Theoretical Population Biology 117 (2017) 43–50

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population

Hisashi Ohtsuki, Hideki Innan *

SOKENDAI, The Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan

ARTICLE INFO

Article history:
Received 9 May 2017
Available online 1 September 2017

Keywords:
Allele frequency spectrum
Cancer
Population genetics
Branching theory
Coalescent theory

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 provide a theoretical framework of the stochastic process in a cancer cell population and obtain 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 divisions 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 is 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.

1. Introduction

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; The Cancer Genome Atlas Research Network, 2008, 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). Even without such desirable data available, the frequencies of mutations in bulk-sequencing data are informative to infer the history of a tumor (Williams et al., 2016).

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 in providing 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 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

* Correspondence to: Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan.
E-mail address: innan_hideki@soken.ac.jp (H. Innan).

http://dx.doi.org/10.1016/j.tpb.2017.08.006
0040-5809/

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 reproduction 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. Fig. 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.

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.

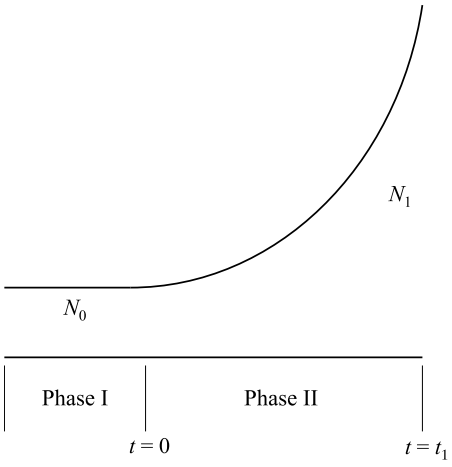

Fig. 1. Illustrating the model of the growth of a cancer cell population.

The behavior of mutations in an exponentially growing has been well studied since Luria and Delbrück (1943) who investigated the evolutionary process of resistance mutations in a bacterial population (see also Kessler and Levine, 2013). The model handles neutral mutations in an exponentially growing population, which will confer selective advantages after an environmental change (e.g., viral infection). Models with stochastic processes taken into account have been explored by Kessler and Levine (2013) and Antal and Krapivsky (2011). The Luria–Delbrück model thus provides the basis for exploring the behavior of driver and passenger mutations in a cancer cell population (e.g., Kansal et al., 2000; Haeno et al., 2007; Antal and Krapivsky, 2011; Durrett et al., 2011; Foo and Leder, 2013; Bozic et al., 2016). Most of these works focus on the number of mutations per cell, the evolutionary "waves" of driver mutations or more complicated tree structure (but see Durrett, 2013, 2015), which may not be straightforward to apply cancer genomic data, especially when single cell-based sequences are not available.

To be more applicable to recent cancer genomic data, we here ask how the well established theory of organismal population genetics can be applied to Phase II assuming 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 the 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 (Nielsen, 2000; 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 first present some derivations for the expected AFS of derived mutations in the final tumor (at t_1 in Fig. 1), by considering the process forward in time, following the previous theories cited above. To be simple enough to apply to cancer genome data and to compare the result with that of organismal population genetics, we use the assumption of infinitely many sites (Kimura, 1969), so that our model is a special case of Kessler and Levine (2013) and Antal and Krapivsky (2011). We have here obtained analytical expressions of the expected AFS in a near exact form (only continuous approximation is involved), which agrees with the solution of Kessler and Levine (2013) at the very large population limit. Also, our results are numerically in agreement with approximate formulas, which were rather intuitively obtained by Durrett (2013, 2015).

Second, we consider how this process can be described backward in time, to compare with the commonly used coalescent theory in organismal population genetics (e.g., Wakeley, 2009). Through the establishment of the modern theory of molecular population genetics, a number of classic theoretical results (mostly by diffusion theory Crow and Kimura, 1970; Ewens, 1979) have been re-obtained under the coalescent theory (e.g., Tajima, 1990a,b). It is usually the case that the agreement between the forward (diffusion theory) and backward (coalescent theory) is excellent; the difference is on the order of $1/N$, due to what kind of approximation was involved (e.g., diffusion approximation vs. continuous approximation of coalescent time). This article demonstrates that this also applies to the case of a cancer cell population. We obtained the density distribution of the coalescent time, from which the expected AFS was derived. Indeed, the AFS obtained by both forward (branching theory) and backward (coalescent theory) treatments are in a very good agreement. The density distribution of the coalescent time would be very useful for coalescent simulations, which allow one to generate a number of random realizations of the genealogy of cancer cells with less computational load. Thus, this work presents a solid population genetic theoretical framework of a cancer cell population, in which more complicated models could be explored.

