bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Allele Frequency Spectrum in a Cancer Cell Population

H. Ohtsuki[*] **and H. Innan**[*,1]

[*]SOKENDAI, The Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan.

ABSTRACT A cancer grows from a single cell, thereby constituting a large cell population. In this work, we are interested in how mutations accumulate in a cancer cell population. We provided a theoretical framework of the stochastic process in a cancer cell population and obtained near exact expressions of allele frequency spectrum or AFS (only continuous approximation is involved) from both forward and backward treatments under a simple setting; all cells undergo cell division and die at constant rates, b and d, respectively, such that the entire population grows exponentially. This setting means that once a parental cancer cell is established, in the following growth phase, all mutations are assumed to have no effect on b or d (i.e., neutral or passengers). Our theoretical results show that the difference from organismal population genetics is mainly in the coalescent time scale, and the mutation rate is defined per cell division, not per time unit (e.g., generation). Except for these two factors, the basic logic are very similar between organismal and cancer population genetics, indicating that a number of well established theories of organismal population genetics could be translated to cancer population genetics with simple modifications.

KEYWORDS allele frequency spectrum; cancer; population genetics; branching theory; coalescent theory

A tumor grows from a single cell, as has been well recognized for several decades (Muller 1950; Nowell 1976; Fidler 1978; Dexter *et al.* 1978; Merlo *et al.* 2006). Through the growth process, cells accumulate various kinds of mutations, from simple point mutations to more drastic changes at the chromosomal level, such as deletions and amplifications (Sjöblom *et al.* 2006; Wood *et al.* 2007; Network 2008; Network *et al.* 2012, 2014; Garraway and Lander 2013; Vogelstein *et al.* 2013). There are two major categories of mutations in cancer cells, driver and passenger mutations. The former are generally cell autonomous, that is, they increase the reproductive ability of the carrier cell (i.e., adaptive), while the latter have no effect on the reproductive ability (i.e., neutral). A new technology for genome sequencing from a single cell opened a new window in cancer genetics, because sequencing a number of cells from a single tumor makes it possible to identify heterogeneity in the catalog of driver and passenger mutations between cells, from which we are able to infer when and how the tumor has grown (Navin 2015).

Population genetics provides a solid theoretical framework for a wide variety of such inference methods (e.g., Nielsen and Slatkin 2013; Wakeley 2009). The coalescent (Kingman 1982; Hudson 1983; Tajima 1983) plays the central role to make the theoretical predictions of the pattern of genetic variation, which can be used to compute the likelihood of the observed variation data (Donnelly 1996; Tavaré *et al.* 1997). It concerns the history of the sampled individuals, by tracing their ancestral lineages up to the MRCA, most recent common ancestor (e.g., Nielsen and Slatkin 2013; Wakeley 2009).

One might think that the coalescent theory can be directly applied to cancer cells due to the obvious analogy; all cancer cells should follow a simple genealogy up to their MRCA. However, the direct

Copyright © 2017 by the Genetics Society of America
doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Sunday 29th January, 2017

[1] *Corresponding author:* Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan. E-mail: innan_hideki@soken.ac.jp

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

application of the standard population genetics (i.e., organismal population genetics) to a cancer cell population may not be exactly correct because of some fundamental differences in the propagation system, as we explain below (see also Sidow and Spies 2015).

In organismal population genetics, the process can be specified by the expected number of offsprings for each individual, namely, the fitness (e.g., Crow and Kimura 1970; Ewens 1979). In the Wright-Fisher model with N haploids (Fisher 1930; Wright 1931), all individuals are randomly replaced every generation, and individuals with higher fitness likely produce more offsprings. In the Moran model (Moran 1962), individuals are replaced one by one, that is, one step consists of a coupling event of birth and death; one dead individual is replaced by the offspring of one randomly chosen individual from the population allowing self-replacement. Consequently, all individuals are on average replaced in N steps, which roughly correspond to one generation in the Wright-Fisher model. It has been well known that theoretical results under the two models are nearly identical in various cases (e.g., Crow and Kimura 1970; Ewens 1979; Wakeley 2009; Bhaskar and Song 2009). Through this random mating process either in the Wright-Fisher or Moran model, mutations that arise in the population will fix or get extinct by the joint action of random genetic drift and selection. A mutation is defined as adaptive when it increases the fitness of the carrier individual.

The evolutionary process of a cancer cell population does not follow such a simple replacement system. Figure 1 illustrates the process from cancer initiation, progression to the following rapid growth, which may be roughly divided into two major phases, and the applicability of organismal population genetics may differ depending on the phase. The first phase (Phase I) from cancer initiation to initial progression could be well handled under the organismal population genetic framework (Komarova *et al.* 2003; Iwasa *et al.* 2004; Michor *et al.* 2004). This phase is commonly modeled in a constant-size population of cells. Most theoretical models for cancer initiation suppose that a tissue consists of a number of small compartments of cells and that cancer initiation can occur in a compartment. The system starts with a normal compartment with a certain number of asexually reproducing normal cells, which is denoted by N_0. N_0 is usually assumed to be constant because the number of cells in a healthy tissue is maintained roughly constant by homeostatic systems, that is, cell division occurs when needed. The Moran model is more suitable to apply to this process than the Wright-Fisher model because it can be modeled such that one cell death asks for one cell division. Indeed, the Moran model has been frequently used to explore a number of problems on cancer initiation (reviewed in Michor *et al.* 2004). One of the major problems is how a cancer initiates. A compartment of a normal tissue could become a cancer when oncogenes are activated and/or tumor-suppressor genes (TSGs) are inactivated. It is believed that at least several mutational alternations in cancer genes (oncogenes and TSGs) are required for the formation of a parental cancer cell. Such accumulation of mutations in cancer genes could allow a cell to acquire typical behaviors of cancer cells, for example, avoiding apoptosis (programmed cell death) that makes it difficult to maintain the equilibrium between birth and death in the compartment, thereby shifting towards uncontrolled proliferation (neoplasia). There are a large body of theory only for the fixation process of mutations in cancer genes, especially for the inactivation of TSGs, perhaps because the problem is mathematically too simple for the activation of oncogenes (Michor *et al.* 2004). Inactivation of a TSG involves the fixation of a double-mutant, that is, both alleles have to be silenced according to Knudson's two-hit model (Knudson 1971). This situation is very similar to the fixation process of a pair of compensatory mutations in organismal population genetics (Innan and Stephan 2001), and the results are indeed in good agreement (Iwasa *et al.* 2004). Thus, it can be considered that the applicability of organismal population genetics is quite good in Phase I because the assumption of a constant-size population roughly holds so that the stochastic process through random genetic drift works as organismal population genetics predicts.

