Department of Computational Biology and Medical Sciences
Graduate School of Frontier Sciences

The University of Tokyo

2016
Master's Thesis

Inferring tumor clonal evolution utilizing population genetics
（集団遺伝学を利用したがんのクローン進化過程の推定）

Submitted January 27, 2017

Advisor: Associate Professor Hisanori Kiryu

Yutaro Konta

Abstract

Tumor is caused by the somatic mutation accumulation. Every cell within a tumor has derived from a single founder cell, whose subsequent accumulation of advantageous mutations causes clonal expansions. In the course of clonal expansion, a driver mutation gives rise to another type of clone, which is called a subtype. As a result, a tumor is a mixture of various subtypes. The emergence of the next-generation sequencers (NGSs) has enabled us to analyse whole cancer genomes at a single nucleotide resolution. However, the subtype reconstruction using bulk sequencing reads has many difficulties because the observed variant allele frequencies (VAFs) does not directly reflect those of each subtypes. The observed VAFs are intertwined with the abundance ratio of each subtypes. Thus our problem is to identify what kinds of subtypes the tumor consists of and to identify the characteristics of each subtype from NGS reads of the bulk tumor. To solve this problem, several methods such as PyClone and AncesTree have been proposed in previous works. However, we cannot estimate how rapidly each subtype proliferates and when these subtypes arose using these methods. Here we provide a statistical model to estimate birth time and growth rate of each subtype. Our method models the allele frequency drift in each subtype with diffusion equation applying Wright-Fisher process, enabling the inference of the birth time and growth rate of each subtype. From the model observation varying the birth time parameters, the earlier the subtype arose, the higher the variant allele frequency (VAF) fixation probability was. Conversely, we exploit the shape of VAF distribution to estimate the birth time and abundance ratio of each subtype. We integrated this population genetics model with mixture modeling to infer the birth time of each subtype. Maximum likelihood estimates of the birth time and abundance ratio can be obtained using Expectation-Maximization algorithm. Using the simulated NGS reads, we could estimate the birth time and abundance ratio of all subtypes if there are large number of SNVs in the given data.

Contents

1 Introduction

Cancer arises as a result of the somatic mutation accumulation. Every cell within a tumor has derived from a single founder cell, whose subsequent accumulation of advantageous mutations causes clonal expansion. In the course of clonal expansion, a driver mutation gives rise to another type of clone, which is called a subtype. As a result, a tumor is a mixture of various subtypes. Such a mechanism of cancer progression is called clonal evolution [1].

In the latest cancer treatment, it is important to identify what subtypes the tumor consists of and to identify the growth rates of these subtypes. For example, breast cancer subtypes have been studied well, and the clinical practice guidelines depending on each subtype have been established. Although well studied subtypes such as luminal A, luminal B, and HRE2+ could be identified biochemically and immunohistochemically [2], however, there is little as yet known about the growth rates of them. Furthermore, patient specific subtypes and unrevealed subtypes of other cancers cannot be identified with such chemical tests.

The emergence of the next-generation sequencers (NGSs) has enabled us to analyse whole cancer genomes at a single nucleotide resolution. For example, two different clonal evolution patterns were revealed in the relapsed acute myeloid leukemia by whole-genome sequencing [3], and highly individual evolutional trajectories were identified in the high-grade serous ovarian cancer using exome sequencing [4].

Furthermore, using the latest single-cell sequencing technology, we can investigate the copy number and the genotype of each cell to identify whole subtypes in a tumor [5]. However, sequencing a bulk tumor is still common because of technical difficulties and high cost of the single cell sequencing. Thus, our problem is to identify what kinds of subtypes the tumor consists of and to identify the characteristics of each subtype from NGS reads of the bulk tumor.

However, the subtype reconstruction using bulk sequencing reads has many difficulties because the observed variant allele frequencies (VAFs) does not directly refrect those of each subtypes. The observed VAFs are interwined with the normal cell contamination, copy number alterations (CNAs), and the abundance ratio of each subtypes.

One of the first computational methods which tackled the first problem mentioned above is ABSOLUTE [6], which enabled the estimation of the tumor purity and the ploidy avaraged over all the subtypes. However, they did not attempt to decompose these subtypes. Then THetA [7] and TITAN [8] were desined to tackle the second and third problem mentioned above. However, they can infer the copy number and the abundance ratio of each subtype only if the CNAs distinguish the subtypes.This is because they only use the number of reads mapped to each genomic interval in the case of tumor and normal sample as inputs. On the other hand, subtype abundance ratio estimation using the single nucleotide variations (SNVs) was conducted by SciClone [9] and PyClone [10]. They utilize Bayesian mixture modeling to cluster mutations which have similar VAF to infer the abundance ratio of each subtype. However, neither of them infers the phylogenetic relationship between subtypes. After that, SCHISM [11], LICHeE [12] and AncesTree [13] enabled us to reconstruct phylogenetic relationship as an acyclic directed graph using multiple spacially distinct samples from the same tumor. They made use of the infinite-sites assumption, which assumes that no genomic position, or locus, mutates more than once in the course of clonal evolution. This assumption resolves the ambiguity of the phylogenetic relationship, because the mutations harbored by smaller number of cancer cells cannot be ancestral to the mutations harbored by larger number of cancer cells in all the tumor samples. While LICHeE [12] and AncesTree [13] are deterministic algorithms, PhyloSub [14] and BitPhylogeny [15] are the Bayesian approaches to detect major phylogeny by sampling trees using tree-structured stick-breaking process.

The subtype reconstruction from bulk sequencing reads has been progressed in this way, however, none of these methods can infer the birth time and growth rate of each subtype, which are the important subtype characteristics in the latest cancer treatment. This is fundamentally because they only make use of driver mutations and neglect passenger mutations. While the formar are the advantageous mutations which cause the phenotype of each cancer subtype, the latter have no effect in the cancer phenotype and obeys neutral evolution. Although driver mutations have higher VAFs compared to passenger mutations because they are fixed in each subtype, passenger mutations account for the overwhelming majority of somatic mutation events [16].

Theoretical studies of the passenger mutations is carried out using population genetics. Wright-Fisher process [17], in which each offspring's allele is drawn at random from its parent's allele, is the basis of the allelic frequency drift. However, it cannot be directly applied to the exponentially growing tumor cell population because Wright-Fisher process assumes that the population is constant. The fraction of the variant cells in the exponentially growing tumor cell population is calculated using branching process [18]. After that, the intratumor heterogeneity in the clonal evolution is simulated by the multitype brancing process [19]. Mathematical framework of the passenger mutation in exponentially growing cell population is given in [20].

However, these theoretical studies ended up with simulation, thus these studies could not exploit the sequencing data to infer the model parameters. Here we can make use of the statistical analysis of passenger mutations to reveal the more detailed characteristics of each cancer subtype. More particularly, the VAF distribution of the passenger mutations changes depending on the duration of their genetic drift process. Using this relationship conversely, we can infer the birth time of each subtype from the VAF distribution of the passenger mutations.

In a previous study, if there exists a single subtype within a tumor, we can infer the birth time using linear regression using this relationship [21]. However, if there exists multiple subtypes, linear regression does not work because the VAF distribution has multiple peaks. More over, the VAF distribution modeling in the previous study is not acculate because it neglects the effect of loss and fixation of the VAF because they dose not consider the stochastic nature of the genetic drift.

Here I provide a mixture modeling to estimate the birth time of each subtype to estimate birth time and growth rate of each subtype. Our method models the allele frequency drift in each subtype with diffusion equation applying Wright-Fisher process [17], enabling the inference of the birth time and growth rate of each subtype.

2 Methods

In this section, I describe a model for the clonal evolution of cancer cell populations and the generation of NGS data from cancer. I assume that each cancer cells within a tumor has derived from a single founder cell (Figure 1), and I only modeled somatic single nucleotide variants (SNVs) which are not influenced by copy number aberrations or rearrangements. I assume, as in previous studies, that each SNVs follow infinites sites assumption; i.e. no genomic position, or locus, mutates more than once in the course of clonal evolution. Thus, all SNVs are heterozygous, and there are only two genotypes, *AA* and *AB* at every locus, where *A* and *B* each denotes the normal and variant allele.

Each SNVs are divided into two types of mutations, driver mutations and passenger mutations. The former are the mutations which are subject to the natural selection and the latter are the ones which obey neutral evolution. I modeled that a driver mutation which occurred within each subtype give rise to another subtype, thus there are ancestral relationships among subtypes. The anestral relationships could be described using a phylogenetic tree where each nodes represent different cancer subtypes and edges represent the parent-child relationships between subtypes.

On the other hand, numerous passenger mutations occur within each subtypes. The variant allele frequency (VAF) of each passenger mutation within each subtypes diffuses according to genetic drift because the cells with that passenger mutation and the cells without it divide randomly. Each passenger mutation occurred within each subtypes could be inherited by its child subtypes if and only if the driver mutation which triggers the child subtype occurs in the cells which carries that passenger mutation.

I assumed that each cancer cell populations grow exponentially, and the growth rate differs from subtype to subtype but remains unchanged in the course of clonal evolution. Under these conditions, there is no clonal interference between subtypes.

I denote the fraction of subtype i's cells among all cells at observation time as n_i, the birth time of subtype i as t_i, and the growth rate of subtype i as α_i. Time is defined so that observation time is $t = 0$ and the time when all the cells within a tumor coalesce into a single normal cell is $t = 1$. Then, the subtype i's population at time t, $Nn_i(t)$, can be calculated as follows.

$$n_i(t) = \frac{1}{N}\mathrm{e}^{\alpha_i(t_i - t)},$$

where N is the number of cells at the observation time ($t = 0$).

And I denote the fraction of subtype i's cells among all cells at the observation time as n_i, which is defined as follows,

$$n_i \coloneqq n_i(0) = \frac{1}{N}\mathrm{e}^{\alpha_i t_i}. \tag{1}$$

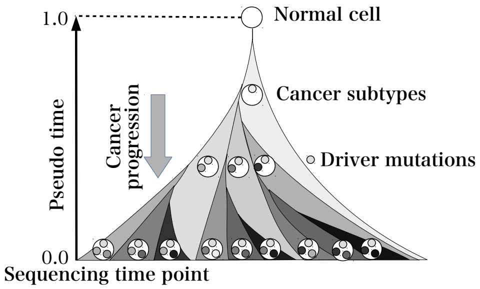


Figure 1: Clonal evolution of the tumor. A driver mutation gives rise to another subtype, causing a clonal expansion.

2.1 Variant allele frequency distribution under neutral evolution

To derive the mean and the variance of the allelic frequency drift from generation g to generation $g+1$, I denote the number of alleles which double during generation g to $g+1$ as B_g, the number of alleles which remain unchanged as L_g, and the number of alleles which dissapear as D_g. Putting that there are a_{B_g} variant alleles within B_g doubling alleles, a_{L_g} variant alleles within L_g alleles, a_{D_g} variant alleles within D_g dissapearing alleles, the probability that we choose a_{B_g}, a_{L_g}, and a_{D_g} among B_g, L_g, and D_g obeys the following hypergeometric distribution.

$$p(a_{B_g}, a_{L_g}, a_{D_g} | A_g, B_g, L_g, D_g) = \frac{\binom{B_g}{a_{B_g}}\binom{L_g}{a_{L_g}}\binom{D_g}{a_{D_g}}}{\binom{N_g}{A_g}},$$

where N_g is the number of total alleles in generation g ($N_g = B_g + L_g + D_g$) and A_g is the number of total variant alleles in generation g ($A_g = a_{B_g} + a_{L_g} + a_{D_g}$). N_g is equivalent to the number of subtype i's cells in generation g.

Then, the mean and the variance of a_{B_g}, a_{L_g}, and a_{D_g} are derived as follows.

$$E[a_{B_g}|A_g, B_g, L_g, D_g] = \frac{B_g}{N_g}A_g,\ \ E[a_{L_g}|A_g, B_g, L_g, D_g] = \frac{L_g}{N_g}A_g,\ \ E[a_{D_g}|A_g, B_g, L_g, D_g] = \frac{D_g}{N_g}A_g$$

$$\begin{aligned}
V[a_{B_g}|A_g, B_g, L_g, D_g] &= \frac{B_g}{N_g}\left(1 - \frac{B_g}{N_g}\right)A_g\frac{N_g - A_g}{N_g - 1} \\
V[a_{L_g}|A_g, B_g, L_g, D_g] &= \frac{L_g}{N_g}\left(1 - \frac{L_g}{N_g}\right)A_g\frac{N_g - A_g}{N_g - 1} \\
V[a_{D_g}|A_g, B_g, L_g, D_g] &= \frac{D_g}{N_g}\left(1 - \frac{D_g}{N_g}\right)A_g\frac{N_g - A_g}{N_g - 1}.
\end{aligned}$$

And the covariance between a_{B_g} and a_{D_g} is calculated as follows.

$$Cov[a_{B_g}, a_{D_g}|A_g, B_g, L_g, D_g] = -\frac{B_g}{N_g}\frac{D_g}{N_g}A_g\frac{N_g - A_g}{N_g - 1}.$$

I assume that there is no overlapping generation and that each allele in generation $g+1$ is drawn independently from the alleles in generation g. These assumptions are the same as those of the Wright-Fisher process, however, the constant population Wright-Fisher process does not apply to this clonal evolution problem because the tumor cell population grows exponentially. Thus, I must derive the mean and the variance of this variant allele frequency drift.