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; The Cancer Genome Atlas Research Network, 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 established and most mutations occurring in the following growth phase may be passengers (Ling et al., 2015; Uchi et al., 2016; Sottoriva et al., 2015; Bozic et al., 2016), 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.

2. Model

Our model (Fig. 1) considers an exponentially growing population starting with N_0 asexually reproductive cells ($N_0 \geq 1$). 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.

It should be important to notice that the setting of time unit is arbitrary throughout this work. In other words, the rates of birth and death can be defined in any time unit. Therefore, the following theory hold with any measure of time such as hours and days. On one hand, the mutation rate is not given per time unit but per cell division, because mutation is involved only when a cell division occurs.

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, Eq. (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 (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

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

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 $i \geq 1$:

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

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. Eq. (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., $i = (1, 2, 3, \ldots, N_1)$) as

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

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

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

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

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

and

$$
AFS(i,t_1) = \frac{\frac{1}{i(i+1)}}{\sum_{i'=1}^{N_1}\frac{1}{i'(i'+1)}} \overset{N_1\to\infty}{\longrightarrow} \frac{1}{i(i+1)}, \qquad (8)
$$

which agrees with the solution of Kessler and Levine (2013) at the very large population limit. At this limit, it is interesting to note that $S(i, \mu, t_1)$ is independent of b or d.

While knowing that our assumption of $b \gg d$ is strongly violated, we may consider the opposite extreme, $b \sim d$. If we formally proceed our calculation by taking the limit $b \to d$, we obtain

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

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

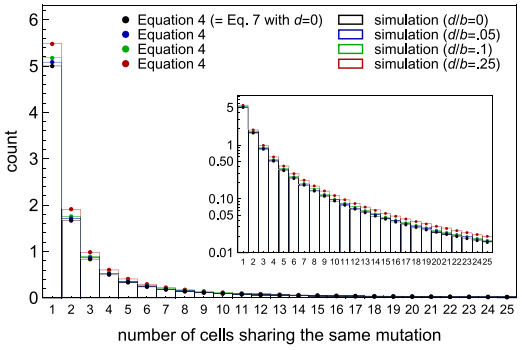

Fig. 2. Population allele frequency spectra, $AFS(i, t_1)$, when $d/b = \{0, 0.05, 0.1, 0.25\}$. The theoretical results from Eqs. (4) and (7) are compared with simulations. Log-scaled spectra are shown in the inner panel. Forward simulations were performed with $N_0 = 10$, $N_1 = 10^5$, $\mu = 10^{-4}$, $b = 4$, and $d = \{0, 0.2, 0.4, 1\}$. It should be noted that Eq. (4) with $d = 0$ is identical to Eq. (7).

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

where $\gamma \equiv 0.577215\cdots$ is 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 $d = \{0, 0.2, 0.4, 1\}$. Each simulation run was performed such that all cells randomly underwent cell division and death with probabilities b and d. As we did not assume an exponential growth of the cell population, simulation was terminated when N first hit N_1. Mutation was introduced at rare μ per cell division such that only one of the daughter cell received a novel mutation. Fig. 2 shows the average spectra (up to $i = 25$) over 10^5 simulation runs. Theoretical results based on Eq. (4) are shown in closed circles. It is demonstrated that Eq. (4) is in excellent agreement with the simulation results (colored open boxed) for all four cases. We further compare the values computed by Eq. (7), which is a simple approximation to Eq. (4) when $b \gg d$. We find that Eqs. (4) and (7) produce almost identical numerical values, which are indistinguishable in Fig. 2, indicating that the simple approximation works very well when $d = 0$. Furthermore, the simple form (Eq. (7)) may be a useful approximation for a sufficiently small d because the spectra for $d/b = 0$, 0.05, and 0.1 are quite similar.

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 $(1 \le i \le n)$ cells in a sample of size n is given by

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

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

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

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

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

and if fixed mutations are ignored

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

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 sample of cells from a tumor, and obtain essentially the same theoretical results as those from the forward treatment (i.e., Eqs. (11)–(14)).