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

By contrast, in the second phase (Phase II) where cells have acquired extraordinary high proliferative ability, the population grows very rapidly, and the stochastic process is less important for changing allele frequencies because most cells have very low death rates by avoiding apoptosis and their cell divisions occur independently of each other. As a consequence, a fixation of adaptive mutation hardly occurs in a cancer cell population because the spread of an adaptive mutation does not necessarily kill other cells with lower reproductive rates, as has been pointed out by Sidow and Spies (2015). This reproducing system is quite different from that organismal population genetics supposes.

We here ask how the well established theory of organismal population genetics can be applied to Phase II that presumably involves an exponential growth. In particular, we are interested in the allele frequency spectrum (AFS, or SFS: site frequency spectrum) of passenger mutations in a cancer cell population. AFS is summarized information of genotype data that are frequently used in organismal population genetics. Under the basic neutral theory of the coalescent for a constant size population (Kingman 1982; Hudson 1983; Tajima 1983) with the assumption of infinitely many sites (Kimura 1969), the expected AFS can be described in a simple form (Fu 1995), but for a non-constant size population, it is not very straightforward to obtain the expected AFS in a simple closed form. Even with any complicated demographic setting, the expected AFS can be written as a function of the expectations of coalescent times (Griffiths and Tavaré 1994, 1998), but these expectations are not easy to derive in a simple form in many cases although possible computationally (Williamson *et al.* 2005; Polanski and Kimmel 2003; Polanski *et al.* 2003). AFS provides substantial information on the past demography, making it possible to infer various demographic parameters including population size changes and migration rates (Adams and Hudson 2004; Williamson *et al.* 2005; Gutenkunst *et al.* 2009; Bhaskar *et al.* 2015; Gao and Keinan 2016).

In this article, we consider a model of a rapidly growing cancer cell population for exploring how mutations accumulate within the cancer cell population. We present some derivations for the expected AFS of derived mutations in the final tumor (at t_1 in Figure 1), which could be useful to infer when the exponential growth started and how fast the tumor has grown. There has been extensive works on a cancer cell population by Durrett (Durrett 2013, 2015), who provided approximate formulas to the sample-based AFS. We have here obtained analytical expressions of the expected AFS in a near exact form (only continuous approximation is involved) by both forward (branching theory) and backward (coalescent theory) treatments. The former is in a simpler form that is useful for intuitive understanding of the process, while the latter provides a solid theoretical framework for coalescent (backward) simulations of a cancer cell population. Our near exact result is compared with Durrett's approximate formulas, together with some simulation results.

It should be noted that our interest is in passenger mutations in the second phase with the assumption of no driver mutations so that the increase of the cancer cell population size can be approximated by an exponential function. There is no doubt that a number of driver mutations are involved in the first phase (e.g., Knudson 1971; Michor *et al.* 2004; Sjöblom *et al.* 2006; Network *et al.* 2014), but there are extensive debates on the potential role of driver mutations in the second phase. Some authors suggest that the role of driver mutations may be quite limited after the original cancer cell is stablished and most mutations occurs in the following growth phase may be passengers (Uchi *et al.* 2016; Sottoriva *et al.* 2015), whereas some point out the importance of driver mutations (Williams *et al.* 2016; Waclaw *et al.* 2015; Marusyk *et al.* 2014). Because our model assumes no driver mutations in the second growth phase, the theoretical result could be used as a null model for testing the role of driver mutations in the second phase.

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

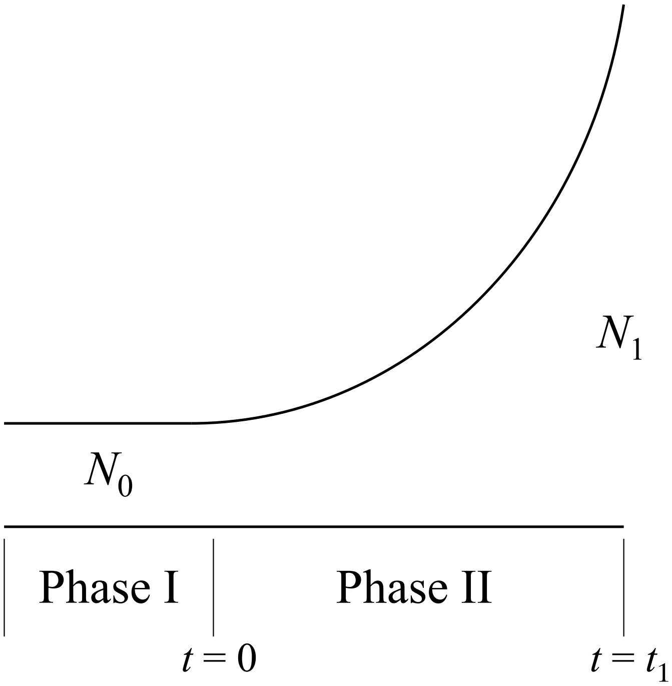

Figure 1 Illustrating the model of the growth of a cancer cell population.