Noting that $A_{g+1} = 2a_{B_g} + a_{L_g}$, $A_{g+1} - A_g = a_{B_g} - a_{D_g}$, $A_{g+1} - A_g$ has the following variance.

$$\begin{aligned} V[A_{g+1} - A_g | A_g, B_g, L_g, D_g] &= V[a_{B_g} - a_{D_g} | A_g, B_g, L_g, D_g] \\ &= V[a_{B_g} | A_g, B_g, L_g, D_g] + V[a_{D_g} | A_g, B_g, L_g, D_g] - 2Cov[a_{B_g}, a_{D_g} | A_g, B_g, L_g, D_g] \\ &= \left[\left(\frac{B_g}{N_g} + \frac{D_g}{N_g} \right) - \left(\frac{B_g}{N_g} - \frac{D_g}{N_g} \right)^2 \right] A_g \frac{N_g - A_g}{N_g - 1} \end{aligned}$$

When I denote the variant allele frequency (VAF) in generation g as X_g, the difference of the VAF from generation g to $g + 1$ is,

$$\begin{aligned} X_{g+1} - X_g &= \frac{A_{g+1}}{N_{g+1}} - \frac{A_g}{N_g} \\ &= \frac{A_g + a_{B_g} - a_{D_g}}{N_g + B_g - D_g} - \frac{A_g}{N_g} \\ &= \frac{A_g}{N_g} \left\{ \left(1 + \frac{B_g - D_g}{N_g} \right)^{-1} - 1 \right\} + \frac{a_{B_g} - a_{D_g}}{N_g} \left(1 + \frac{B_g - D_g}{N_g} \right)^{-1} . \end{aligned}$$

Thus, noting that X_g is given because A_g and $N_g = B_g + L_g + D_g$ are given, the mean difference of the VAF from generation g to $g + 1$ is calculated as follows,

$$\begin{aligned} E[X_{g+1} - x | X_g = x] &= \frac{A_g}{N_g} \left\{ \left(1 + \frac{B_g - D_g}{N_g} \right)^{-1} - 1 \right\} + \frac{E\left[a_{B_g} - a_{D_g} \right]}{N_g} \left(1 + \frac{B_g - D_g}{N_g} \right)^{-1} \\ &= \frac{A_g}{N_g} \left\{ \left(1 + \frac{B_g - D_g}{N_g} \right)^{-1} - 1 \right\} + \frac{B_g - D_g}{N_g} \frac{A_g}{N_g} \left(1 + \frac{B_g - D_g}{N_g} \right)^{-1} \\ &= 0 \end{aligned} \tag{2}$$

On the other hand, the variance of the difference of the VAF from generation g to $g + 1$ is calculated as follows,

$$\begin{aligned} V[X_{g+1} - x | X_g = x] &= \frac{1}{N_g^2} \left(1 + \frac{B_g - D_g}{N_g} \right)^{-2} V[a_{B_g} - a_{D_g}] \\ &= \frac{1}{N_g^2} \left(1 + \frac{B_g - D_g}{N_g} \right)^{-2} \left[\left(\frac{B_g}{N_g} + \frac{D_g}{N_g} \right) - \left(\frac{B_g}{N_g} - \frac{D_g}{N_g} \right)^2 \right] A_g \frac{N_g - A_g}{N_g - 1} \\ &= \frac{1}{N_{g+1}^2} \left[\frac{2B_g + N_g - N_{g+1}}{N_g} - \left(\frac{N_{g+1} - N_g}{N_g} \right)^2 \right] A_g \frac{N_g - A_g}{N_g - 1} \\ &= \frac{1}{N_{g+1}} \left(1 - \frac{1}{N_g} \right)^{-1} \left(\frac{2B_g}{N_{g+1}} - \frac{N_{g+1} - N_g}{N_g} \right) \frac{A_g}{N_g} \left(1 - \frac{A_g}{N_g} \right) \\ &= \frac{N_g}{N_{g+1}} \frac{1}{N_g - 1} \left(\frac{2B_g}{N_{g+1}} - \frac{N_{g+1}}{N_g} + 1 \right) X_g \left(1 - X_g \right) . \end{aligned}$$

If we assume that there are N generations in total, and choose unit time $\varDelta t$ (i.e. the time between generation g to $g+1$) to be $1/N$,

$$\begin{aligned} N_g &= N n_i((N-g)\varDelta t) \\ &= N n_i\left(1 - \frac{g}{N}\right) \\ &= \exp\left[\alpha_i\left\{t_i - \left(1 - \frac{g}{N}\right)\right\}\right]. \end{aligned}$$

Thus,

$$\frac{N_{g+1}}{N_g} = \mathrm{e}^{\alpha_i/N} \sim 1 \ \ (\because N >> 1)\,.$$

$2B_g/N_{g+1}$ represents the fraction of the alleles which double between the generation and we denote it as β. If β is large, diffusion of the allele frequency is greater, thus β characterize the the strength of the genetic drift.

$$\begin{aligned} V[X_{g+1} - x|X_g = x] &\sim \frac{\beta}{N_g} X_g\left(1 - X_g\right), \\ \frac{1}{\varDelta t} V[X_{g+1} - x|X_g = x] &\sim \frac{\beta}{N_g/N} X_g\left(1 - X_g\right) \\ &= \frac{\beta}{n_i(t)} x(t)\left(1 - x(t)\right), \end{aligned} \tag{3}$$

where $t = (N-g)\varDelta t$.

In the case of the mean difference of the VAF from generation g to $g+1$,

$$\frac{1}{\varDelta t} E[X_{g+1} - x|X_g = x] = 0. \tag{4}$$

where I used the equation (2).

This process of the variant allele frequency drift is a continuous Markov process because the variant allele frequency in generation $g+1$ is only dependent on that of generation g. If I denote the transition probability density that the random variable changes from y to x during time t as $f(x|y,t)$, the diffusion equation of the continuous time Markov process is expressed as the following Kolmogorov forward equation [22, 23],

$$\frac{\partial}{\partial t} f(x|y,t) = -\frac{\partial}{\partial x}\left(\mu(x) f(x|y,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(V(x) f(x|y,t)\right), \tag{5}$$

where $\mu(x)$ and $V(x)$ represents the first and the second moment of $\varDelta x$ in the infinitesimal time interval $(t, t+\varDelta t)$.

In this case, $f(x|y,t)$ represents the probability density that the variant allele frequency changes from y to x during time t. And from equation (4) and equation (3), $\mu(x) = 0$ and $V(x) = \frac{\beta}{n(t)} x(t)\left(1 - x(t)\right)$. Substituting them into the equation (5) yields the following diffusion equation,

$$\frac{\partial}{\partial t} f(x|y,t) = \frac{1}{2}\frac{\beta}{n(t)}\frac{\partial^2}{\partial x^2}\left(x(1-x) f(x|y,t)\right). \tag{6}$$

Changing the variable from t to τ using the following equation,

$$\tau(t) = \int_0^t \frac{\beta}{n(t')}\, dt',$$

(6) becomes the simpler form,

$$\frac{\partial}{\partial\tau} f(x|y,\tau) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(x(1-x)f(x|y,\tau)\right). \tag{7}$$

If I assume that the solution can be written in the separated form $f(x|y,\tau) = T(\tau; y)X(x; y)$, substituting it into (7) and dividing both sides with $T(\tau; y)X(x; y)$ yields the following equation,

$$\frac{1}{T}\frac{\partial T}{\partial\tau} = \frac{1}{2}\frac{1}{X}\frac{\partial}{\partial x^2}\left(x(1-x)X\right) = -\lambda, \tag{8}$$

where I put the both sides as a constant, λ, because the left and the right hand side have the same value though they are only dependent on τ or x respectively.

From the left hand side of equation (8),

$$T \propto \mathrm{e}^{-\lambda\tau}.$$

On the other hand, the right hand side can be written in the following form,

$$x(1-x)\frac{d^2X}{dx^2} + 2(1-2x)\frac{dX}{dx} - 2(1-\lambda)X = 0. \tag{9}$$

This is a special case of the hypergeometric equation (35). And I denote the solution of this hypergeometric equation as $F(a, b, c; x)$, which is called a hypergeometric function.

Comparing equation (9) with equation (35),

$$c = 2,\ a + b = 3,\ ab = 2(1-\lambda). \tag{10}$$

Solving the simultaneous equation (10),

$$a = \frac{3 + \sqrt{1+8\lambda}}{2},\ b = \frac{3 - \sqrt{1+8\lambda}}{2}. \tag{11}$$

The hypergeometric function has the following relationship,

$$F(a, b; 2; x) = \frac{\Gamma(2)\Gamma(2-a-b)}{\Gamma(2-a)\Gamma(2-b)}F(a, b, -1+a+b, 1-x) + \frac{\Gamma(2)\Gamma(a+b-2)}{\Gamma(a)\Gamma(b)}F(2-a, 2-b, 3-a-b, 1-x).$$

And

$$2 - a \leq 0 \text{ and } b \leq 0\ (a, b \in \mathbf{Z}),$$

where $\mathbf{Z}$ represents the integer set, is required for $F(a, b, 2, x)$ to have the finite value.

Then,

$$a = 1 + i\ (i = 1, 2, \cdots)$$
$$b = 1 - j\ (j = 1, 2, \cdots)$$

Using (11),

$$\lambda = \frac{(i-1)i}{2} = \frac{j(j+1)}{2}$$

Finally, noting that j gives the tighter condition, a, b, and λ are expressed as follows,

$$a = 2 + j\ , b = 1 - j\ (j = 1, 2, \cdots).$$

Thus,

$$X(x; y) \propto F(2 + i, 1 - i, 2; x).$$

Using the following relationship between the hypergeometric fuction and Gegenbauer polynomial $C_{i-1}^{(3/2)}(z)$,

$$C_{i-1}^{(3/2)}(z) = \frac{i(i+1)}{2} F\left(i + 2, 1 - i, 2; \frac{1-z}{2}\right),$$

$$X(x; y) \propto C_{i-1}^{(3/2)}(z)$$

where $z = 1 - 2x$.

Therefore, the overall solution can be written as follows putting the coefficient a_i,

$$f(x|y, t) = \sum_{j=1}^{\infty} a_j C_{j-1}^{(3/2)}(z) \mathrm{e}^{-\frac{j(j+1)}{2}\tau}$$

Under the initial condition $f(x|y, 0) = \delta(x - y) = 2\delta(z - w)$, where $w = 1 - 2y$,

$$2\delta(z - w) = \sum_{j=1}^{\infty} a_j C_{j-1}^{(3/2)}(z).$$

When I multiply both sides with $(1 - z^2)C_{k-1}^{(3/2)}(z)$, and take the integral with respect to z,

$$2\int_{-1}^{1} \delta(z - w)(1 - z^2)C_{k-1}^{(3/2)}(z)dz = \sum_{j=1}^{\infty} a_j \int_{-1}^{1} (1 - z^2)C_{j-1}^{(3/2)}(z)C_{k-1}^{(3/2)}(z)dz.$$

Using the orthogonality of Gegenbauer polynomial, $\int_{-1}^{1}(1-z^2)C_{j-1}^{(3/2)}(z)C_{k-1}^{(3/2)}(z)dz = \frac{2k(k+1)}{2k+1}\delta_{jk}$,

$$a_k = (1 - w^2)\frac{2k+1}{k(k+1)} C_{k-1}^{(3/2)}(w).$$

Therefore, the solution of the diffusion equation (7) is,

$$\begin{aligned} f(x|y, \tau) &= (1 - w^2) \sum_{j=1}^{\infty} \frac{2j+1}{j(j+1)} C_{j-1}^{(3/2)}(w) C_{j-1}^{(3/2)}(z) \mathrm{e}^{-\frac{j(j+1)}{2}\tau} \\ &= 4y(1 - y) \sum_{j=1}^{\infty} \frac{2j+1}{j(j+1)} C_{j-1}^{(3/2)}(1 - 2y) C_{j-1}^{(3/2)}(1 - 2x) \mathrm{e}^{-\frac{j(j+1)}{2}\tau}, \end{aligned} \tag{12}$$

where $z = 1 - 2x$, $w = 1 - 2y$, and $\tau(t) = \int_0^t \frac{\beta}{n(t')}\, dt'$.

However, allelic frequency of a variant allele might be fixed to 1 (fixation) or 0 (loss) during the diffusion process because if all the cells in generation g have that variant allele, all the cells in generation $g + 1$ must have that variant allele. Thus, $f(1|y, \tau)$ and $f(0|y, \tau)$ are probabilities, though $f(x|y, \tau)$ $(0 < x < 1)$ is a probability density.

From the definition of probability, the following relationship holds,

$$\int_0^1 f(x|y, \tau)dx + f(1|y, \tau) + f(0|y, \tau) = 1.$$

If I put $P(x|y, \tau) = -\frac{1}{2}\frac{\partial}{\partial x}(x(1 - x)f(x|y, \tau))$, from (7),

$$\frac{\partial f(x|y, \tau)}{\partial \tau} = -\frac{\partial P(x|y, \tau)}{\partial x}.$$

This is the continuity equation with regard to the variant allele frequency distribution. $\frac{\partial f(x|y,\tau)}{\partial \tau}$ is the rate of net flow of the probability density, and $P(x|y, \tau)$ corresponds to the flux of the variant allele frequency. Because probability flows into $x = 0$ and $x = 1$ at the rate $-P(0|y, \tau)$ and $P(1|y, \tau)$ respectively, and there is no outflow,

$$\frac{\partial f(0|y, \tau)}{\partial \tau} = -P(0|y, \tau), \tag{13}$$

$$\frac{\partial f(1|y, \tau)}{\partial \tau} = P(1|y, \tau). \tag{14}$$

Integrating the differential equation (13) and (14), we can calculate the fixation probability $f(1|y, \tau)$ and loss probability $f(0|y, \tau)$. Noting that $z = 1 - 2x$,

$$\begin{aligned}
P(x|y, \tau) &= \frac{1}{4}\frac{\partial}{\partial z}\left((1 - z^2)f(x|y, \tau)\right) \\
&= \frac{1 - w^2}{4}\frac{\partial}{\partial z}\sum_{j=1}^{\infty} C_{j-1}^{(3/2)}(w)\frac{2j + 1}{j(j + 1)}(1 - z^2)C_{j-1}^{(3/2)}(z)\mathrm{e}^{-\frac{j(j+1)}{2}\tau} \\
&= -\frac{1 - w^2}{4}\sum_{j=1}^{\infty} C_{j-1}^{(3/2)}(w)\frac{\partial}{\partial z}\left(P_{j+1}(z) - P_{j-1}(z)\right)\mathrm{e}^{-\frac{j(j+1)}{2}\tau} \\
&= -\frac{1 - w^2}{4}\sum_{j=1}^{\infty} C_{j-1}^{(3/2)}(w)(2j + 1)P_j(z)\mathrm{e}^{-\frac{j(j+1)}{2}\tau}
\end{aligned}$$

Thus, using equation (41), (42), and $z = 1 - 2x$,

$$\begin{aligned}
P(0|y, \tau) &= -\frac{1 - w^2}{4}\sum_{j=1}^{\infty}(2j + 1)C_{j-1}^{(3/2)}(w)\mathrm{e}^{-\frac{j(j+1)}{2}\tau}, \\
P(1|y, \tau) &= -\frac{1 - w^2}{4}\sum_{j=1}^{\infty}(-1)^j(2j + 1)C_{j-1}^{(3/2)}(w)\mathrm{e}^{-\frac{j(j+1)}{2}\tau}.
\end{aligned}$$

Integrating equation (13),

$$\begin{aligned}
f(0|y,\tau) &= \frac{1-w^2}{2}\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}(w)\left(1-\mathrm{e}^{-\frac{j(j+1)}{2}\tau}\right) \\
&= \frac{1+w}{2} - \frac{1-w^2}{2}\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}(w)\mathrm{e}^{-\frac{j(j+1)}{2}\tau} \\
&= (1-y) - 2y(1-y)\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}(1-2y)\mathrm{e}^{-\frac{j(j+1)}{2}\tau}, \qquad (15)
\end{aligned}$$

where we used equation (44).