Let us consider a pair of random (different) cells from the final tumor with N cells, where N is assumed to be very large. 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, which is defined such that the present time is set to $\tau = 0$ and τ increases as the process goes backward in time.

Let T_2 be the time for a pair of cell lineages to coalesce to their common ancestor. To obtain the pdf of T_2, we first consider the process from $\tau = 0$ to $\tau = \Delta\tau$. We assume $\Delta\tau$ is an infinitesimally small time interval, such that the number of birth and death events in the entire cell population is at most one. We first consider the population size at time $\tau = \Delta\tau$ conditioned on that the population size at $\tau = 0$ is N, $P(N_{\tau=\Delta\tau} = N-1|N_{\tau=0} = N)$, which is given by

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

where the probabilities $P(N_{\Delta\tau} = N-1)$ and $P(N_0 = N)$ can be easily calculated based on the forward process. Note that, for a large N, it is expected that the difference between these two probabilities is negligibly small (i.e., at most on the order of $\Delta\tau$). Therefore, the leading term of (15) is given by $b(N-1)\Delta\tau$, which represents the probability that one birth event occurred in the entire population in this time interval, $\Delta\tau$.

We next consider the coalescent rate in this time interval. Because a birth event causes the coalescence of a particular pair of lineages with probability

$$\frac{2}{N}\frac{1}{N-1}, \quad (16)$$

the rate of coalescence is, up to the first order of $\Delta\tau$, given by

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

It is obvious that these arguments on the coalescent rate holds at any time (τ), and by taking 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

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

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

$$\rho_{2,\tau} = \frac{2}{N}, \quad (19)$$

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

With these results, it is quite straightforward to obtain the pdf of T_2. Let $P_2(\tau)$ be the probability that the coalescence between a particular pair of lineages have not occurred yet by time τ. Then, $P_2(\tau)$ satisfies the following differential equation,

$$\frac{dP_2(\tau)}{d\tau} = -\rho_{2,\tau}P_2(\tau) = -\frac{2b}{Ne^{-r\tau}}P_2(\tau), \quad (20)$$

with $P_2(0) = 1$ as an initial condition. The solution is given by a double exponential function:

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

Therefore, the probability density function of coalescent time T_2, $F(T_2 = \tau_2)$, is given by

$$\begin{aligned} F(T_2 = \tau_2) &= -\frac{dP_2(\tau_2)}{d\tau_2} \\ &= \frac{2b}{N}\exp\left[-\frac{2b}{rN}(e^{r\tau_2} - 1) + r\tau_2\right] \\ &= \frac{2be^{r\tau_2}}{N}\exp\left[-\frac{2b}{rN}(e^{r\tau_2} - 1)\right]. \end{aligned} \quad (22)$$

In order to consider the coalescent process of n sampled cells up to their MRCA, we need to generalize (23) to the cases with k up to n. Following the above logic, we can derive the pdf of time interval during which $k(>2)$ lineages coalesce into $k-1$ lineages, which is denoted by T_k. Then, we have

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

Then, the joint pdf of $\{T_2, T_3, \ldots, T_{n-1}, T_n\}$ is given by

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

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 $\sum_{\ell=j+1}^{n}\tau_\ell$ (see Slatkin and Hudson, 1991). For a coalescent simulation, we can generate a $(n-1)$-tuple of coalescent time, $\{\tau_2, \tau_3, \ldots, \tau_{n-1}, \tau_n\}$, from the joint distribution (24) in the following way. First we set $N_{k=n} = N_1$, that is the size of the population from which 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.

Using the formula of Griffiths and Tavaré (1998), it is straightforward to obtain the expected normalized AFS under this coalescent process:

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

where ET_k is the expectation of T_k that can be obtained from (24). 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

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

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 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 Eqs. (25) and (26).

In Fig. 3, the numerical results from our forward and backward derivations (i.e., Eqs. (11) and (26)) are compared with Durrett's two approximations (Eqs. (27) and (28)). 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

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

Fig. 3 demonstrates that Eqs. (11) and (26) are in excellent agreement, which is not very surprising according to our experience in organismal population genetic theory. In addition, we find that Durrett's improved approximation (28) is extremely good, while the first approximation (27) would overestimate the singleton frequency (not shown).