MODEL

Our model (Figure 1) considers an exponentially growing population starting with N_0 asexually reproductive cells. The reproductive ability of a cell is specified by the cell division rate (birth rate) and death rate per time unit, denoted by b and d, respectively, which are assumed to be constant over time. The tumor starts growing at time $t = 0$, and let $N(t)$ be the number of cells at time t. For convenience, we define t_1 such that $N(t_1) = N_1$ is satisfied for the first time. Under this setting, because it is obvious that the Moran model does not work, we use the branching process.

We assume $b \gg d$ so that the tumor grows approximately exponentially at rate $r = b - d$ and the number of cells at t is approximately given by

$$N(t) = N_0 \exp[(b-d)t] \tag{1}$$

This equation is a very good approximation unless N_0 is very small. Note that in reality $N(t)$ follows some distribution, but our deterministic treatment on $N(t)$ does not affect the following results much.

The rate of passenger mutation is given such that at each cell division one of the daughter cells receives a novel mutation at rate μ. We assume a very small rate per site so that the assumption of the infinite-site model (Kimura 1969) holds.

Forward Treatment by Branching Process: We aim to obtain the expected derived allele frequency spectrum (AFS) when the total number of cells is N_1 (i.e., $t = t_1$), where we assume that $N_1 \gg N_0$. The expected number of passenger mutations that are shared by i cells at time $t = t_1$ is denoted by $S(i, \mu, t_1)$. Because of our deterministic assumption (i.e, Equation (1)), t_1 is given such that it satisfies $N_1/N_0 = \exp[(b-d)t_1]$.

We first consider how many cells at $t = t_1$ share a particular mutation that occurred at $t = t_1 - t'$. We here use the well-known formula under the branching process: the probability density function

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(pdf) of the number of daughter cells (i) of a particular single individual after t' time units is given by:

$$P(i,b,d,t') = \begin{cases} x(t') & \text{if } i = 0 \\ \{1 - x(t')\}\{1 - y(t')\}y(t')^{i-1} & \text{if } i \geq 1 \end{cases} \tag{2}$$

(Bailey 1964), where

$$\begin{aligned} x(t') &= \frac{de^{(b-d)t'} - d}{be^{(b-d)t'} - d}, \\ y(t') &= \frac{be^{(b-d)t'} - b}{be^{(b-d)t'} - d}. \end{aligned} \tag{3}$$

This formula provides an unconditional distribution of the number of individuals having a specific origin, which is independent of the total population size. Nevertheless, we use this formula by ignoring the effect of the total population size. This simplification is reasonable and the effect on the theoretical treatments is negligible even though it is technically possible that i exceeds the total population size. This is because i is usually not a large number unless $N(t_1 - t')$ is unrealistically small.

We then obtain $S(i, \mu, t_1)$, the expected number of mutations with frequency i in the final tumor by considering all potential mutations that occur $0 < t < t_1$. Because the population mutation rate at time t is $N(t)b\mu$, we obtain $S(i, \mu, t_1)$ for $i \geq 1$:

$$\begin{aligned} S(i,\mu,t_1) &= \int_0^{t_1} P(i,b,d,t_1 - t) \cdot N(t)b\mu \mathrm{d}t \\ &= \int_{y(t_1)}^{0} \underbrace{\left(1 - \frac{d}{b}w\right)(1-w)w^{i-1}}_{=P(i,b,d,t_1-t)} \cdot \underbrace{N_0 e^{(b-d)t_1} \frac{b(1-w)}{b-dw} b\mu}_{=N(t)b\mu} \underbrace{\left(-\frac{1}{(b-dw)(1-w)}\right) \mathrm{d}w}_{=\mathrm{d}t} \\ &= N_1\mu \int_0^{y(t_1)} \frac{1-w}{1-\frac{d}{b}w} w^{i-1} \mathrm{d}w \\ &\approx N_1\mu \int_0^{1} \frac{1-w}{1-\frac{d}{b}w} w^{i-1} \mathrm{d}w \\ &= N_1\mu \sum_{k=0}^{\infty} \frac{1}{(i+k)(i+k+1)} \left(\frac{d}{b}\right)^k, \end{aligned} \tag{4}$$

where we set $w = y(t_1 - t)$ and assume $y(t_1) \approx 1$ and N_0 is very small. We again note that because of the nature of our approximation, it is possible to compute $S(i, \mu, t_1)$ even for $i > N_1$. For a practical calculation of $S(i, \mu, t_1)$, however, this treatment should not matter so much as mentioned above. Equation (4) means that the relative frequency distribution of $S(i, \mu, t_1)$ is determined by the ratio of d to b, while $N_1\mu$ determines the absolute number of mutations.

It is straightforward to obtain the expected normalized AFS (pdf of i given a segregating mutation, i.e., $i = (1, 2, 3, ..., N_1)$) as

$$AFS(i,t_1) = \frac{S(i,\mu,t_1)}{\sum_{i'=1}^{N_1} S(i',\mu,t_1)}, \tag{5}$$

where, for a large N_1, the denominator of eq.(5) is approximated by

$$\sum_{i'=1}^{N_1} S(i',\mu,t_1) \approx \sum_{i'=1}^{\infty} S(i',\mu,t_1) = \int_0^1 \frac{1}{1-\frac{d}{b}w} \mathrm{d}w = -\frac{b}{d} \log\left(1 - \frac{d}{b}\right). \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Of particular importance is the case of $b \gg d$, that is, the population grows very rapidly, where Equation (4) becomes

$$S(i, \mu, t_1) = N_1 \mu \cdot \frac{1}{i(i+1)} \tag{7}$$

and

$$AFS(i, t_1) = \frac{\frac{1}{i(i+1)}}{\sum_{i'=1}^{N_1} \frac{1}{i'(i'+1)}} \xrightarrow{N_1 \to \infty} \frac{1}{i(i+1)}. \tag{8}$$

At this limit, it is interesting to note that $S(i, \mu, t_1)$ is independent of b or d.