On the other hand, integrating equation (14),

$$\begin{aligned}
f(1|y,\tau) &= -\frac{1-w^2}{2}\sum_{j=1}^{\infty}(-1)^j\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}(w)\left(1-\mathrm{e}^{-\frac{j(j+1)}{2}\tau}\right) \\
&= \frac{1-w}{2} + \frac{1-w^2}{2}\sum_{j=1}^{\infty}(-1)^j\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}(w)\mathrm{e}^{-\frac{j(j+1)}{2}\tau} \\
&= y + 2y(1-y)\sum_{j=1}^{\infty}(-1)^j\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}(1-2y)\mathrm{e}^{-\frac{j(j+1)}{2}\tau}, \qquad (16)
\end{aligned}$$

where we used equation (45).

In the case of subtype i,

$$\tau(t) = \int_0^t \frac{\beta}{n_i(t')}\,dt' = \frac{\beta}{\alpha_i n_i}\left(\mathrm{e}^{\alpha_i t}-1\right).$$

Every single nucleotide mutation which occurs within one of subtype i's cell at time t have the allelic frequency $\frac{1}{Nn_i(t)}$. The probability density that the allelic frequency grows from $\frac{1}{Nn_i(t)}$ to x_i during time t is expressed as $f(x_i|\frac{1}{Nn_i(t)},\tau(t'))$.

The probability that the variant allele frequency x_i is observed at time 0 can be calculated integrating $f\left(x_i|\frac{1}{Nn_i(t)},\tau(t)\right)$ weighed by the number of the cell at t from the birth time of subtype i to the observation time,

$$\int_0^{t_i} f\left(x_i|\frac{1}{Nn_i(t)},\tau(t')\right)(Nn_i(t')-1)dt'. \qquad (17)$$

$$\begin{aligned}
& f\left(x_i|\frac{1}{Nn_i(t)},\tau(t)\right)(Nn_i(t)-1) \\
&= 4\frac{1}{Nn_i(t)}\left(1-\frac{1}{Nn_i(t)}\right)\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}\left(1-2\frac{1}{Nn_i(t)}\right)C^{(3/2)}_{j-1}(1-2x_i)\mathrm{e}^{-\frac{j(j+1)}{2}\tau}\,(Nn_i(t)-1) \\
&\sim 2\frac{1}{Nn_i(t)}\sum_{j=1}^{\infty}(2j+1)C^{(3/2)}_{j-1}(1-2x_i)\exp\left(-\frac{j(j+1)}{2}\frac{\beta}{\alpha_i n_i}\left(\mathrm{e}^{\alpha_i t}-1\right)\right)(Nn_i(t)-1) \\
&= 2(1-\mathrm{e}^{\alpha_i(t-t_i)})\sum_{j=1}^{\infty}(2j+1)C^{(3/2)}_{j-1}(1-2x_i)\exp\left(-\gamma_j\left(\mathrm{e}^{\alpha_i t}-1\right)\right),
\end{aligned}$$

where I used $1 - \frac{1}{Nn_i(t)} \sim 1$ and $C^{(3/2)}_{j-1}\left(1 - 2\frac{1}{Nn_i(t)}\right) \sim C^{(3/2)}_{j-1}(1) = \frac{j(j+1)}{2}$ and denoted $\frac{j(j+1)}{2}\frac{\beta}{\alpha_i n_i} = \frac{j(j+1)}{2}\frac{\beta}{n_i}\frac{t_i}{\ln(Nn_i)}$ as γ_j.

Using

$$\int_0^{t_i} \exp(-\gamma_j \mathrm{e}^{\alpha_i t'})dt' = \frac{1}{\alpha_i}\int_{\gamma_j}^{\gamma_j \mathrm{e}^{\alpha_i t_i}} \frac{\mathrm{e}^{-\zeta}}{\zeta}d\zeta = \frac{1}{\alpha_i}\left[E_1(\gamma_j) - E_1(\gamma_j N n_i)\right]$$

$$\int_0^{t_i} \mathrm{e}^{\alpha_i t'}\exp(-\gamma_j \mathrm{e}^{\alpha_i t'})dt' = \frac{1}{\alpha_i\gamma_j}\int_{\gamma_j}^{\gamma_j \mathrm{e}^{\alpha_i t_i}} \mathrm{e}^{-\zeta}d\zeta = \frac{1}{\alpha_i\gamma_j}\left[\mathrm{e}^{-\gamma_j} - \mathrm{e}^{\gamma_j N n_i}\right],$$

(17) can be calculated as follows,

$$\begin{aligned}
&\int_0^{t_i} f\left(x_i | \frac{1}{Nn_i(t)}, \tau(t')\right)(Nn_i(t') - 1)dt' \\
&= 2\sum_{j=1}^{\infty}(2j+1)C^{(3/2)}_{j-1}(1-2x_i)\mathrm{e}^{\gamma_j}\left[\frac{1}{\alpha_i}\left\{E_1(\gamma_j) - E_1(\gamma_j N n_i)\right\} - \frac{\mathrm{e}^{-\alpha_i t_i}}{\alpha_i\gamma_j}\left(\mathrm{e}^{-\gamma_j} - \mathrm{e}^{\gamma_j N n_i}\right)\right] \\
&= -\frac{2\mathrm{e}^{-\alpha_i t_i}}{\alpha_i}\sum_{j=1}^{\infty}(2j+1)C^{(3/2)}_{j-1}(1-2x_i)\frac{1}{\gamma_j} \\
&+ \frac{2}{\alpha_i}\sum_{j=1}^{\infty}(2j+1)\frac{1}{\gamma_j}C^{(3/2)}_{j-1}(1-2x_i)\left[\gamma_j \mathrm{e}^{\gamma_j}\left\{E_1(\gamma_j) - E_1(\gamma_j N n_i)\right\} + \frac{1}{Nn_i}\mathrm{e}^{-\gamma_j(Nn_i - 1)}\right]
\end{aligned} \tag{18}$$

The first term of (18) is,

$$-\frac{2\mathrm{e}^{-\alpha_i t_i}}{\alpha_i}\sum_{j=1}^{\infty}(2j+1)C^{(3/2)}_{j-1}(1-2x_i)\frac{1}{\gamma_j} = -\frac{4}{N\beta}\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}(z_i) = -\frac{4}{N\beta}\frac{1}{1-z_i} = -\frac{2}{N\beta}\frac{1}{x_i} \tag{19}$$

On the other hand, the second term of (18) is,

$$\frac{4n_i}{\beta}\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}(1-2x_i)\left[\gamma_j \mathrm{e}^{\gamma_j}\left\{E_1(\gamma_j) - E_1(\gamma_j N n_i)\right\} + \frac{1}{Nn_i}\mathrm{e}^{-\gamma_j(Nn_i-1)}\right].$$

However, the element of this series does not converge to 0 in the limit $j \to \infty$ because $C^{(3/2)}_{j-1}(1-2x_i) \to \frac{j(j+1)}{2}$ in the limit $x_i \to 0$ and

$$\gamma_j \mathrm{e}^{\gamma_j}\left\{E_1(\gamma_j) - E_1(\gamma_j N n_i)\right\} + \frac{1}{Nn_i}\mathrm{e}^{-\gamma_j(Nn_i-1)} = 1 - \frac{1}{\gamma_j} + O(\gamma_j^{-2}) = 1 - \frac{2}{j(j+1)}\frac{n_i}{\beta}\frac{\ln(Nn_i)}{t_i} + O(j^{-4}). \tag{20}$$

Thus, I must evaluate the sum of the series with regard to the first and second term of (20) using (44) and (40) respectively. As a result, the second term of (18) can be expressed as follows,

$$\begin{aligned}
&\frac{2n_i}{\beta}\left[\frac{1}{x_i} + \frac{4n_i}{\beta}\frac{\ln(Nn_i)}{t_i}\int_0^1 \frac{(1-s)\log s}{(1-2zs+s^2)^{3/2}}ds\right. \\
&\left. + 2\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}C^{(3/2)}_{j-1}(1-2x_i)\left\{\gamma_j \mathrm{e}^{\gamma_j}\left(E_1(\gamma_j) - E_1(\gamma_j N n_i)\right) - 1 + \frac{1}{\gamma_j} + \frac{1}{Nn_i}\mathrm{e}^{-\gamma_j(Nn_i-1)}\right\}\right]
\end{aligned} \tag{21}$$

Finally, (18) can be expressed as the sum of (19) and (21),

$$\begin{aligned}&\int_0^{t_i} f\left(x_i\middle|\frac{1}{Nn_i(t')},\tau(t')\right)(Nn_i(t')-1)dt'\\&=2\frac{n_i}{\beta}\Bigg[(1-\frac{1}{Nn_i})\frac{1}{x_i}+4\frac{n_i}{\beta}\frac{\ln(Nn_i)}{t_i}\int_0^1 ds\frac{(1-s)\log s}{(1-2zs+s^2)^{3/2}}\\&+2\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}C^{(3/2)}(1-2x_i)\left\{\gamma_j\mathrm{e}^{\gamma_j}\left(E_1(\gamma_j)-E_1(\gamma_jNn_i)\right)-1+\frac{1}{\gamma_j}+\frac{1}{Nn_i}\mathrm{e}^{-\gamma_j(Nn_i-1)}\right\}\Bigg]\end{aligned}\tag{22}$$

In the same way, the fixation probability can be calculated as follows using (16),

$$\begin{aligned}&\int_0^{t_i} f\left(1\middle|\frac{1}{Nn_i(t')},\tau(t')\right)(Nn_i(t')-1)dt'\\&=t_i\left(1-\frac{1}{\ln(Nn_i)}\left(1-\frac{1}{Nn_i}\right)\right)\\&+2\frac{n_i}{\beta}\Bigg[-(1-\frac{1}{Nn_i})+\sum_{j=1}^{\infty}\frac{(-1)^j(2j+1)}{j(j+1)}\left\{\gamma_j\mathrm{e}^{\gamma_j}\left(E_1(\gamma_j)-E_1(\gamma_1Nn_i)\right)-1+\frac{1}{Nn_i}\mathrm{e}^{-\gamma_j(Nn_i-1)}\right\}\Bigg].\end{aligned}\tag{23}$$

Noting that all the observed variant NGS reads are derived from SNVs whose allele frequency $x_i > \frac{1}{Nn_i}$, I should normalize (22) and (23) with $\Pr\left\{x_i > \frac{1}{Nn_i}\right\}$, which is calcuated as follows,

$$\begin{aligned}&\Pr\left\{x_i>\frac{1}{Nn_i}\right\}\\&=\int_{\frac{1}{Nn_i}}^1 dx_i\int_0^{t_i} dt' f\left(x_i\middle|\frac{1}{Nn_i(t')},\tau(t')\right)(Nn_i(t')-1)+\int_0^{t_i} dt' f\left(1\middle|\frac{1}{Nn_i(t')},\tau(t')\right)(Nn_i(t')-1)\\&=t_i\left(1-\frac{1}{\ln(Nn_i)}\left(1-\frac{1}{Nn_i}\right)\right)\\&+2\frac{n_i}{\beta}\Bigg[(\ln(Nn_i)-1)(1-\frac{1}{Nn_i})+\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}\{\gamma_j\mathrm{e}^{\gamma_j}\left(E_1(\gamma_j)-E_1(\gamma_1Nn_i)\right)-1+\frac{1}{Nn_i}\mathrm{e}^{-\gamma_j(Nn_i-1)}\}\Bigg],\end{aligned}\tag{24}$$

where $C^{(3/2)}(1-2x_i)$ is integrated using equation (48).

Finally, the variant allele frequency distribution $p(x_i|t_i,n_i)$ is calculated as follows using (22), (23), and (24),

$$p(x_i|t_i,n_i)=\begin{cases}\int_0^{t_i} f\left(x_i\middle|\frac{1}{Nn_i(t')},\tau(t')\right)(Nn_i(t')-1)dt'\Big/\Pr\left\{x_i>\frac{1}{Nn_i}\right\} & \left(\frac{1}{Nn_i}<x_i<1\right)\\ \int_0^{t_i} f\left(1\middle|\frac{1}{Nn_i(t')},\tau(t')\right)(Nn_i(t')-1)dt'\Big/\Pr\left\{x_i>\frac{1}{Nn_i}\right\} & (x_i=1)\,.\end{cases}\tag{25}$$

Above equation holds if the mutation is an isolated mutation of subtype i, i.e. the mutation is not inherited by descendant subtypes. However, in reality, the mutation is inherited to the descendant subtypes if the mutation occured in the lineage of the child subtype's founder cell. In this case, the variant allele frequency x_i can be zero even if the variant NGS reads are observed because $x_{\text{child}(i)} = 1$.