3. Discussion

This article considers a model of a rapidly growing cancer cell population for exploring how mutations accumulate within the population. We are particularly interested in how the process can be described in the framework of forward and backward theory of population genetics. Our main results (i.e., (4) and (18)) demonstrate that both birth and death rates determine the evolutionary process. More specifically, the two rates specify the frequency of a mutation at the final tumor in the forward treatment and the coalescent rate in the backward treatment. We also provide several approximations when $b \gg d$, which provide some intuitive interpretation of our theoretical results.

For a large $r = b - d$, our forward expression (4) can be approximated to a very simple form:

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

This means that $N_1\mu$ determines the absolute number of mutations and the relative frequency is converged to $\frac{1}{i(i+1)}$ when r is very large. In such a case, AFS 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 the relative spectrum, the total number of observed mutations 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.

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 Eq. (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 Eq. (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 Eq. (20), for example, changes to

$$\frac{dP_2'(\tau')}{d\tau'} = -\rho_{2,\tau'}P_2'(\tau') = -\frac{2}{Ne^{-(r/b)\tau'}}P_2'(\tau'), \tag{30}$$

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.

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 described by a single exponential function. 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

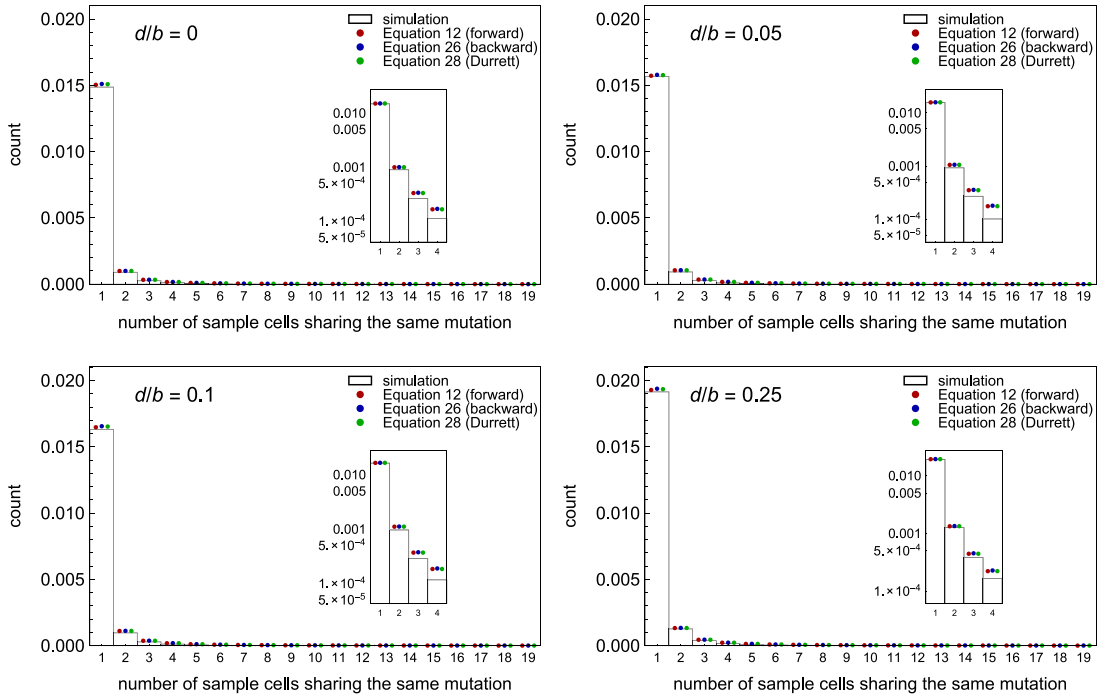

Fig. 3. Population allele frequency spectra, $AFS(i, t_1)$, when $d/b = \{0, 0.05, 0.1, 0.25\}$. Log-scaled spectra are shown in inner panels. The theoretical results from our forward and backward treatments and Durrett's approximation (28) are compared with simulations. The simulation results are identical to those used in Fig. 2.

be incorporated by replacing Eq. (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, this article demonstrates that the theoretical logics commonly used in organismal population genetics can be translated into 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 that is defined per cell division, not per time unit (e.g., generation). Except for these two factors, the basic logic is very similar between organismal and cancer population genetics. Therefore, a number of well established theories of organismal population genetics could be modified for cancer population genetics such as inferring demographic history and selection (e.g., Nielsen and Slatkin, 2013).

One of the implications is that most mutations in a cancer cell population appear as singletons. Considering the current situation of single-cell sequencing, it may not be easy to identify singletons, but in the near future, the quality and quantity of single-cell sequence data will increase dramatically (Navin, 2015).