We can consider the opposite extreme, $b \sim d$, where the underlying assumption of our calculation (*i.e.*, the population grows exponentially) is obviously broken. Nevertheless, if we formally proceed our calculation by taking the limit $b \to d$, we obtain

$$S(i, \mu, t_1) \approx N_1 \mu \cdot \frac{1}{i}, \tag{9}$$

which reproduces the result for a Moran process in a constant-size population (Fu 1995; Griffiths and Tavaré 1998; Wakeley 2009). This is not a coincidence because our assumption $b = d$ with deterministic treatment simply means a constant size population, but this equation does not work well in our randomly reproductive population without keeping the population size constant. For AFS, we have

$$AFS(i, t_1) = \frac{\frac{1}{i}}{\sum_{i'=1}^{N_1} \frac{1}{i'}} \overset{N_1 \gg 1}{\approx} \frac{1}{i} \cdot \frac{1}{\log N_1 + \gamma}, \tag{10}$$

where $\gamma \equiv 0.577215 \cdots$ is the Euler's constant.

We performed forward simulation to check how our equations work. Our simulations assumed that $N_0 = 10$, $N_1 = 10^5$, $\mu = 10^{-4}$, $b = 4$, and $d = \{0, 0.2, 0.4, 1\}$, and Figure 2 shows the average spectra (up to $i = 25$) over 10^5 simulation runs. Theoretical results based on Equation (4) are shown in closed circles. It is demonstrated that Equation (4) is in excellent agreement with the simulation results (colored open boxed) for all four cases. We further compare the values computed by Equation 7 (filled triangles in Figure 2), which is a simple approximation to Equation (4) when $b \gg d$. We find that Equations (4) and (7) produce almost identical numerical values, which are indistinguishable in Figure 2, indicating that the simple approximation works very well when $d = 0$. Furthermore, Equation (7) could be in fairly good agreement with the results of Equation (4) with $d/b = 0.05$, indicating that the simple form (Equation (7)) can be a good approximation when $d/b \ll 0.05$.

For applying our theoretical result to data, it is more convenient to consider a sample rather than the entire cell population. Suppose that n random cells are sampled from the population. Then, the expected number of mutations that are shared by i $(1 \leq i \leq n)$ cells in a sample of size n is given by

$$S_{sample}(i, \mu, t_1 | n) = \sum_{i'=i}^{N_1} \frac{\binom{i'}{i}\binom{N_1 - i'}{n - i}}{\binom{N_1}{n}} S(i', \mu, t_1), \tag{11}$$

which is, for a large N_1, approximated by using a Poisson distribution as

$$S_{sample}(i, \mu, t_1 | n) = \sum_{i'=1}^{N_1} Poisson_{\frac{ni'}{N_1}}(i) \cdot S(i', \mu, t_1), \tag{12}$$

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

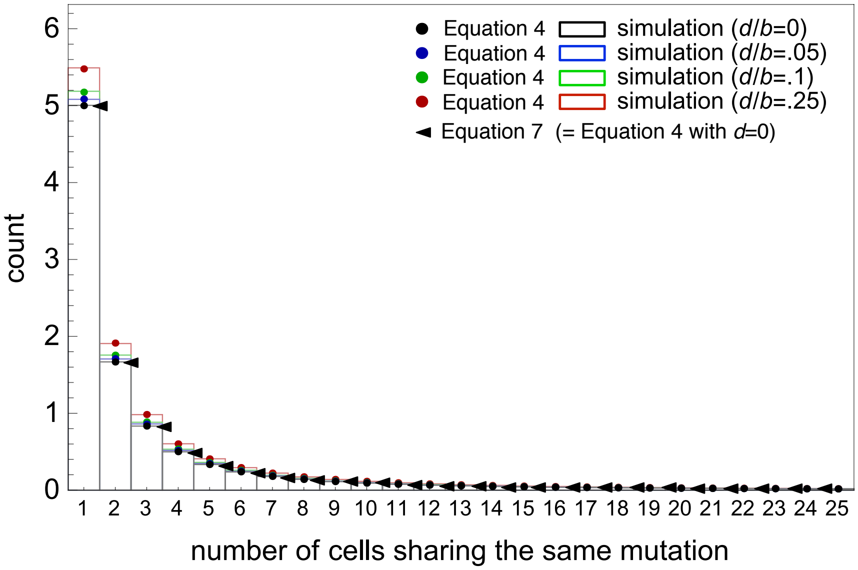

Figure 2 Population allele frequency spectra, $AFS(i, t_1)$, when $d/b = \{0, 0.05, 0.2, 0.25\}$. The theoretical results from Equations (4) and (7) are compared with simulations. Forward simulations were performed with $N_0 = 10$, $N_1 = 10^5$, $\mu = 10^{-4}$, $b = 4$, and $d = \{0, 0.2, 0.4, 1\}$. It should be noted that Equation (4) with $d = 0$ is identical to Equation (7).

where $Poisson_\lambda(i) = \frac{\lambda^i}{i!}e^{-\lambda}$. Then, it is straightforward to obtain normalized sample AFS. If we include fixed mutations, the normalized sample AFS is given by

$$AFS_{sample}(i, t_1|n) = \frac{S_{sample}(i, \mu, t_1|n)}{\sum_{i'=1}^{n} S_{sample}(i', \mu, t_1|n)} = \frac{\sum_{i'=1}^{N_1} Poisson_{\frac{ni'}{N_1}}(i) \cdot S(i', \mu, t_1)}{\sum_{i'=1}^{n} \sum_{i''=1}^{N_1} Poisson_{\frac{ni''}{N_1}}(i') \cdot S(i'', \mu, t_1)}, \qquad (13)$$

and if fixed mutations are ignored

$$AFS_{sample}(i, t_1|n) = \frac{S_{sample}(i, \mu, t_1|n)}{\sum_{i'=1}^{n-1} S_{sample}(i', \mu, t_1|n)} = \frac{\sum_{i'=1}^{N_1} Poisson_{\frac{ni'}{N_1}}(i) \cdot S(i', \mu, t_1)}{\sum_{i'=1}^{n-1} \sum_{i''=1}^{N_1} Poisson_{\frac{ni''}{N_1}}(i') \cdot S(i'', \mu, t_1)}. \qquad (14)$$

Backward Treatment by the Coalescent: The coalescent is one of the major theories in organismal population genetics. It is a sample-based theory: The lineages of sampled individuals are traced backward in time until they coalesce into their MRCA (most recent common ancestor). We here apply this logic to a sampled cells from a tumor, and obtain essentially the same theoretical results as those from the forward treatment (i.e., Equations (11 - 14)).