Since the mutation must occur in a single lineage before the child subtype is born ($t = t_{\text{child}(i)}$), The probability that the variant allele frequency x_i is observed at time 0 can be calculated integrating $f\left(x_i|\frac{1}{Nn_i(t)}, \tau(t)\right)$ from the birth time of subtype i to the birth time of the i's child subtype,

$$\begin{aligned}&\int_{t_{\text{child}(i)}}^{t_i} f\left(x_i\middle|\frac{1}{Nn_i(t')}, \tau(t')\right) dt' \\ &= \frac{4}{N\beta}\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}C^{(3/2)}(1-2x_i)\mathrm{e}^{\gamma_j}\left\{\exp\left(-\gamma_j\,(Nn_i)^{\frac{t_{\text{child}(i)}}{t_i}}\right) - \exp(-\gamma_j Nn_i)\right\}. \qquad (26)\end{aligned}$$

The fixation probability is,

$$\begin{aligned}&\int_{t_{\text{child}(i)}}^{t_i} f\left(1\middle|\frac{1}{Nn_i(t')}, \tau(t')\right) dt' \\ &= \frac{t_i}{\ln(Nn_i)}\left(1-(Nn_i)^{\frac{t_{\text{child}(i)}}{t_i}-1}\right) + \frac{2}{N\beta}\sum_{j=1}^{\infty}\frac{(-1)^{-j}(2j+1)}{j(j+1)}\mathrm{e}^{\gamma_j}\left\{\exp\left(-\gamma_j\,(Nn_i)^{\frac{t_{\text{child}(i)}}{t_i}}\right) - \exp(-\gamma_j Nn_i)\right\}. \\ &\qquad (27)\end{aligned}$$

And the loss probability is,

$$\begin{aligned}&\int_{t_{\text{child}(i)}}^{t_i} f\left(0\middle|\frac{1}{Nn_i(t')}, \tau(t')\right) dt' \\ &= t_i - t_{\text{child}(i)} - \frac{t_i}{\ln(Nn_i)}\left(1-(Nn_i)^{\frac{t_{\text{child}(i)}}{t_i}-1}\right) - \frac{2}{N\beta}\sum_{j=1}^{\infty}\frac{2j+1}{j(j+1)}\mathrm{e}^{\gamma_j}\left\{\exp\left(-\gamma_j\,(Nn_i)^{\frac{t_{\text{child}(i)}}{t_i}}\right) - \exp(-\gamma_j Nn_i)\right\}. \\ &\qquad (28)\end{aligned}$$

Normalization factor is calculated using equation (26), (27), and (28),

$$\begin{aligned}&\int_0^1 dx_i \int_{t_{\text{child}(i)}}^{t_i} dt' f\left(x_i\middle|\frac{1}{Nn_i(t')}, \tau(t')\right) + \int_{t_{\text{child}(i)}}^{t_i} dt f\left(1\middle|\frac{1}{Nn_i(t')}, \tau(t')\right) + \int_{t_{\text{child}(i)}}^{t_i} dt f\left(0\middle|\frac{1}{Nn_i(t')}, \tau(t')\right) \\ &= t_i - t_{\text{child}(i)} \qquad (29)\end{aligned}$$

Finally, the variant allele frequency distribution $p(x_i|t_i, t_{\text{child}(i)}, n_i)$ is calculated as follows using (26), (27), (28), and (29),

$$p(x_i|t_i, t_{\text{child}(i)}, n_i) = \begin{cases} \int_0^{t_i} f\left(x_i\middle|\frac{1}{Nn_i(t')}, \tau(t')\right) dt' \Big/ (t_i - t_{\text{child}(i)}) & (0 < x_i < 1) \\ \int_0^{t_i} f\left(1\middle|\frac{1}{Nn_i(t')}, \tau(t')\right) dt' \Big/ (t_i - t_{\text{child}(i)}) & (x_i = 1) \\ \int_0^{t_i} f\left(0\middle|\frac{1}{Nn_i(t')}, \tau(t')\right) dt' \Big/ (t_i - t_{\text{child}(i)}) & (x_i = 0)\,. \end{cases} \qquad (30)$$

2.2 Simulation of the NGS reads

In this section, I will explain how NGS reads are simulated in this study.

First, I discretized the time to H time points, thus time difference $\varDelta t = \frac{1}{H}$. Then I put the length of the captured region as B. The mutation rate of each subtype is denoted as r_i. The algorithm of NGS read count simulation is written as Algorithm 1 using the the variant allele frequency transition probability (12), (16), and (15), where mark_inherited($child(j)$) denotes the recursive subroutine to set $x = 1$ for all the subtype j's descendants. How each passenger mutation is inherited to descendant subtypes is explained in Figure 2. I generated the reads using $H = 1000, N = 10^6, r_i = 1.67 \times 10^{-7}$ with varying B.

```
Reads = {}
for $h = H - 1, \cdots, 1$ do
    $t = h\varDelta t$
    for $Subtype\ i \in \{i | t_i > t\}$ do
        $\lambda_i = Nn_i(t) \times B \times r_i \varDelta t$
        $s_i \sim Poisson(s_i | \lambda_i)$
        for $SNVs\ \ k_i \in \{1, \cdots, s_i\}$ do
            $x_i \sim f(x_i | 1/Nn_i(t), t)$
            for $Subtype\ \ j \in child(i)$ do
                if $t_j < t$ then
                    $x_j \sim Bernoulli(x_j | 1/Nn_i(t))$
                    if $x_j = 1$ then
                        mark_inherited($child(j)$)
                    end
                end
            end
            $\mu_{k_i} = \sum_{i=1}^{I} n_i x_i / 2$
            $m_{k_i} \sim Binom(m_{k_i} | M, \mu_{k_i})$
            Reads.append(($m_{k_i}, M$))
        end
    end
end
return Reads
```

Algorithm 1: NGS read count simulation

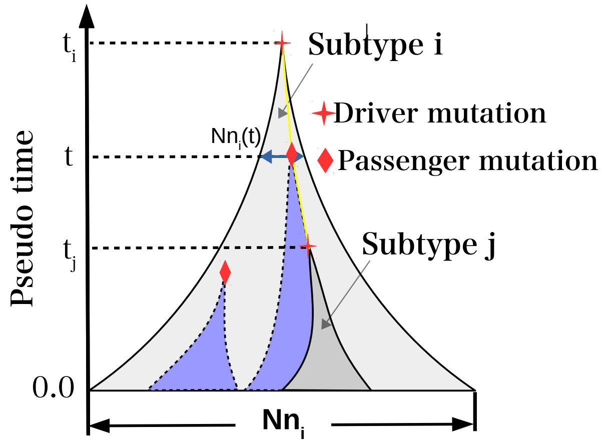


Figure 2: Schematic diagram to show how the passenger mutation which originated in subtype i is inherited to its descendant subtypes. The yellow line represents the lineage of the subtype j's founder cell. If the passenger mutation occurs on the lineage of the subtype j's founder cell, that mutation is inherited to subtype j and all the descendants of subtype j. Thus, all the subtype j's cells and some fraction of subtype i's cells harbor the passenger mutation at sequencing time. However, if the passenger mutation occurs in other lineages, that mutation is not inherited to its desendants.

2.3 Birth time and abundance ratio estimation using NGS reads

In this section, I define the problem of inferring the abundance ratio n_i and the birth time t_i of each subtype from NGS reads of the bulk tumor sample. When the number of total reads M_k and the number of variant reads m_k at the each SNV locus k, $(m, M) = \{(m_1, M_1), \cdots, (m_K, M_K)\}$ is given, the abundance ratio n_i and the birth time t_i of each subtype can be estimated via the maximum likelihood problem.

2.3.1 Estimation modeling using binomial distribution

Given the birth time $t = (t_1, \cdots, t_I)$ and the abundance ratio $n = (n_1, \cdots, n_I)$, the probability that we get the read count (m, M) as a result of NGS sequencing and the variant calling is,

$$p(m|M,n,t) = \prod_{k=1}^{K} \sum_{i_k=1}^{I} p(i_k) \sum_{h_k=0}^{H_k-1} p(h_k|i_k) \sum_{x_{i_k}} p(x_{i_k}|i_k, h_k, n_{i_k}, t_{i_k}, t_{\mathrm{child}(i_k)}) p(m_k|M_k, n, x) \tag{31}$$

where $p(m_k|M_k, n, x)$ represents the probability that m_k variant reads are observed among M_k total reads, which is expressed as the following binomial distribution.

$$p(m_k|M_k, n, x) = Binom(m_k|M_k, \mu_k) = \binom{m_k}{M_k} \mu_k^{m_k} (1 - \mu_k)^{M_k - m_k} .$$

$\mu_k = \sum_{i_k=1}^{I} n_i x_i / 2$ represents the ratio of variant copies among all the copies exist in the tumor sample, assuming that all the tumor subtype are diploids and all the variants are heterozygous.

And $p(i_k)$ represents the probability that the SNV k originated in subtype i_k. $p(h_k|i_k)$ represents the probability that SNV k is inherited to the child subtypes in the pattern $h_k \in \{(0, \cdots, 0), (0, \cdots, 1), \cdots, (1, \cdots, 1)\}$, where the j th element of h_k represents whether the j th child of subtype i inherits the numation or not. H_{i_k} denotes the number of possible inheritance patterns, which is equal to $2^{\kappa_{i_k}}$ where κ_{i_k} = (number of subtype i_k's children) because we can choose wheter each i_k's child inherits the mutation or not. $p(x_{i_k}|i_k, h_k, n_{i_k}, t_{i_k}, t_{\mathrm{child}(i_k)})$ represents the probability that the variant allele frequency of subtype i_k is x_{i_k}, which corresponds to equation (25) or (30) depending on whether SNV k is an isolated mutation (i.e. $h_k = (0, \cdots, 0)$) or not. Though the fraction of the mutated cell x_i takes any value between 0 and 1 in reality, however, I assumed that x_i takes the discrete value for the estimation efficiency. That is, $x_{i_k} \in \{0.1, 0.2, \cdots, 1.0\}$ if $h_k = (0, \cdots, 0)$ and $x_{i_k} \in \{0.0, 0.1, 0.2, \cdots, 1.0\}$ otherwise. x_{i_k} starts from 0.1 in the case of $h_k = (0, \cdots, 0)$ because at least one subtype must have a positive variant allele frequency for the mutated reads to be observed.

Then, the maximum likelihood extimation of the abundance ratio n_i and the birth time t_i of each subtype can be expressed as follows,

$$(t, n) = \arg\max_{t,n} \ln p(m|M, n, t).$$

2.3.2 Estimation modeling using NGS read emission probability

Given the birth time $t = (t_1, \cdots, t_I)$ and the abundance ratio $n = (n_1, \cdots, n_I)$, the probability that m_k variant reads and $M_k - m_k$ normal reads are observed as a result of NGS sequencing is,

$$p(D|t, n) = \prod_{k=1}^{K} \Bigg[\sum_{i_k=1}^{I} p(i_k) \sum_{h_k=0}^{H_k-1} p(h_k|i_k) \sum_{x_{i_k}} p(x_{i_k}|i_k, h_k, n_{i_k}, t_{i_k}, t_{\mathrm{child}(i_k)}) \\ \times \prod_{l=0}^{m_k-1} \left(\sum_{q_{kl}=1}^{I} n_{q_{kl}} \frac{x_{q_{kl}}}{2} \right) \prod_{l=m_k}^{M_k-1} \left(\sum_{q_{kl}=1}^{I} n_{q_{kl}} \left(1 - \frac{x_{q_{kl}}}{2} \right) \right) \Bigg] \tag{32}$$

where the l th read at the SNV locus k is derived from the subtype q_{kl}'s copy and $D = \{D_1, \cdots D_K\} = \{(m_1, M_1), \cdots, (m_K, M_K)\}$.

The Expectation-Maximization (EM) algorithm to derive the maximum likelihood estimate of t and n can be formulated as follows.

