,3) = \frac{1}{3}(|12-10| + |7-7| + |8-8|) = \frac{1}{3}(2+0+0) = \frac{2}{3}$$

The result of the NJ tree reconstruction algorithm on these samples is depicted in S1 Fig.

We note that alternatives to distance-based methods for phylogeny estimation exist, which might yield better results or improve the cost efficiency; however, analyzing or developing such methods is not in the scope of this paper and is a subject of an ongoing research in our lab.

Measuring Significant Depth Differences

Given two groups of cells $A = \{a_i\}$ and $B = \{b_j\}$ we define a binary classifier f that decides whether there is a depth difference between them or not. We define $D(x)$ as the distance between the cell x and the root of the tree where D is calculated using the absolute distance function as defined above. We define the set $D(X) = \{D(x)\}_{x \in X}$ where X is a group of cells. We define $ttest(D(A),D(B))$ as the p-value obtained from a t-test between the set of distances $D(A)$ and $D(B)$. The classifier f is defined as follows:

$$f(A,B) = \begin{cases} 1 \; if \; ttest(D(A), D(B)) \leq 0.05 \\ 0 \; otherwise \end{cases}$$

where $f(A,B) = 1$ means that there is a significant distance between the cell groups A and B.

In the words of hypothesis testing, if we define the null hypothesis to be that there is no depth difference between A and B then from the definition of f if the depth of the two populations is equally distributed the probability of incorrectly rejecting the null hypothesis, i.e., the type I error, is 5% and the statistical power is depicted in Fig 2B.

Measuring Significant Independent Clustering

Similarly to the case of depth differences, we define $D(x,y)$ as the distance between cell x and cell y, and $D(X,Y) = \{D(x,y)\}_{x \in X, y \in Y}$. We define the clustering classifier f to be:

$$f(A,B) = \begin{cases} 1 \; if \; ttest(D(A,A), D(A,B)) \leq 0.05 \\ 0 \; otherwise \end{cases}$$

i.e., we measure the difference in the average distance of cells within the group A and the distance of cells between group A and group B. Similarly to the case of the depth differences, the type I error is 5% and the statistical power is depicted in Fig 3B.

We note that the measures presented here for identifying significant depth differences and clustering are used for proof of concept and there may be better ones. However, finding better measures is not in the scope of this paper and is a subject of future research.

Supporting Information

S1 File. eSTG programs of the simulated lineage trees.
(ZIP)

S2 File. Results of cell lineage analysis. Results include figures similar to Figs 1B, 2B and 3B but using STR mutations with mutation rates 10^{-3} and 10^{-5} and SNV mutations with mutation rates 10^{-7}, 10^{-8}, 10^{-9} using current and improved values for ADO and noise corresponding to future quality enhancements of single cell genomics.
(ZIP)

S1 Fig. Example of NJ tree reconstruction of a triplet.
(TIF)

Acknowledgments

We would like to thank Tamir Biezuner for performing the *ex-vivo* experiments and helping in calibrating the different models parameters, and Yoni Herzog for helping in the writing of the eSTG programs. Ehud Shapiro is the Incumbent of The Harry Weinrebe Professorial Chair of Computer Science and Biology.

Author Contributions

Conceived and designed the experiments: AS ES. Performed the experiments: AS. Analyzed the data: AS. Contributed reagents/materials/analysis tools: AS. Wrote the paper: AS ES.