Let us consider a pair of random (different) cells from the final tumor with N cells, where N is already a large number. Because the following argument works at any time in Phase II (assuming N_0 is very small), we shall use N for the population size rather than N_1. We consider backward time τ from the present, such that the present time is set to $\tau = 0$. Let T_2 be the time it takes for the two lineages to coalesce. We consider an infinitesimally small time interval $\Delta\tau$ such that at most one event (birth or death) can occur. The conditional probability that the population size was $N - 1$ at

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

time $\Delta\tau$ backward, conditioned on that the present population size is N is given by

$$\begin{aligned} P(N_{\tau=\Delta\tau} = N-1 | N_{\tau=0} = N) &= \frac{P(N_{\tau=\Delta\tau} = N-1) P(N_{\tau=0} = N | N_{\tau=\Delta\tau} = N-1)}{P(N_{\tau=0} = N)} \\ &= \frac{P(N_{\tau=\Delta\tau} = N-1)}{P(N_{\tau=0} = N)} b(N-1)\Delta\tau. \end{aligned} \tag{15}$$

The probabilities $P(N_{\Delta\tau} = N-1)$ and $P(N_0 = N)$ can be calculated based on the forward process, but for a large N it is expected that their difference is at most of order $\Delta\tau$, so the leading term of the expression above is $b(N-1)\Delta\tau$. This equation represents the probability that a birth event occurred in the interval $\Delta\tau$. The birth event can cause the coalescence between two specific lineages, with probability

$$\frac{2}{N}\frac{1}{N-1}, \tag{16}$$

and therefore the probability of coalescence is, up to the first order of $\Delta\tau$, given by

$$b(N-1)\Delta\tau \cdot \frac{2}{N(N-1)} = \frac{2b}{N}\Delta\tau. \tag{17}$$

In the mean time, we must take into account the fact that the population is shrinking at rate $r = b - d$ backward in time (Slatkin and Hudson 1991). The rate of coalescence between two lineages at time τ is approximated by

$$\rho_{2,\tau} = \frac{2b}{Ne^{-r\tau}}. \tag{18}$$

Note that this formula is consistent with a well-known formula for the Moran process when the population size is fixed (e.g., Wakeley 2009), namely, setting $b = d = 1$ reproduces

$$\rho_{2,\tau} = \frac{2}{N}, \tag{19}$$

which is the per-generation rate of coalescence for the Moran model.

Let $P_2(\tau)$ be the probability that the coalescence between the two lineages have not occurred yet by time τ, for which the following differential equation holds:

$$\frac{\mathrm{d}P_2(\tau)}{\mathrm{d}\tau} = -\rho_{2,\tau} P_2(\tau) = -\frac{2b}{Ne^{-r\tau}} P_2(\tau). \tag{20}$$

With $P_2(0) = 1$ as a boundary condition, the solution is given by a double exponential function:

$$P_2(\tau) = \exp\left[-\frac{2b}{rN}(e^{r\tau} - 1)\right]. \tag{21}$$

Therefore, the density function of coalescent time T_2 is given by

$$\begin{aligned} \text{Density}(T_2 = \tau_2) &= -\frac{\mathrm{d}P_2(\tau_2)}{\mathrm{d}\tau_2} \\ &= \frac{2b}{N} \exp\left[-\frac{2b}{rN}(e^{r\tau_2} - 1) + r\tau_2\right] \\ &= \frac{2be^{r\tau_2}}{N} \exp\left[-\frac{2b}{rN}(e^{r\tau_2} - 1)\right]. \end{aligned} \tag{22}$$

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Following the same logic, for $k(> 2)$ cells, we have

$$\begin{cases} \rho_{k,\tau} = \dfrac{k(k-1)b}{Ne^{-r\tau}} \\ P_k(\tau) = \exp\left[-\dfrac{k(k-1)b}{rN}(e^{r\tau}-1)\right], \\ \text{Density}(T_k = \tau_k) = \dfrac{k(k-1)be^{r\tau_k}}{N}\exp\left[-\dfrac{k(k-1)b}{rN}(e^{r\tau_k}-1)\right]. \end{cases} \tag{23}$$

In order to consider the coalescent process of n sampled cells up to their MRCA, we are interested in the joint pdf of $\{T_2, T_3, ..., T_{n-1}, T_n\}$, which is given by

$$\begin{aligned} &\text{Density}(\{T_2, T_3, ..., T_{n-1}, T_n\} = \{\tau_2, \tau_3, ..., \tau_{n-1}, \tau_n\}) = \\ &\quad \frac{2be^{r\tau_2}}{N_{k=2}}\exp\left[-\frac{2b}{rN_{k=2}}(e^{r\tau_2}-1)\right] \\ &\quad \times\frac{6be^{r\tau_3}}{N_{k=3}}\exp\left[-\frac{6b}{rN_{k=3}}(e^{r\tau_3}-1)\right] \\ &\quad \times\dots \\ &\quad \times\frac{(n-1)(n-2)be^{r\tau_{n-1}}}{N_{k=n-1}}\exp\left[-\frac{(n-1)(n-2)b}{rN_{k=n-1}}(e^{r\tau_{n-1}}-1)\right] \\ &\quad \times\frac{n(n-1)be^{r\tau_n}}{N_{k=n}}\exp\left[-\frac{n(n-1)}{rN_{k=n}}(e^{r\tau_n}-1)\right], \end{aligned} \tag{24}$$

where $N_{k=j}$ is the population size at the moment when the original n lineages coalesce up to j lineages. In other words, $N_{k=j}$ is the population size $\sum_{\ell=j+1}^{n}\tau_\ell$ time units before the present. Thus, the coalescent times are not independent one another, that is, T_j is given conditional on $\sum_{\ell=j+1}^{n}\tau_\ell$.