For each SVN locus k, i_k, h_k, x_{i_k}, and $q_{k0:(M_k-1)} = (q_{k0}, \cdots, q_{k(M_k-1)})$ are assumed to be hidden states because they cannot be observed. Then the hidden indicator variable $z^{(k)}_{i_k, h_k, x_{i_k}, q_{k0:(M_k-1)}}$ can be defined as,

$$z^{(k)}_{i_k, h_k, x_{i_k}, q_{k0:(M_k-1)}} \in \{0, 1\}, \ \sum_{i_k=1}^{I} \sum_{h_k=0}^{H_k-1} \sum_{x_{i_k}} \sum_{q_{k0}=1}^{I} \cdots \sum_{q_{k(M_k-1)}=1}^{I} z^{(k)}_{i_k, h_k, x_{i_k}, q_{k0:(M_k-1)}} = 1.$$

Thus the joint probability distribution of observables (m_k, M_k) and hidden variables $z^{(k)}_{i_k, h_k, x_{i_k}, q_{k0:(M_k-1)}}$

is,

$$p\left(D_k, z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\middle|t,n\right) = \prod_{i_k=1}^{I}\prod_{h_k=0}^{H_k-1}\prod_{x_{i_k}}\prod_{q_{k0:(M_k-1)}}\left\{p(i_k)p(h_k|i_k)p(x_{i_k}|i_k,h_k,n_{i_k},t_{i_k},t_{\text{child}(i_k)})\right.$$
$$\left.\times\prod_{l=0}^{m_k-1}\left(n_{q_{kl}}\frac{x_{q_{kl}}}{2}\right)\prod_{l=m_k}^{M_k-1}\left(n_{q_{kl}}\left(1-\frac{x_{q_{kl}}}{2}\right)\right)\right\}^{z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}}.$$

And the posterior distribution of the hidden variables $z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}$ given observables (m_k, M_k) can be calculated using Bayes theorem,

$$p\left(z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\middle|D_k,t,n\right) = \frac{p\left(D_k, z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\middle|n,t\right)}{\sum_{z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}} p\left(D_k, z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\middle|n,t\right)}.$$

The responsibility that the hidden state $z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}$ takes for explaining the observation (m_k, M_k) is,

$$\begin{aligned}
&\gamma\left(z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\right)\\
&= E_{z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\middle|D_k,t^{\text{old}},n^{\text{old}}}\left[z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\right]\\
&= \frac{p(i_k)p(h_k|i_k)p(x_{i_k}|i_k,h_k,n^{\text{old}}_{i_k},t^{\text{old}}_{i_k},t^{\text{old}}_{\text{child}(i_k)})\prod_{l=0}^{m_k-1}\left(n^{\text{old}}_{q_{kl}}\frac{x_{q_{kl}}}{2}\right)\prod_{l=m_k}^{M_k-1}\left(n^{\text{old}}_{q_{kl}}\left(1-\frac{x_{q_{kl}}}{2}\right)\right)}{\sum_{i_k=1}^{I}p(i_k)\sum_{h_k=0}^{H_k-1}p(h_k|i_k)\sum_{x_{i_k}}p(x_{i_k}|i_k,h_k,n^{\text{old}}_{i_k},t^{\text{old}}_{i_k},t^{\text{old}}_{\text{child}(i_k)})\prod_{l=0}^{m_k-1}\left(\sum_{q_{kl}=1}^{I}n^{\text{old}}_{q_{kl}}\frac{x_{q_{kl}}}{2}\right)\prod_{l=m_k}^{M_k-1}\left(\sum_{q_{kl}=1}^{I}n^{\text{old}}_{q_{kl}}\left(1-\frac{x_{q_{kl}}}{2}\right)\right)}.
\end{aligned}$$

And the Q-function is defined as follows,

$$\begin{aligned}
&Q(t,n;t^{\text{old}},n^{\text{old}})\\
&= \sum_{k=1}^{K}E_{z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\middle|D_k,t^{\text{old}},n^{\text{old}}}\left[\ln p\left(D_k, z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\middle|t,n\right)\right]\\
&= \sum_{k=1}^{K}\sum_{i_k=1}^{I}\sum_{h_k=0}^{H_k-1}\sum_{x_{i_k}}\sum_{q_{k0:(M_k-1)}}\gamma(z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}})\Bigg[\ln p(i_k)+\ln p(h_k|i_k)+\ln p(x_{i_k}|i_k,h_k,n_{i_k},t_{i_k},t_{\text{child}(i_k)})\\
&+\sum_{l=0}^{M_k-1}\ln n_{q_{kl}}+\sum_{l=0}^{m_k-1}\ln\left(\frac{x_{i_{kl}}}{2}\right)+\sum_{l=m_k}^{M_k-1}\ln\left(1-\frac{x_{i_{kl}}}{2}\right)\Bigg]
\end{aligned}$$

The procedure of EM algorithm for estimating the birth time t and abundance ratio n is described as Algorithm 2.

```
1. Select initial parameters $(t^{(0)}, n^{(0)})$
2. for $iter = 1, 2, \cdots$ do
    E step: Calcuate responsibility $\gamma\left(z^{(k)}_{i_k,h_k,x_{i_k},q_{k0:(M_k-1)}}\right)$ based on $(t^{(iter-1)}, n^{(iter-1)})$
    M step: $(t^{(iter)}, n^{(iter)}) = \arg\min_{(t,n)} Q(t, n; t^{(iter-1)}, n^{(iter-1)})$
    if $(t^{(iter)}, n^{(iter)})$ converge then
        break
    end
end
```

Algorithm 2: EM algorithm for estimating the birth time *t* and abundance ratio *n*

Minimization of $Q(t, n; t^{(iter-1)}, n^{(iter-1)})$ in each M step is conducted using the L-BFGS-B gradient descent algorithm [24].

3 Results and Discussion

3.1 Birth time estimation using linear regression

First, I conducted the double check of the neutral tumor evolution which is identified in the previous work [21]. In their study, "if neutral evolution theory applies to passenger mutation, the number of passenger mutations per allelic frequency $M(f)$ follows $1/f$ power law,

$$M(f) = \mu_e \left(\frac{1}{f} - \frac{1}{f_{max}} \right),$$

where μ_e denotes the mutation rate per effective cell division. And they set $R^2 \leq 0.98$ as the stringent criteria to decide whether each tumor sample obeys neutral evolution or not." If we consider that we cannot estimate the mutation rate and the birth time of a subtype at the same time as long as only a single time point data is available because there is a trade-off between the mutation rate and the birth time, the inverse of μ_e has a information about the birth time of the subtype. Thus, "if a tumor sample obeys neutral evolution, we can estimate the birth time of the tumor sample using linear regression", which is conducted as Figure 3, 4.

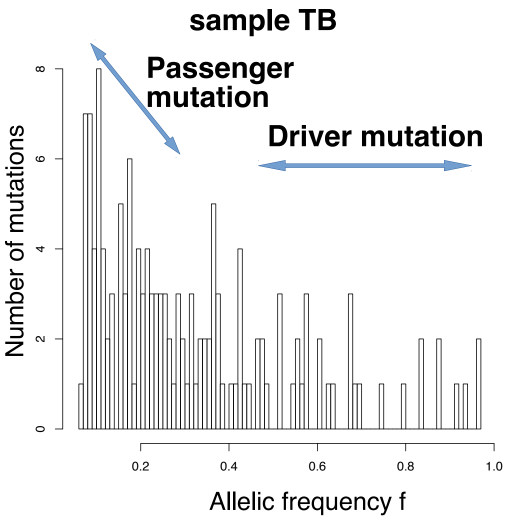


Figure 3: Variant allelic frequency distribution of the colorectal tumor sample TB [21]. Mutations with higher frequency and lower frequency each correspond to driver mutations and passenger mutations respectively.

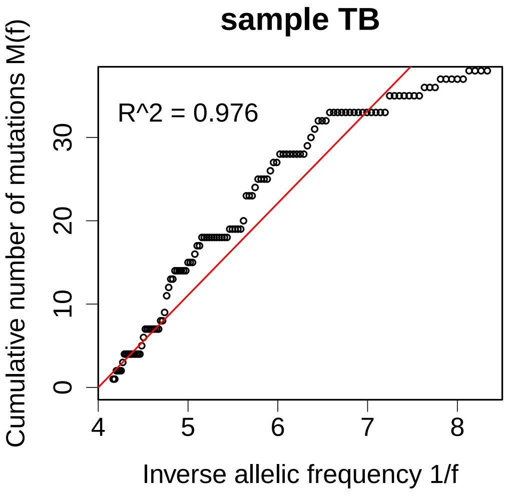


Figure 4: The cumulative number of mutations $M(f)$ plotted against the inverse allelic frequency $1/f$ using the sample TB [21]. The red line represents the result of the linear regression. Though $R^2 = 0.976 < 0.98$ deviates from the stringent criteria [21], this tumor almost obeys the neutral evolution.

However, "if there are multiple subtypes, $M(f)$ generally does not follow $1/f$ power law as the previous work showed using simulated data" [21]. We conducted double check of their result using our simulated reads using Algorithm 1. As a result, our simulation also showed deviation from the $1/f$ power law Figure 5, 6.

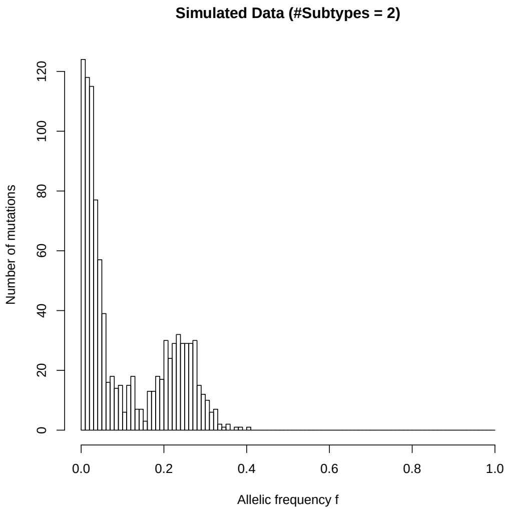


Figure 5: Variant allelic frequency distribution of the simulated reads ($t_1 = 0.5, t_2 = 0.15, n_1 = 0.5, n_2 = 0.5, N = 10^6, \beta = 0.5, \#\text{SNVs} = 1000$).

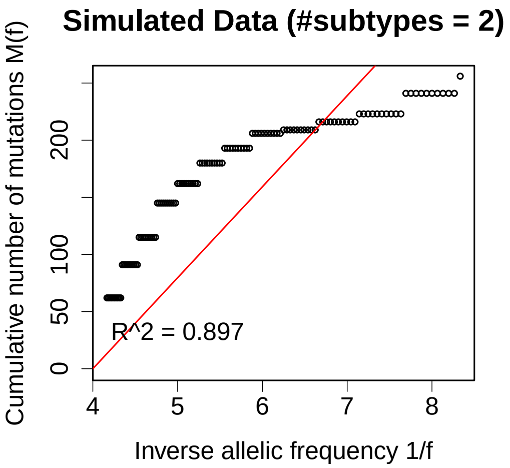


Figure 6: The cumulative number of mutations $M(f)$ plotted against the inverse allelic frequency $1/f$ using the simulated reads ($t_1 = 0.5, t_2 = 0.15, n_1 = 0.5, n_2 = 0.5, \#\text{SNVs} = 1000$). The red line represents the result of the linear regression. The tumor which consists of multiple subtypes cause deviation from the $1/f$ power law though each subtype follows neutral evolution.

Therefore, if a tumor sample consists of multiple subtypes, we cannot estimate the birth time of each subtype using linear regression. I resolved this problem using probabilistic modeling mentioned in the method section.

Moreover, there was a case that the tumor consists of two subtypes though the $M(f)$ and $1/f$ have the linear relationship in our simulated data (Figure 7, 8).

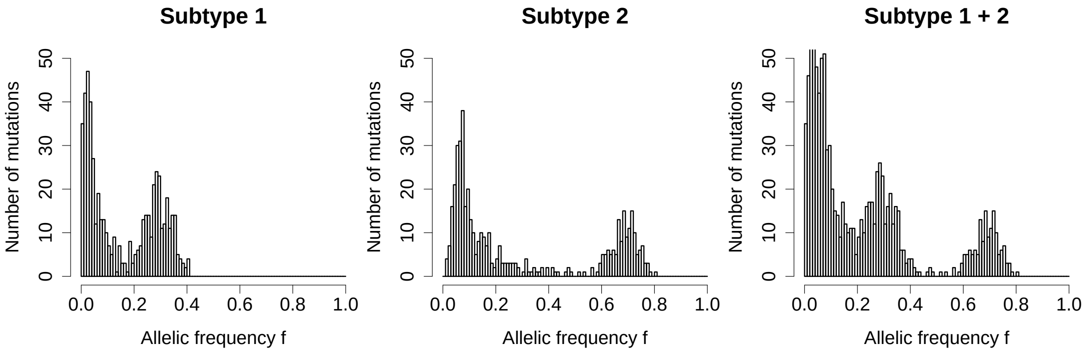


Figure 7: The VAF distribution of the sequencing data which is generated by our probabilistic modeling. NGS read counts are simulated using the birth time $t_1 = 0.5, t_2 = 0.2$, and abundance ratio $n_1 = 0.3, n_2 = 0.7$. Higher allelic frequency peaks seen in subtype 1 and 2 correspond to the fixed variant alleles. Only the VAF distribution of subtype 1+2 can be observed in the bulk sequencing.

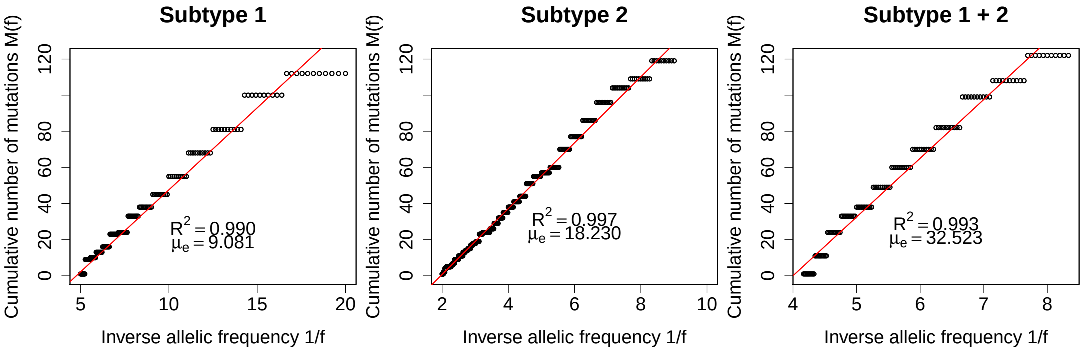


Figure 8: The cumulative number of mutations $M(f)$ plotted against the inverse allelic frequency $1/f$ using the sequencing data which is generated by our probabilistic modeling (birth time $t_1 = 0.5, t_2 = 0.2$, and abundance ratio $n_1 = 0.3, n_2 = 0.7$). If we exclude fixation region which can be seen in Figure 7, subtype 1 and 2 obeys neutral evolution. (I plotetd $M(f)$ against $0.05 < f < 0.2$ for subtype 1, and $0.1 < f < 0.5$ for subtype 2.) This is a concordant result with the previous study [21]. Subtype 1 + 2 also shows linear relationsip between $M(f)$ and $1/f$ with $\mu_e = 32.523$. However, in reality, this is a mixture of two subtypes having different mutation rates per effective cell division ($\mu_e = 9.081$ and $\mu_e = 18.230$). This suggests that the linear regression can cause wrong subtype structure estimation.

When we consider the SNVs originated in subtype 1 and those originated in subtype 2 separately, each subtypes follows the neutral evolution. This shows that our probabilistic modeling of the VAF distribution is concordant with that of the previous work [21] if the fixation and loss region in the VAF disrtibution is excluded. In the previous work, the effect of the genetic drift is not considered, thus the fixation and loss probability is not considered. Thus our Wright-Fisher modeling includes the modeling of the previous study.

However when we consider these SNVs altogether, the estimate of μ_e differed from that of each subtype. This suggests that the linear regression can cause wrong subtype structure and birth time estimation.