References

1. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013; 14(9):618–30. doi: 10.1038/nrg3542 PMID: 23897237
2. Gawad C, Koh W, Quake SR. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proc Natl Acad Sci U S A. 2014; 111(50):17947–52. doi: 10.1073/pnas.1420822111 PMID: 25425670
3. Lohr JG, Adalsteinsson VA, Cibulskis K, Choudhury AD, Rosenberg M, Cruz-Gordillo P, et al. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. Nature Biotechnology. 2014; 32:479–84. doi: 10.1038/nbt.2892 PMID: 24752078
4. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011; 472(7341):90–4. doi: 10.1038/nature09807 PMID: 21399628
5. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014; 512(7513):155–60. doi: 10.1038/nature13600 PMID: 25079324
6. Shlush LI, Chapal-Ilani N, Adar R, Pery N, Maruvka Y, Spiro A, et al. Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and miscrosatellite instability. Blood. 2012.
7. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, et al. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell Rep. 2014; 8(5):1280–9. doi: 10.1016/j.celrep.2014.07.043 PMID: 25159146
8. Evrony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai X, et al. Cell lineage analysis in human brain using endogenous retroelements. Neuron. 2015; 85(1):49–59. doi: 10.1016/j.neuron.2014.12.028 PMID: 25569347
9. Reizel Y, Chapal-Ilani N, Adar R, Itzkovitz S, Elbaz J, Maruvka YE, et al. Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. PLoS Genet. 2011; 7(7):e1002192. Epub 2011/08/11. doi: 10.1371/journal.pgen.1002192 PMID: 21829376
10. Reizel Y, Itzkovitz S, Adar R, Elbaz J, Jinich A, Chapal-Ilani N, et al. Cell lineage analysis of the mammalian female germline. PLoS Genet. 2012; 8(2):e1002477. Epub 2012/03/03. doi: 10.1371/journal.pgen.1002477 PMID: 22383887
11. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012; 148(5):873–85. doi: 10.1016/j.cell.2012.02.028 PMID: 22385957
12. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012; 148(5):886–95. doi: 10.1016/j.cell.2012.02.025 PMID: 22385958
13. Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science. 2015; 350(6256):94–8. doi: 10.1126/science.aab1785 PMID: 26430121
14. Segev E, Shefer G, Adar R, Chapal-Ilani N, Itzkovitz S, Horovitz I, et al. Muscle-bound primordial stem cells give rise to myofiber-associated myogenic and non-myogenic progenitors. PLoS One. 2011; 6(10):e25605. Epub 2011/10/25. doi: 10.1371/journal.pone.0025605 PMID: 22022423
15. Frumkin D, Wasserstrom A, Itzkovitz S, Stern T, Harmelin A, Eilam R, et al. Cell lineage analysis of a mouse tumor. Cancer Res. 2008; 68(14):5924–31. Epub 2008/07/18. doi: 10.1158/0008-5472.CAN-07-6216 PMID: 18632647
16. Wasserstrom A, Adar R, Shefer G, Frumkin D, Itzkovitz S, Stern T, et al. Reconstruction of cell lineage trees in mice. PLoS One. 2008; 3(4):e1939. doi: 10.1371/journal.pone.0001939 PMID: 18398465
17. Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E. Genomic variability within an organism exposes its cell lineage tree. PLoS Comput Biol. 2005; 1(5):e50. PMID: 16261192
18. Luo T, He X, Xing K. Lineage analysis by microsatellite loci deep sequencing in mice. Mol Reprod Dev. 2016.
19. Spiro A, Cardelli L, Shapiro E. Lineage grammars: describing, simulating and analyzing population dynamics. BMC Bioinformatics. 2014; 15:249. doi: 10.1186/1471-2105-15-249 PMID: 25047682

20. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004; 5(6):435–45. Epub 2004/05/22. PMID: 15153996

21. Spiro A, Shapiro E. eSTGt: a programming and simulation environment for population dynamics. BMC Bioinformatics. 2016; 17(1):187. doi: 10.1186/s12859-016-1004-y PMID: 27117841

22. Ohta T, Kimura M. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res. 1973; 22(2):201–4. PubMed PMID: 4777279

23. Imai K, Yamamoto H. Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. Carcinogenesis. 2008; 29(4):673–80. PMID: 17942460

24. Lynch M. Evolution of the mutation rate. Trends Genet. 2010; 26(8):345–52. doi: 10.1016/j.tig.2010.05.003 PMID: 20594608

25. Ming GL, Song H. Adult neurogenesis in the mammalian brain: significant answers and significant questions. Neuron. 2011; 70(4):687–702. doi: 10.1016/j.neuron.2011.05.001 PMID: 21609825

26. Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature. 2014; 513(7518):422–5. doi: 10.1038/nature13448 PMID: 25043003

27. Salipante SJ, Kas A, McMonagle E, Horwitz MS. Phylogenetic analysis of developmental and postnatal mouse cell lineages. Evol Dev. 2010; 12(1):84–94. doi: 10.1111/j.1525-142X.2009.00393.x PMID: 20156285

28. Huang L, Ma F, Chapman A, Lu S, Xie XS. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. Annu Rev Genomics Hum Genet. 2015; 16:79–102. doi: 10.1146/annurev-genom-090413-025352 PMID: 26077818

29. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4(4):406–25. PMID: 3447015