We can generate a $(n-1)$-turple of coalescent time, $\{\tau_2, \tau_3, ..., \tau_{n-1}, \tau_n\}$, from the joint distribution (24) in the following way. First we set $N_{k=n} = N_1$, that is the size of the population where n samples are originally taken. Then, generate a random number τ_n according to the density distribution given by (23). Next, set $N_{k=n-1} = N_1\exp[-r\tau_n]$, and generate a random number τ_{n-1} according to the density distribution given by (23). The value of $N_{k=n-2}$ is then set to $N_{k=n-2} = N_1\exp[-r(\tau_n + \tau_{n-1})]$ and τ_{n-2} is generated, and so on.

The expected normalized AFS under this coalescent process can be described as

$$AFS_{sample}(i, t_1|n) = \frac{(n-i-1)!(i-1)!\sum_{k=2}^{n-i+1}k(k-1)\binom{n-k}{i-1}ET_k}{(n-1)!\sum_{k=2}^{n}kET_k} \qquad (1 \le i \le n-1), \tag{25}$$

where ET_k is the expectation of T_k that can be obtained from (24) (Griffiths and Tavaré 1998; Wakeley 2009). For the absolute number of mutations that exactly i individuals in a sample of size n have, $S_{sample}(i, \mu, t_1|n)$, it is not difficult to see that

$$S_{sample}(i, \mu, t_1|n) = b\mu\frac{(n-i-1)!(i-1)!\sum_{k=2}^{n-i+1}k(k-1)\binom{n-k}{i-1}ET_k}{(n-1)!} \qquad (1 \le i \le n-1), \tag{26}$$

holds. This is because ET_n contributes to $S_{sample}(1, \mu, t_1|n)$ in the form of $2b \cdot \mu \cdot (1/2) \cdot nET_n = b\mu nET_n$, where $2b$ is the backward rate of birth event per lineage, μ is the mutation rate, $(1/2)$ is the chance that the focal lineage receives a mutation at a single birth event, n is the total number

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

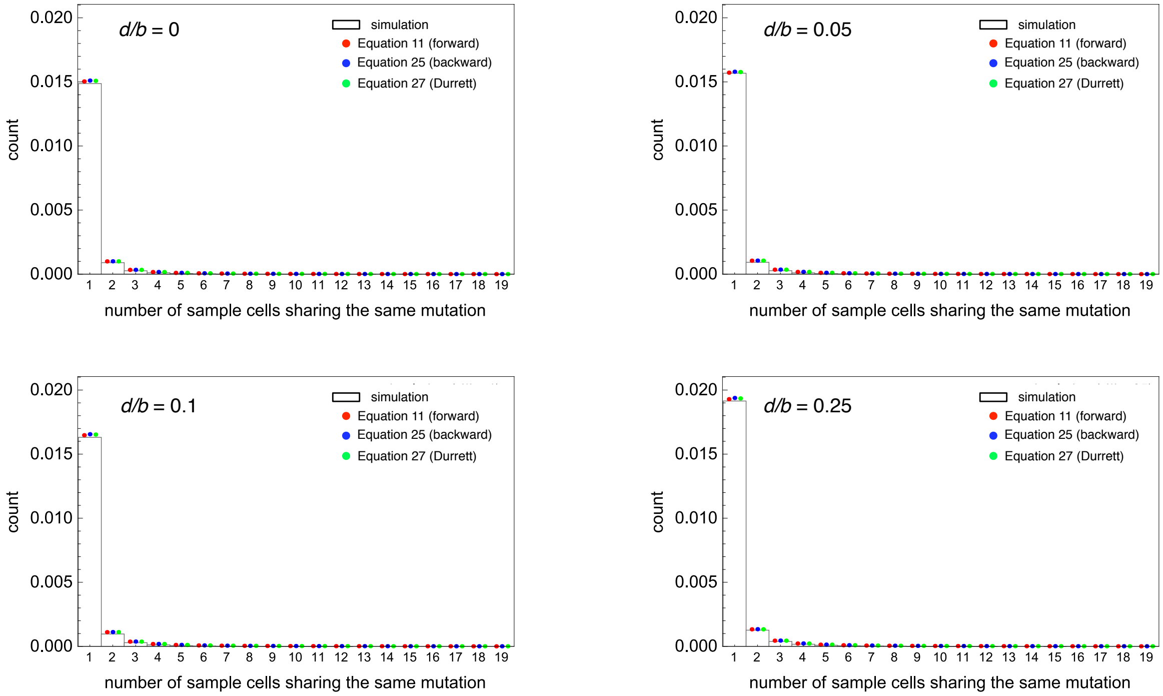

Figure 3 Population allele frequency spectra, $AFS(i, t_1)$, when $d/b = \{0, 0.05, 0.2, 0.25\}$. The theoretical results from our forward and backward treatments and Durrett's approximation (28) are compared with simulations. The simulation results are icentical to those used in Figure 2.

of independent lineages, and ET_n is the expected duration during which there are n independent lineages. As the expected coalescent time ET_k can be computed based on the numerical procedure provided above, it is straightforward to numerically calculate the sample AFS with Equations (25) and (26).