3.2 Relationship between the birth time and the variant allele frequency distribution

VAF distribution of subtype i, $p(x_i|t_i, n_i)$ can be calculated given the birth time t_i and the abundance ratio n_i using equation (25) mentioned in the method section. To investigate the time (t_i) dependency and the drift strength (β_i) dependency of the VAF distribution, I showed the violin plots against these variables (Figure 9, 10). From these plots, we can see that the VAF is more likely to be fixed ($x_i = 1$) in the tumor cell population if the subtype arose earlier. Also, the VAF is more likely to be fixed ($x_i = 1$) in the tumor cell population if the effect of the genetic drift is stronger.

Using this relationship conversely, we can estimate the birth time of each subtype from the observed VAF distribution. However, the observable VAF frequency distribution is a mixture of various subtypes as the figure “Subtype 1 + 2” shown in Figure 7. Thus, we must conduct the subtype decomposition using mixture modeling described in the method section.

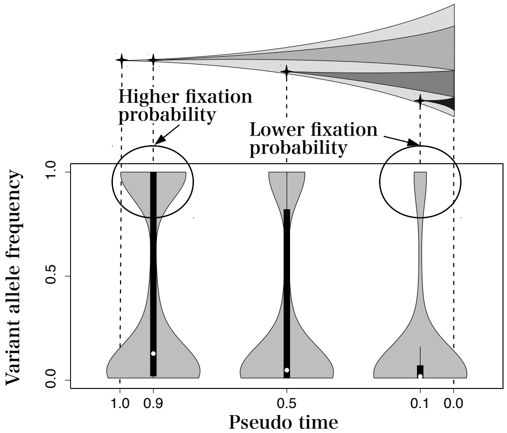


Figure 9: VAF distribution $p(x_i|t_i, n_i)$ plotted against pseudo time t_i ($n_i = 0.1, N = 10^6, \beta_i = 0.5$). We can see that the earlier the subtype arose, the higher the VAF fixation probability is.

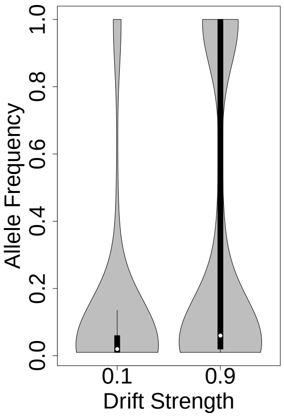


Figure 10: VAF distribution $p(x_i|t_i, n_i)$ plotted against the drift strength β_i ($t_i = n_i = 0.1, N = 10^6$). We can see that the earlier the subtype arose, the higher the VAF fixation probability is.

3.3 Estimation of the birth time and the abundance ratio of each subtype using binomial distribution

Using the parameter estimation modeling represented as equation (31), I conducted the estimation of the birth time t_i and the abundance ratio n_i of each subtype.

First, I started with solving more easier problem than that presented in the method section (equation (31)). I assumed that we know the subtype i_k in which each SNV k is originated as well as the number of variant and total reads (m, M). In this case, the probability that we get the read count along with additional information $(m, M, i, h) = \{(m_1, M_1, i_1, h_1), \cdots, (m_K, M_K, i_K, h_K)\}$ as a result of NGS sequencing and the variant calling is,

$$p(m_{1:K}|M_{1:K}, i_{1:K}, n_{1:I}, t_{1:I}) = \prod_{k=1}^{K} \sum_{h_k=0}^{H_k-1} \frac{1}{H_{i_k}} \sum_{x_{i_k}} p(x_{i_k}|i_k, h_k, n_{i_k}, t_{i_k}, t_{\mathrm{child}(i_k)}) p(m_k|M_k, n, x),$$

where I put $p(h_k|i_k) = \frac{1}{H_{i_k}}$ in the equation (31) to marginalize the unknown inheritance patterns with equal probabilities. Using this marginal probability, I estimated the birth time t and abundance ratio n using Fletcher-Reeves conjugate gradient algorithm [25] according to the following maximum likelihood estimation,

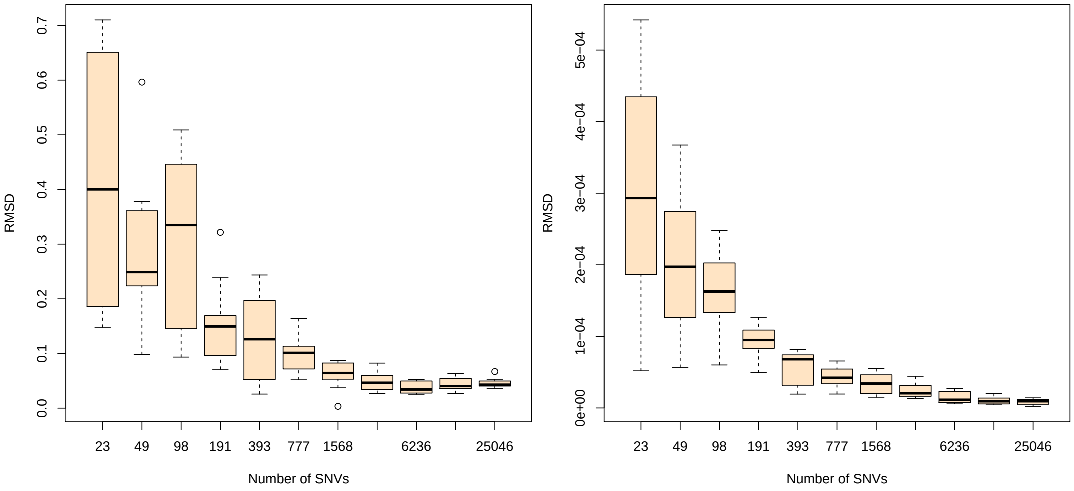
$$(t_{1:I}, n_{1:I}) = \underset{t_{1:I}, n_{1:I}}{\arg\max} \ln p(m_{1:K}|M_{1:K}, i_{1:K}, n_{1:I}, t_{1:I})$$

Simulated data was generated using Algorithm 1 represented in the method section. For each number of SNVs, sequencing data was generated 10 times independently. And for each simulated data, maximum likelihood estimation was conducted 10 times using randomly selected initial parameters t, n. The parameters with the least Root-Mean-Square deviation

(RMSD) was selected as the best estimate for that simulated data. The birth time of the normal cell is defined as $t_0 = 1$, and the abundance ratio $n_0 = 0.1$ is given.

Table 1: True parameters which is used to generate sequencing data (#Subtypes=4)

Population type	Tumor subtype 1	Tumor subtype 2	Tumor subtype 3	Tumor subtype 4
Birth time	9.000460e-01	8.100828e-01	6.481035e-01	7.291118e-01
Abundance ratio	1.000000e-01	2.000000e-01	4.000000e-01	2.000000e-01



(11.1) Accuracy of the birth time (t) estimation

(11.2) Accuracy of the abundance ratio (n) estimation

Figure 11: Accuracy of the birth time (t) and abundance ratio (n) estimation against the number of SNVs. Outliers are detected by means of deviation from the $1.5 \times IQR$, where IQR is the interquartile range. Whiskers show the maximum and minimum estimates except for the outliers. We can see that if there are sufficient number of SNVs, we can accurately estimate the birth time and the abundance ratio as long as the subtype in which each SNV is originated is given.

The results of the parameter estimation shows that we can infer the birth time and abundance ratio accurately if there are larger number of SNVs (Figure 11). In reality, however, we cannot tell the subtype in which each SNV is originated only from the sequencing reads. To move on to the next step, I investigated whether we can remove this unrealistic assumption by marginalizing the subtypes with equal probabilities, which is formulated as follows,

$$p(m_{1:K}|M_{1:K}, n_{1:I}, t_{1:I}) = \prod_{k=1}^{K} \sum_{i_k=1}^{I} \frac{1}{I} \sum_{h_k=0}^{H_k-1} \frac{1}{H_{i_k}} \sum_{x_{i_k}} p(x_{i_k}|i_k, h_k, n_{i_k}, t_{i_k}, t_{\mathrm{child}(i_k)}) p(m_k|M_k, n, x) \tag{33}$$

$$(t_{1:I}, n_{1:I}) = \underset{t_{1:I}, n_{1:I}}{\arg\max} \ln p(m_{1:K}|M_{1:K}, n_{1:I}, t_{1:I}),$$

where I put $p(i_k) = \frac{1}{I}$ and $p(h_k|i_k) = \frac{1}{H_{i_k}}$ in the equation (31). True parameters of t and n and the simulated reads are the same as that of mentioned above (Table 1). As a result of parameter estimation using equation (34), I could not estimate the true birth time and abundance ratio even if there are thousands of SNVs (Supplementary Figure 18). This is because gradient descent estimates fall into the optima which are different from the true parameters.

Also, I tried to estimate the birth time and abundance ratio by adding only the maximum probability component with regard to the subtype in which each SNV is originated. The formulation is given as follows.

$$p(m_{1:K}|M_{1:K}, n_{1:I}, t_{1:I}) = \prod_{k=1}^{K} \max_{i_k=1}^{I} \left[\sum_{h_k=0}^{H_k-1} \frac{1}{H_{i_k}} \sum_{x_{i_k}} p(x_{i_k}|i_k, h_k, n_{i_k}, t_{i_k}, t_{\mathrm{child}(i_k)}) p(m_k|M_k, n, x) \right], \quad (34)$$

$$(t_{1:I}, n_{1:I}) = \underset{t_{1:I}, n_{1:I}}{\arg\max} \ln p(m_{1:K}|M_{1:K}, n_{1:I}, t_{1:I}),$$

In this case too, the gradient descent estimates fall into the optima which are different from the true parameters, thus I could not infer the true birth time and abundance ratio even if there are thousands of SNVs (Supplementary Figure 19). These problems was essentially caused by the miss subtype-labeling of each SNV, that is, the subtype which gives the maximum probabilistic component does not match the true subtype. Though we cannot expect the perfect subtype-labeling because we only have a single data with regard to each SNV, systematic labeling error cause the estimation failure. In the next subsection, I conducted the estimation of the birth time and abundance ratio using the generative model (equation (32)), which is concordant with the read simulation in this study, thus there is a guarantee that the maximum likelihood estimates converge to the true parameters given sufficient number of data.

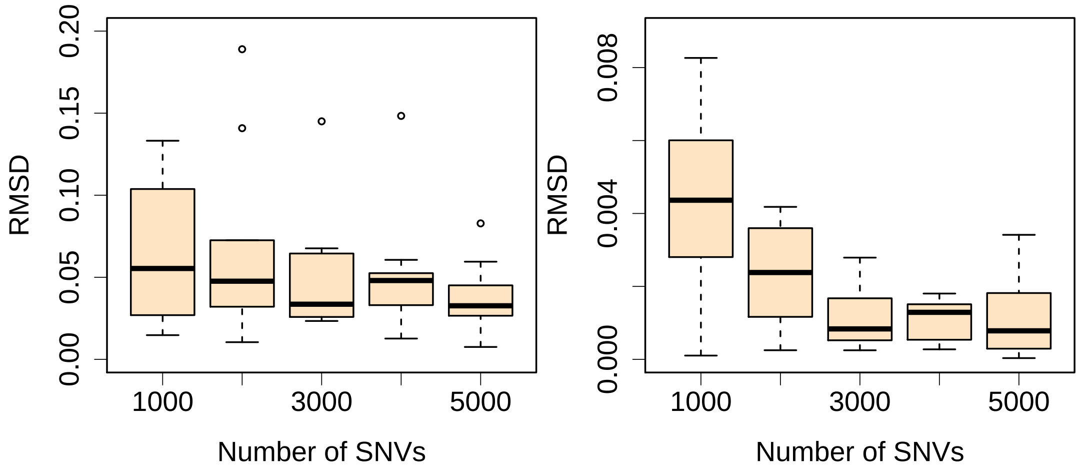
3.4 Estimation of the birth time and the abundance ratio of each subtype using NGS read emission probability

Based on the EM algorithm derived in the method section (Algorithm 2), I conducted the maximum likelihood estimation of the birth time and abundance ratio of each subtype. Simulated reads were generated with the true parameter shown in Table 2. Normal cell population birth time is defined as $t_0 = 1$, and the abundance ratio $n_0 = 0$ is given.

Table 2: True parameters which is used to generate sequencing data (#Subtypes=2)

Population type	Tumor subtype 1	Tumor subtype 2
Birth time	5.000000e-01	2.000000e-01
Abundance ratio	3.000000e-01	7.000000e-01

The result of the birth time and abundance ratio estimation varying the number of SNVs from 1000 to 5000 while the sequencing depth is fixed to 100 is shown in (Figure 12).



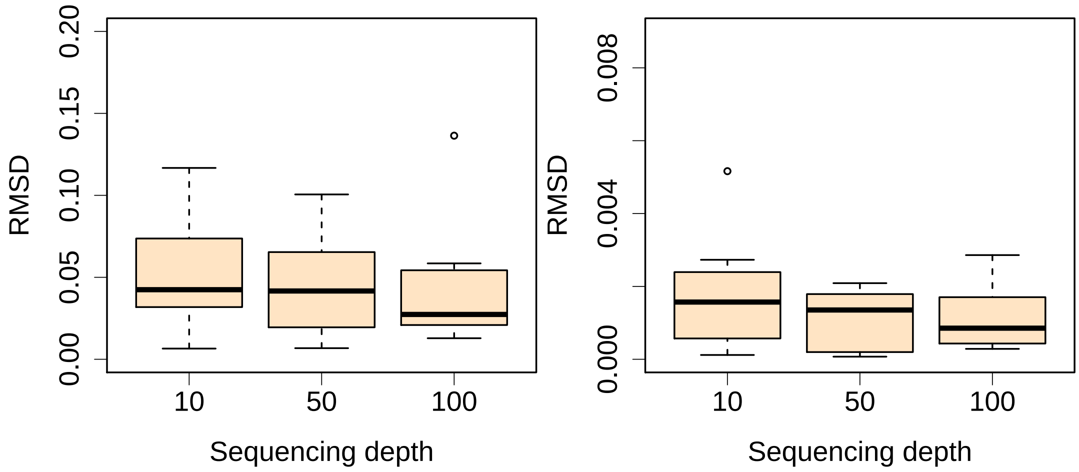
(12.1) Accuracy of the birth time (*t*) estimation

(12.2) Accuracy of the abundance ratio (*n*) estimation

Figure 12: Accuracy of the birth time (*t*) and abundance ratio (*n*) estimation against the number of SNVs. Outliers are detected by means of deviation from the $1.5 \times IQR$, where IQR is the interquartile range. Whiskers show the maximum and minimum estimates except for the outliers. We can see that if there are sufficient number of SNVs, we can accurately estimate the birth time and abundance ratio even if we do not know the subtype in which each SNV is originated.