DISCUSSION

This article considers a model of a rapidly growing cancer cell population for exploring how mutations accumulate within the population. The expected AFS of derived mutations is obtained in a near exact form by both forward (branching theory) and backward (coalescent theory) treatments. Durrett (2013, 2015) obtained two approximate formulas to the sample-based AFS:

$$S_{sample}(i, \mu, t_1|n) = \begin{cases} \frac{n\mu}{1-(d/b)} \log[N_1\{1-(d/b)\}] & \text{if } i = 1 \\ \frac{n\mu}{1-(d/b)} \frac{1}{i(i-1)} & \text{if } 2 \le i \le n-1, \end{cases} \tag{27}$$

which was ultimately improved to be

$$S_{sample}(i, \mu, t_1|n) = \begin{cases} \frac{\mu}{1-(d/b)} \sum_{k=1}^{N_1\{1-(d/b)\}} \frac{n}{n+k} \frac{k}{n+k-1} & \text{if } i = 1 \\ \frac{n\mu}{1-(d/b)} \frac{1}{i(i-1)} & \text{if } 2 \le i \le n-1. \end{cases} \tag{28}$$

In Figure 3, the numerical results from our forward and backward derivations (i.e., Equations (11) and Equation (26) are compared with Durrett's two approximations (Equations (27) and (28)). There is nothing surprising that Equations (11) and (26) are in excellent agreement because they are in near

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

exact forms. In addition, we find Durrett's improved approximation (28) is extremely good, while the first approximation (27) would overestimate the singleton frequency (not shown).

The advantage of our near exact expressions over Durret's great approximation is that our theory would provide some mathematical intuitions, which could be useful for data analysis. (i) First, our forward expression (4) can be approximated to a very simple form for a large $r = b - d$:

$$S(i, \mu, t_1) \approx N_1\mu \cdot \frac{1}{i(i+1)}.$$

This means that $N_1\mu$ determines the absolute number of mutations and the relative frequency is converged to $\frac{1}{i(i+1)}$ with $r \to \infty$, which is independent of the growth rate. Provided that the growth rate of a typical cancer cell population is very large, AFS may not be very informative to estimate the growth rate. Rather, $N_1\mu$ may be more informative biologically because the mutation rate (μ) may be easily estimated if N_1 is given. It may not be very difficult to obtain a rough estimate of N_1 from the size of tumor. One might think that this implication seems odd: What if a tumor has grown from $N_0 = 1$ to $N_1 = 10^{10}$ in an hour or so? An hour could be too short to accumulate mutations. To address this question, we should note that how short the time is taken, it has to have involved at least $N_1 - N_0$ cell divisions and that the mutation rate is defined per cell division, not per time unit.

(ii) Second, our backward expression for the coalescent time is

$$\frac{2be^{r\tau_2}}{N} \exp\left[-\frac{2b}{rN}(e^{r\tau_2} - 1)\right]$$

(identical to Equation (22)), which is in a similar form to that under the standard coalescent in organismal population genetics:

$$\frac{e^{r\tau_2}}{N} \exp\left[-\frac{1}{rN}(e^{r\tau_2} - 1)\right] \tag{29}$$

(Slatkin and Hudson 1991). The difference between those two expressions can easily be explained; the factor 2 in the former equation reflects the fact that the our model assumes overlapping generation, while Slatkin and Hudson (1991) did not (e.g., Wakeley 2009). After neglecting this factor 2, these two equations are completely equivalent when the birth rate of a cell is $b = 1$ in our model. By comparing these two equations, the expression of Slatkin and Hudson (1991) for organismal population genetics is a special case of our expression. In other words, the well-established backward theory of organismal population genetics can be directly used to a cancer cell population by introducing a scale factor b that determines the relative rate of coalescent and population shrinkage (in backward).

One may think from our formulas of coalescent time (22) that the absolute values of b and $r = b - d$ jointly specifies the process. This is indeed true if we are interested in the absolute length of waiting time until coalescence. On one hand, if only allele frequency spectrum is of interest, those absolute values are much less important. Rather, the ratio of d to b, namely d/b, is a crucial determinant of the spectrum, as is obvious in Equation (4), which explicitly tells us that it is the case because it depends on b and d only through d/b. It may be difficult to see this fact in our backward formula (e.g., (22)), but if we rescale backward time and introduce a new timescale τ' by $\tau' = b\tau$, then Equation (20), for example, changes to

$$\frac{\mathrm{d}P_2'(\tau')}{\mathrm{d}\tau'} = -\rho_{2,\tau'}P_2'(\tau') = -\frac{2}{Ne^{-(r/b)\tau'}}P_2'(\tau'), \tag{30}$$

which depends on b and d only through $r/b = 1 - (d/b)$ and therefore only through d/b. This intuitively makes sense because in our cancer model, mutation occurs only at birth events, so the absolute waiting time until a birth event occurs is irrelevant when we focus on AFS of a population/sample.

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(iii) Third, our expressions are based on solid derivations. Therefore, the basic logic behind our derivations can be applied to more complex growth pattern as long as $b \gg d$. It can be considered that the growth of a cancer cell population may not be necessarily exponential. Driver mutations could increase the growth rate, while the growth process may slow down if the availability of resources such as space, oxygen, and other nutrients is limited. Such change of the growth curve may be incorporated by replacing Equation (1), which will be involved in the integration in (4) in the forward treatment and in the rate of coalescent specified by (18) in the backward treatment.

In summary, assuming an exponentially growing cancer cell population, we obtained near exact expressions of AFS (only continuous approximation is involved) from both forward and backward treatments. The former is in a simpler form and enhance our intuitive understanding of the process, while the latter provides a theoretical framework for coalescent (backward) simulations of a cancer cell population. Our theoretical results show that the difference from organismal population genetics is mainly in the coalescent time scale and the mutation rate is defined per cell division, not per time unit (e.g., generation). Except for these two factors, the basic logic are very similar between organismal and cancer population genetics. Therefore, a number of well established theories of organismal population genetics could be translated to cancer population genetics with simple modifications.

Acknowledgements

This work is in part supported by grants from the Japan Society for the Promotion of Science (JSPS) to HO and HI.

Literature Cited

Adams, A. M. and R. R. Hudson, 2004 Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics **168**: 1699–1712.

Bailey, N. T., 1964 *The Elements of Stochastic Processes with Applications to the Natural Sciences*. John Wiley & Sons, New York.

Bhaskar, A. and Y. S. Song, 2009 Multi-locus match probability in a finite population: a fundamental difference between the Moran and Wright–Fisher models. Bioinformatics **25**: i187–i195.

Bhaskar, A., Y. R. Wang, and Y. S. Song, 2015 Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Res. **25**: 268–279.

Crow, J. F. and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Dexter, D. L., H. M. Kowalski, B. A. Blazar, Z. Fligiel, R. Vogel, and G. H. Heppner, 1978 Heterogeneity of tumor cells from a single mouse mammary tumor. Cancer Res. **38**: 3174–3181.

Donnelly, P., 1996 Interpreting genetic variability: the effects of shared evolutionary history. Variation in the human genome pp. 25–50.

Durrett, R., 2013 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**: 230.

Durrett, R., 2015 Branching process models of cancer. In *Branching Process Models of Cancer*, pp. 1–63, Springer.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fidler, I. J., 1978 Tumor heterogeneity and the biology of cancer invasion and metastasis. Cancer Research **38**: 2651–2660.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

Fu, Y.-X., 1995 Statistical properties of segregating sites. Theor. Pop. Biol. **48**: 172–197.

Gao, F. and A. Keinan, 2016 Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. Genetics **202**: 235–245.

Garraway, L. A. and E. S. Lander, 2013 Lessons from the cancer genome. Cell **153**: 17–37.

Griffiths, R. and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Stochastic Models **14**: 273–295.

Griffiths, R. C. and S. Tavaré, 1994 Ancestral inference in population genetics. Stat. Science **9**: 307–319.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS Genet. **5**: e1000695.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. **23**: 183–201.

Innan, H. and W. Stephan, 2001 Selection intensity against deleterious mutations in rna secondary structures and rate of compensatory nucleotide substitutions. Genetics **159**: 389–399.

Iwasa, Y., F. Michor, and M. A. Nowak, 2004 Stochastic tunnels in evolutionary dynamics. Genetics **166**: 1571–1579.

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

Kingman, J. F. C., 1982 The coalescent. Stochast. Proc. Appl. **13**: 235–248.

Knudson, A. G., 1971 Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. USA **68**: 820–823.

Komarova, N. L., A. Sengupta, and M. A. Nowak, 2003 Mutation–selection networks of cancer initiation: tumor suppressor genes and chromosomal instability. J. Theor. Biol. **223**: 433–450.

Marusyk, A., D. P. Tabassum, P. M. Altrock, V. Almendro, F. Michor, and K. Polyak, 2014 Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature **514**: 54–58.

Merlo, L. M., J. W. Pepper, B. J. Reid, and C. C. Maley, 2006 Cancer as an evolutionary and ecological process. Nat. Rev. Cancer **6**: 924–935.

Michor, F., Y. Iwasa, and M. A. Nowak, 2004 Dynamics of cancer progression. Nat. Rev. Cancer **4**: 197–205.

Moran, P. A. P., 1962 *The Statistical Processes of Evolutionary Theory*. Clarendon Press; Oxford University Press, Oxford.

Muller, H. J., 1950 Radiation damage to the genetic material. Am. Scientist **38**: 33.

Navin, N. E., 2015 The first five years of single-cell cancer genomics and beyond. Genome Res. **25**: 1499–1507.

Network, C. G. A. R. *et al.*, 2012 Comprehensive genomic characterization of squamous cell lung cancers. Nature **489**: 519–525.

Network, C. G. A. R. *et al.*, 2014 Comprehensive molecular characterization of gastric adenocarcinoma. Nature **513**: 202–209.

Network, T. C. G. A. R., 2008 Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature **455**: 1061–1068.

Nielsen, R. and M. Slatkin, 2013 *An introduction to population genetics: theory and applications*. Sinauer Associates Sunderland, MA.

Nowell, P. C., 1976 The clonal evolution of tumor cell populations. Science **194**: 23–28.

Polanski, A., A. Bobrowski, and M. Kimmel, 2003 A note on distributions of times to coalescence,

bioRxiv preprint doi: https://doi.org/10.1101/104158; this version posted January 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

under time-dependent population size. Theor. Popul. Biol. **63**: 33–40.

Polanski, A. and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics **165**: 427–436.

Sidow, A. and N. Spies, 2015 Concepts in solid tumor evolution. Trends Genet. **31**: 208–214.

Sjöblom, T., S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, *et al.*, 2006 The consensus coding sequences of human breast and colorectal cancers. Science **314**: 268–274.

Slatkin, M. and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**: 555–562.

Sottoriva, A., H. Kang, Z. Ma, T. A. Graham, M. P. Salomon, J. Zhao, P. Marjoram, K. Siegmund, M. F. Press, D. Shibata, *et al.*, 2015 A big bang model of human colorectal tumor growth. Nat. Genet. **47**: 209–216.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437–460.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring coalescence times from dna sequence data. Genetics **145**: 505–518.

Uchi, R., Y. Takahashi, A. Niida, T. Shimamura, H. Hirata, K. Sugimachi, G. Sawada, T. Iwaya, J. Kurashige, Y. Shinden, *et al.*, 2016 Integrated multiregional analysis proposing a new model of colorectal cancer evolution. PLoS Genet. **12**: e1005778.

Vogelstein, B., N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, 2013 Cancer genome landscapes. Science **339**: 1546–1558.

Waclaw, B., I. Bozic, M. E. Pittman, R. H. Hruban, B. Vogelstein, and M. A. Nowak, 2015 A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. Nature **525**: 261–264.

Wakeley, J., 2009 *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village.

Williams, M. J., B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva, 2016 Identification of neutral tumor evolution across cancer types. Nat. Genet. **48**: 238–244.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen, and C. D. Bustamante, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA **102**: 7882–7887.

Wood, L. D., D. W. Parsons, S. Jones, J. Lin, T. Sjöblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, *et al.*, 2007 The genomic landscapes of human breast and colorectal cancers. Science **318**: 1108–1113.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16**: 97–159.