From this result, we can inferr the birth time and abundance ratio accurately with larger number of SNVs. Though the birth time estimation requires greater number of SNVs, smaller number of SNVs is sufficient for the abundance ratio estimation. I suppose this is because abundance ratio n appears explicitly in the equation (32) while the birth time t affect the likelihood through VAF distribution.

Finally, I estimated the birth time and abundance ratio varying the sequencing depth from 10 to 100 while the number of SNVs is fixed to 3000 (Figure 13).



(13.1) Accuracy of the birth time (t) estimation (13.2) Accuracy of the abundance ratio (n) estimation

Figure 13: Accuracy of the birth time (t) and abundance ratio (n) estimation against the sequencing depth. Outliers are detected by means of deviation from the $1.5 \times IQR$, where IQR is the interquartile range. Whiskers show the maximum and minimum estimates except for the outliers. The birth time and the abundance ratio can be estimated more acculately with the higher sequencing depth.

The birth time and abundance ratio were estimated more acculately with the higher sequencing depth.

4 Conclusions

For the precise birth time parameter estimation, larger number of SNVs were required. On the other hand, abundance ratio could be estimated using smaller number of SNVs. From the inferred abundance ratio and birth time of each subtype, we can estimate the growth rate of each subtype assuming exponential tumor growth using equation (1).

When we consider the applicable target of our birth time estimation method, circulating tumor DNA (ctDNA) is getting much more attention in recent years. Clinical application of ctDNA is awaited because it is more reliable tool in the cancer diagnosis than the conventional tumor markers. Moreover, it can be collected in non-invasive manner and we can detect single nucleotide variation by sequencing ctDNA, thus it can be used to mutational analysis of a tumor [26]. That is why ctDNA is easier to collect in the time course than the conventional tumor samples, and the time course sequencing data of ctDNA is increasing [27]. Using the time course data of ctDNA, we will be able to reveal the more detailed dynamics of tumor clonal evolution.

On the other hand, the drug resistance is the problem widely observed in the chemotherapy of various tumors. The mechanism of the drug resistance acquisition is closely related to the clonal evolution of the tumor; Even if we could decrease the major tumor subtype by chemotherapy, resistant subtype becomes predominant and prolifereates [28]. That is why complete recovery from a cancer is difficult. There are some studies which propose the control systems for the suppression of prostate cancer using tumor biomarker PSA [29], however, abundance ratio estimation of the multiple subtypes using biomarker is difficult, and it cannot be applicable to the unknown tumor subtypes and patient specific subtypes. Thus, if we integrate our method with the control engineering, we might be able to optimize the dosage of anti-cancer drugs to suppress tumor growth by estimating the abundance ratio and the growth rate in each time point using ctDNA sequencing and feedback control. Sequencing data driven anti-cancer drug dosage optimization would be useful in the future tailor-made cancer treatment.

5 Supplementary Figures

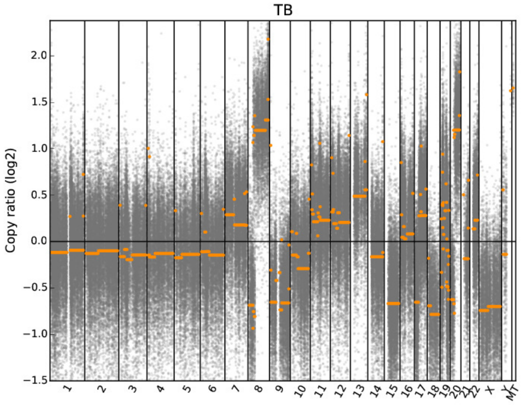


Figure 14: log 2 copy ratio of the tumor TB against the normal sample TN detected using copy number detection tool CNVkit [30].

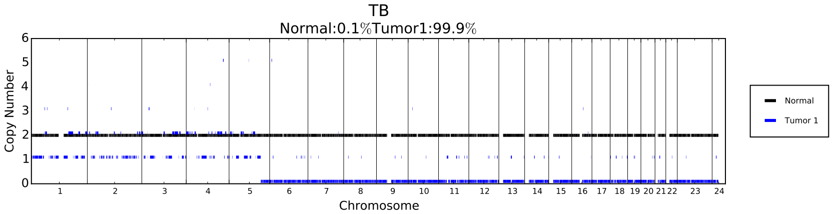


Figure 15: Abundance ratio estimation of the normal and tumor subtypes and tumor copy number estimation of the sample TB using THetA [7].

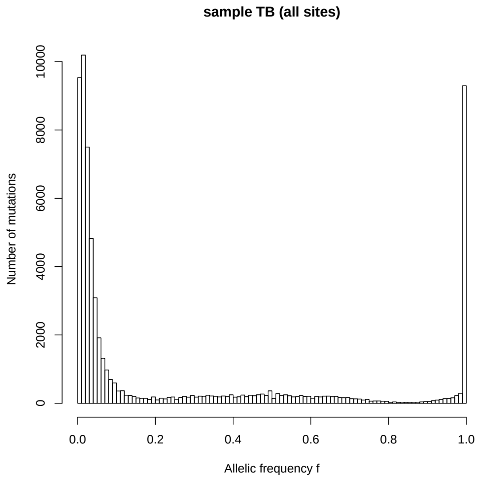


Figure 16: Variant allelic frequency distribution of the mutations detected using a variant caller MuTect [31]. In this figure, non significant SNVs are not excluded.

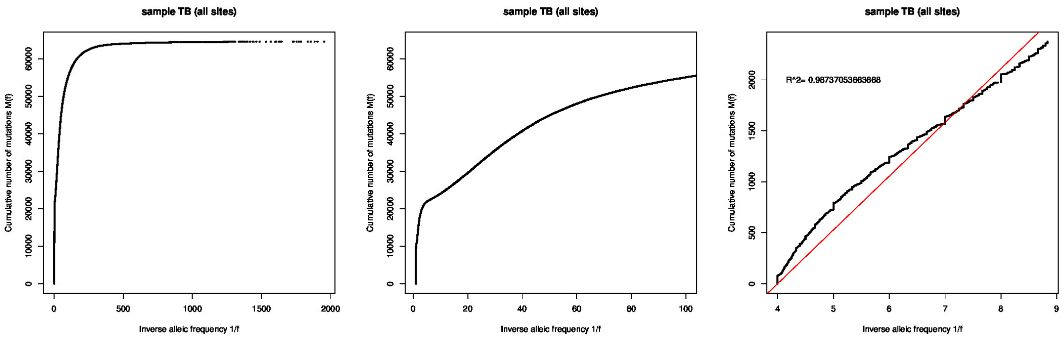
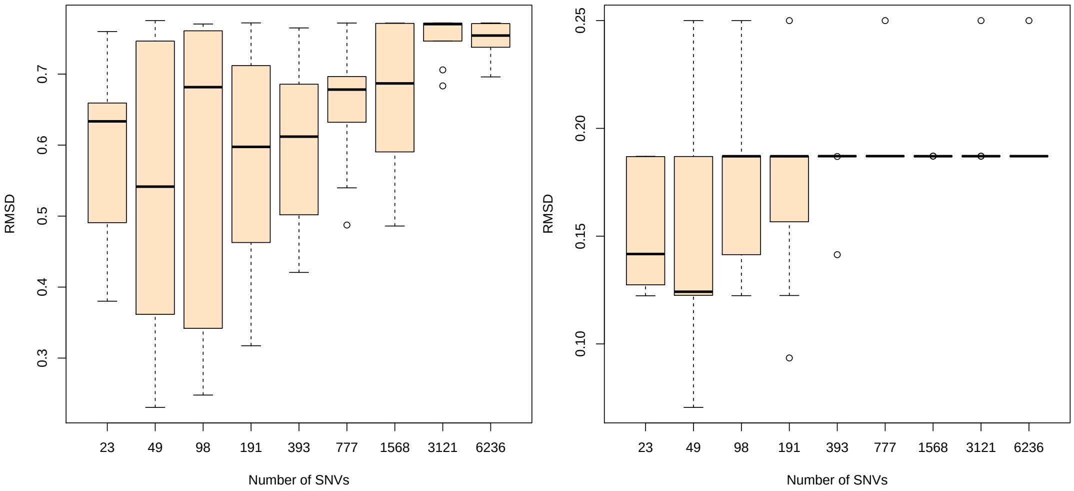


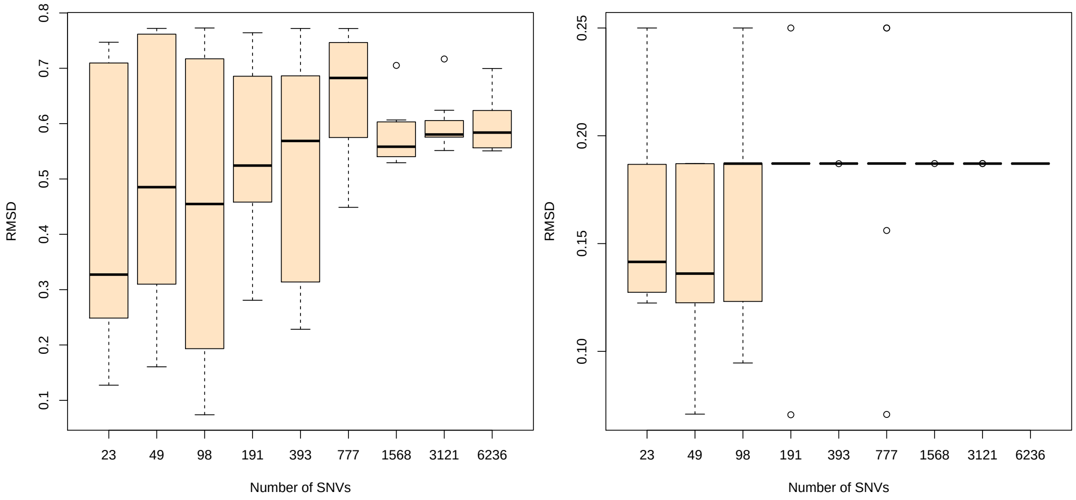
Figure 17: The cumulative number of mutations $M(f)$ plotted against the inverse allelic frequency $1/f$ using the sample TB (non significant SNVs are not excluded). Left figure is the result plotted against all the inverse allelic frequency. In the middle figure, we magnified into the $0 < 1/f < 100$ region. Plateau region can be seen around $1/f = 10$. Further magnifying this region yields the right figure, from which we can see the linear relationship between $M(f)$ and $1/f$. The red line represents the result of the linear regression. From $R^2 > 0.98$, This tumor obeys the neutral evolution.



(18.1) Accuracy of the birth time (t) estimation

(18.2) Accuracy of the abundance ratio (n) estimation

Figure 18: Accuracy of the birth time (t) and abundance ratio (n) estimation against the number of SNVs. Outliers are detected by means of deviation from the $1.5 \times IQR$, where IQR is the interquartile range. Whiskers show the maximum and minimum estimates except for the outliers. If we do not know the subtype in which each SNV was originated, gradient descent estimations fall into the optima which are different from the true parameters. Thus, I could not estimate the true birth time and abundance ratio even if there are thousands of SNVs by marginalizing the subtypes with equal probabilities.



(19.1) Accuracy of the birth time (t) estimation

(19.2) Accuracy of the abundance ratio (n) estimation

Figure 19: Accuracy of the birth time (t) and abundance ratio (n) estimation against the number of SNVs. Outliers are detected by means of deviation from the $1.5 \times IQR$, where IQR is the interquartile range. Whiskers show the maximum and minimum estimates except for the outliers. If we do not know the subtypein which each SNV was originated, gradient descent estimations fall into the optima which are different from the true parameters. Thus, I could not estimate the true birth time and abundance ratio even if there are thousands of SNVs by adding only the maximum probability component with regard to the subtype in which each SNV was originated.

6 Appendix

6.1 Hypergeometric series

Hypergeometric series $F(a, b; c; z)$ is defined as follows [32],

$$F(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!},$$

where

$$(a)_n = \begin{cases} 1 & (n = 0) \\ a(a+1)\cdots(a+n-1) & (n > 0). \end{cases}$$

Also, $X = F(a, b; c; z)$ is the solution of the following differential equation,

$$z(1-z)\frac{d^2X}{dz^2} + [c - (a+b+1)z]\frac{dX}{dz} - abX = 0. \tag{35}$$

6.2 Jacobi polynomials

Jacobi polynomials $P_n^{(\alpha,\beta)}(z)$ is defined using the following generating function $g(t, z)$ [32],

$$g(t, z) = 2^{\alpha+\beta} R^{-1} (1 - t + R)^{-\alpha} (1 + t + R)^{-\beta} = \sum_{n=0}^{\infty} P_n^{(\alpha,\beta)}(z) t^n,$$

where $R = R(t, z) = \left(1 - 2zt + t^2\right)^{1/2}$.

The general form of the Jacobi Polynomials can be expressed as follows,

$$P_n^{(\alpha,\beta)}(z) = \frac{(-1)^n}{2^n n!}(1-z)^{-\alpha}(1+z)^{-\beta}\frac{d^n}{dz^n}\left[(1-z)^{\alpha+n}(1+z)^{\beta+n}\right].$$

Also, Jacobi polynomials are expressed using Hypergeometric series,

$$P_n^{(\alpha,\beta)}(z) = \frac{(\alpha+1)_n}{n!} F\left(-n, 1 + \alpha + \beta + n; \alpha + 1; \frac{1-z}{2}\right).$$

6.3 Gegenbauer polynomials

Gegenbauer polynomials $C_n^{(\alpha)}(z)$ is defined using the following generating function $g(t, z)$ [32],

$$g(t, z) = \frac{1}{(1 - 2zt + t^2)^\alpha} = \sum_{n=0}^{\infty} C_n^{(\alpha)}(z) t^n. \tag{36}$$

Gegenbauer polynomials are expressed using Hypergeometric series or Jacobi polynomials,

$$C_n^{(\alpha)}(z) = \frac{(2\alpha)_n}{n!} F\left(-n, 2\alpha + n; \alpha + \frac{1}{2}; \frac{1-z}{2}\right) = \frac{(2\alpha)_n}{\left(\alpha + \frac{1}{2}\right)_n} P_n^{(\alpha-1/2,\alpha-1/2)}(z)$$

When $\alpha = 3/2$,

$$C_n^{(3/2)}(z) = \frac{n+2}{2} P_n^{(1,1)}(z).$$

Equating $\frac{\partial g}{\partial t} = \frac{2\alpha(z-t)}{1-2zt+t^2} g = \frac{2\alpha(z-t)}{1-2zt+t^2} \sum_{n=0}^{\infty} C_n^{(\alpha)}(z) t^n$ and $\frac{\partial g}{\partial t} = \sum_{n=1}^{\infty} n C_n^{(\alpha)}(z) t^{n-1}$ as an identical equation with respect to t,

$$2(n+\alpha) z C_n^{(\alpha)}(z) = (n+1) C_{n+1}^{(\alpha)}(z) + (n+2\alpha-1) C_{n-1}^{(\alpha)}(z).$$

Differentiating both sides with respect to z yields the following equation,

$$2(n+\alpha)\left(C_n^{(\alpha)}(z) + z\frac{dC_n^{(\alpha)}(z)}{dz}\right) = (n+1)\frac{dC_{n+1}^{(\alpha)}(z)}{dz} + (n+2\alpha-1)\frac{dC_{n-1}^{(\alpha)}(z)}{dz}. \tag{37}$$

On the other hand, equating $\frac{\partial g}{\partial z} = \frac{2\alpha t}{1-2zt+t^2} g = \frac{2\alpha t}{1-2zt+t^2} \sum_{n=0}^{\infty} C_n^{(\alpha)}(z) t^n$ and $\frac{\partial g}{\partial z} = \sum_{n=1}^{\infty} \frac{dC_n^{(\alpha)}(z)}{dz} t^n$ with respect to t as an identical equation with respect to t,

$$2\alpha C_n^{(\alpha)}(z) = \frac{dC_{n+1}^{(\alpha)}(z)}{dz} - 2z\frac{dC_n^{(\alpha)}(z)}{dz} + \frac{dC_{n-1}^{(\alpha)}(z)}{dz}. \tag{38}$$

Cancelling the term $\frac{dC_n^{(\alpha)}(z)}{dz}$ using equation (37) and (38) yields the following relationship,

$$2(n+\alpha) C_n^{(\alpha)} = \frac{d}{dz}\left(C_{n+1}^{(\alpha)}(z) - C_{n-1}^{(\alpha)}(z)\right). \tag{39}$$

$C_n^{(\alpha)}(1)$ and $C_n^{(\alpha)}(-1)$ can be calculated as follows. Noting that $g(t,1) = (1-t)^{-2\alpha}$ and $\frac{d^n g(t,1)}{dt^n} = (2\alpha)_n (1-t)^{-2\alpha-n}$,

$$g(t,1) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{d^n g(0,1)}{dt^n} t^n = \sum_{n=0}^{\infty} \frac{(2\alpha)_n}{n!} t^n = \sum_{n=0}^{\infty} C_n^{(\alpha)}(1) t^n.$$

Thus, $C_n^{(\alpha)}(1) = \frac{(2\alpha)_n}{n!}$. In the same way, $C_n^{(\alpha)}(-1) = \frac{(-1)^n (2\alpha)_n}{n!}$. Especially, when $\alpha = 3/2$, $C_n^{(3/2)}(1) = \frac{(n+1)(n+2)}{2}$ and $C_n^{(3/2)}(-1) = \frac{(-1)^n (n+1)(n+2)}{2}$.

And the following series can be calculated as follows,

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{2n+1}{n^2(n+1)^2} C_{n-1}^{(\alpha)}(z) &= \sum_{n=1}^{\infty} \left(\frac{1}{n^2} - \frac{1}{(n+1)^2}\right) C_{n-1}^{(\alpha)}(z) \\
&= \sum_{n=1}^{\infty} \left(\int_0^{\infty} x\mathrm{e}^{-nx} dx - \int_0^{\infty} x\mathrm{e}^{-(n+1)x} dx\right) C_{n-1}^{(\alpha)}(z) \\
&= \int_0^{\infty} x\mathrm{e}^{-x} \left(1-\mathrm{e}^{-x}\right) \sum_{n=1}^{\infty} C_{n-1}^{(\alpha)}(z) \left(\mathrm{e}^{-x}\right)^{n-1} dx \\
&= \int_0^{\infty} \frac{x\mathrm{e}^{-x}\left(1-\mathrm{e}^{-x}\right)}{\left(1-2z\mathrm{e}^{-x}+\mathrm{e}^{-2x}\right)^{\alpha}} dx \\
&= -\int_0^1 \frac{(1-s)\log s}{\left(1-2zs+s^2\right)^{\alpha}} ds,
\end{aligned} \tag{40}$$

where I used $\int_0^{\infty} x\mathrm{e}^{-nx} dx = 1/n^2$ and equation (36).

6.4 Legendre polynomials

Legendre polynomials $P_n(z)$ is defined using the following generating function $g(t, z)$ [32],

$$g(t,z) = \frac{1}{\sqrt{1-2zt+t^2}} = \sum_{n=0}^{\infty} P_n(z) t^n.$$

The general form of the Legendre Polynomials can be expressed as follows,

$$P_n(z) = \frac{1}{2^n n!} \frac{d^n}{dz^n} \left[(z^2-1)^n \right].$$

And the first and second term are calculated as follows,

$$P_0(z) = 1, \; P_1(z) = z.$$

Also, Legendre polynomials are expressed using Hypergeometric series or Jacobi polynomials, or Gegenbauer polynomials,

$$P_n(z) = C_n^{(1/2)}(z) = P_n^{(0,0)}(z) = F\left(-n, 1+n; 1; \frac{1-z}{2}\right).$$

$P_n(1)$ and $P_n(-1)$ can be calculated as follows,

$$P_n(1) = C_n^{(1/2)}(1) = \frac{1_n}{n!} = 1, \tag{41}$$

$$P_n(-1) = C_n^{(1/2)}(-1) = \frac{(-1)^n 1_n}{n!} = (-1)^n. \tag{42}$$

Using the following relationship,

$$\frac{d^2}{dz^2}\left[(z^2-1)^{n+2}\right] = 2(n+2)\left[(2n+3)(z^2-1)^{n+1} + 2(n+1)(z^2-1)^n\right],$$

There is a relationship,

$$\begin{aligned} P_{n+2}(z) - P_n(z) &= \frac{2n+3}{2(n+1)} \frac{1}{2^n n!} \frac{d^n}{dz^n}\left[(z^2-1)^{n+1}\right] \\ &= -\frac{2n+3}{2(n+1)}(1-z^2) P_n^{(1,1)}(z) \\ &= -\frac{2n+3}{(n+1)(n+2)}(1-z^2) C_n^{(3/2)}(z), \end{aligned} \tag{43}$$

between the Legendre polynomials and Gegenbauer polynomials.

The following series can be calculated using equation (43).

$$\sum_{n=1}^{\infty} \frac{2n+1}{n(n+1)} C_{n-1}^{(3/2)}(z) = \sum_{n=1}^{\infty} \frac{P_{n-1}(z) - P_{n+1}(z)}{1-z^2} = \frac{P_0(z) + P_1(z)}{1-z^2} = \frac{1}{1-z}. \tag{44}$$

$$\sum_{n=1}^{\infty} \frac{(-1)^n(2n+1)}{n(n+1)} C_{n-1}^{(3/2)}(z) = \sum_{n=1}^{\infty} \frac{(-1)^n(P_{n-1}(z) - P_{n+1}(z))}{1-z^2} = \frac{-P_0(z) + P_1(z)}{1-z^2} = -\frac{1}{1+z} \tag{45}$$

Also, using equation (39),

$$(2n+1)P_n(z) = \frac{d}{dz}\left(P_{n+1}(z) - P_{n-1}(z)\right). \tag{46}$$

Using equation (35), $X = P_n(z)$ is the solution of the following differential equation,

$$\frac{d}{dz}\left[(1-z^2)\frac{d}{dz}P_n(z)\right] = -n(n+1)P_n(z). \tag{47}$$

Integrating both sides of (47) and dividing by $1-z^2$,

$$\frac{d}{dz}P_n(z) = -\frac{n(n+1)}{(1-z^2)}\int P_n(z)dz = -\frac{n(n+1)}{(1-z^2)(2n+1)}\left(P_{n+1}(z) - P_{n-1}(z)\right) = C_{n-1}^{(3/2)}(z), \tag{48}$$

where I used equation (46) and (43).

6.5 Exponential integral

The first order exponential integral in defined as follows [32],

$$E_1(z) = \int_z^\infty \frac{\mathrm{e}^{-z}}{z}dz \sim \frac{\mathrm{e}^{-z}}{z}\left(1 - \frac{1!}{z} + \frac{2!}{z^2} - \cdots\right) \text{ (if } z \gg 1).$$

Acknowledgements

We would like to thank my advisor Kiryu Hisanori, and our lab members and alumni for the many constructive advices. I also thank the student room members.

References

[1] Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[2] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[3] Li Ding, Timothy J Ley, David E Larson, Christopher A Miller, Daniel C Koboldt, John S Welch, Julie K Ritchey, Margaret A Young, Tamara Lamprecht, Michael D McLellan, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, 2012.

[4] Ali Bashashati, Gavin Ha, Alicia Tone, Jiarui Ding, Leah M Prentice, Andrew Roth, Jamie Rosner, Karey Shumansky, Steve Kalloger, Janine Senz, et al. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *The Journal of pathology*, 231(1):21–34, 2013.

[5] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.

[6] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, et al. Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology*, 30(5):413–421, 2012.

[7] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. Inferring intra-tumor heterogeneity from high-throughput dna sequencing data. In *Annual International Conference on Research in Computational Molecular Biology*, pages 171–172. Springer, 2013.

[8] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, et al. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome research*, 24(11):1881–1893, 2014.

[9] Christopher A Miller, Brian S White, Nathan D Dees, Malachi Griffith, John S Welch, Obi L Griffith, Ravi Vij, Michael H Tomasson, Timothy A Graubert, Matthew J Walter, et al. Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*, 10(8):e1003665, 2014.

[10] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398, 2014.

[11] Noushin Niknafs, Violeta Beleva-Guthrie, Daniel Q Naiman, and Rachel Karchin. Subclonal hierarchy inference from somatic mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Comput Biol*, 11(10):e1004416, 2015.

[12] Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome biology*, 16(1):1, 2015.

[13] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.

[14] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15(1):1, 2014.

[15] Ke Yuan, Thomas Sakoparnig, Florian Markowetz, and Niko Beerenwinkel. Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology*, 16(1):1, 2015.

[16] Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W Kinzler, Bert Vogelstein, and Martin A Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.

[17] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97–159, 1931.

[18] Yoh Iwasa, Martin A Nowak, and Franziska Michor. Evolution of resistance during clonal expansion. *Genetics*, 172(4):2557–2566, 2006.

[19] Rick Durrett, Jasmine Foo, Kevin Leder, John Mayberry, and Franziska Michor. Intratumor heterogeneity in evolutionary models of tumor progression. *Genetics*, 188(2):461–477, 2011.

[20] Rick Durrett. Population genetics of neutral mutations in exponentially growing cancer cell populations. *The annals of applied probability: an official journal of the Institute of Mathematical Statistics*, 23(1):230, 2013.

[21] Marc J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature genetics*, 2016.

[22] Andrei Kolmogoroff. Über die analytischen methoden in der wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104(1):415–458, 1931.

[23] Motoo Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.

[24] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[25] Reeves Fletcher and Colin M Reeves. Function minimization by conjugate gradients. *The computer journal*, 7(2):149–154, 1964.

[26] Jens G Lohr, Viktor A Adalsteinsson, Kristian Cibulskis, Atish D Choudhury, Mara Rosenberg, Peter Cruz-Gordillo, Joshua M Francis, Cheng-Zhong Zhang, Alex K Shalek, Rahul Satija, et al. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nature biotechnology*, 32(5):479–484, 2014.

[27] Muhammed Murtaza, Sarah-Jane Dawson, Dana WY Tsui, Davina Gale, Tim Forshew, Anna M Piskorz, Christine Parkinson, Suet-Feung Chin, Zoya Kingsbury, Alvin SC Wong, et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma dna. *Nature*, 497(7447):108–112, 2013.

[28] Dan A Landau, Scott L Carter, Gad Getz, and Catherine J Wu. Clonal evolution in hematological malignancies and therapeutic implications. *Leukemia*, 28(1):34–43, 2014.

[29] Aiko Miyamura Ideta, Gouhei Tanaka, Takumi Takeuchi, and Kazuyuki Aihara. A mathematical model of intermittent androgen suppression for prostate cancer. *Journal of nonlinear science*, 18(6):593–614, 2008.

[30] Eric Talevich, A Hunter Shain, Thomas Botton, and Boris C Bastian. Cnvkit: genome-wide copy number detection and visualization from targeted dna sequencing. *PLoS Comput Biol*, 12(4):e1004873, 2016.

[31] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219, 2013.

[32] FWJ Olver, DW Lozier, RF Boisvert, and CW Clark. Nist digital library of mathematical functions. *Online companion to [65]: http://dlmf. nist. gov*, 2010.