Acknowledgment

This work is in part supported by grants from the Japan Society for the Promotion of Science (JSPS) (grant no. #23114004) to HI.

References

Adams, A.M., Hudson, R.R., 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics 168, 1699–1712.
Antal, T., Krapivsky, P., 2011. Exact solution of a two-type branching process: models of tumor progression. J. Stat. Mech. Theory Exp. 2011 (08), P08018.
Bailey, N.T., 1964. The Elements of Stochastic Processes with Applications to the Natural Sciences. John Wiley & Sons, New York.
Bhaskar, A., Song, Y.S., 2009. Multi-locus match probability in a finite population: a fundamental difference between the Moran and Wright–Fisher models. Bioinformatics 25 (12), i187–i195.
Bhaskar, A., Wang, Y.R., Song, Y.S., 2015. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Res. 25 (2), 268–279.
Bozic, I., Gerold, J.M., Nowak, M.A., 2016. Quantifying clonal and subclonal passenger mutations in cancer evolution. PLoS Comput. Biol. 12 (2), e1004731.
Crow, J.F., Kimura, M., 1970. An Introduction to Population Genetics Theory. Harper & Row, New York.
Dexter, D.L., Kowalski, H.M., Blazar, B.A., Fligiel, Z., Vogel, R., Heppner, G.H., 1978. Heterogeneity of tumor cells from a single mouse mammary tumor. Cancer Res. 38 (10), 3174–3181.
Donnelly, P., 1996. Interpreting genetic variability: the effects of shared evolutionary history. In: Variation in the Human Genome. John Wiley Chichester, UK, pp. 25–50.
Durrett, R., 2013. Population genetics of neutral mutations in exponentially growing cancer cell populations. Ann. Appl. Probab. 23 (1), 230.
Durrett, R., 2015. Branching process models of cancer. In: Branching Process Models of Cancer. Springer, pp. 1–63.
Durrett, R., Foo, J., Leder, K., Mayberry, J., Michor, F., 2011. Intratumor heterogeneity in evolutionary models of tumor progression. Genetics 188 (2), 461–477.
Ewens, W.J., 1979. Mathematical Population Genetics. Springer-Verlag, Berlin.
Fidler, I.J., 1978. Tumor heterogeneity and the biology of cancer invasion and metastasis. Cancer Res. 38 (9), 2651–2660.
Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Oxford University Press, Oxford.
Foo, J., Leder, K., 2013. Dynamics of cancer recurrence. Annl. Appl. Probab. 23 (4), 1437–1468.
Fu, Y.-X., 1995. Statistical properties of segregating sites. Theor. Pop. Biol. 48, 172–197.
Gao, F., Keinan, A., 2016. Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. Genetics 202 (1), 235–245.
Garraway, L.A., Lander, E.S., 2013. Lessons from the cancer genome. Cell 153 (1), 17–37.
Griffiths, R., Tavaré, S., 1998. The age of a mutation in a general coalescent tree. Stoch. Models 14 (1–2), 273–295.
Griffiths, R.C., Tavaré, S., 1994. Ancestral inference in population genetics. Statist. Sci. 9, 307–319.
Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5 (10), e1000695.
Haeno, H., Iwasa, Y., Michor, F., 2007. The evolution of two mutations during clonal expansion. Genetics 177 (4), 2209–2221.
Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. 23, 183–201.
Innan, H., Stephan, W., 2001. Selection intensity against deleterious mutations in rna secondary structures and rate of compensatory nucleotide substitutions. Genetics 159 (1), 389–399.

Iwasa, Y., Michor, F., Nowak, M.A., 2004. Stochastic tunnels in evolutionary dynamics. Genetics 166 (3), 1571–1579.
Kansal, A., Torquato, S., Chiocca, E., Deisboeck, T., 2000. Emergence of a subpopulation in a computational model of tumor growth. J. Theoret. Biol. 207 (3), 431–441.
Kessler, D.A., Levine, H., 2013. Large population solution of the stochastic Luria–Delbrück evolution model. Proc. Natl. Acad. Sci. USA 110 (29), 11682–11687.
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. Stochastic Process. Appl. 13, 235–248.
Knudson, A.G., 1971. Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. USA 68 (4), 820–823.
Komarova, N.L., Sengupta, A., Nowak, M.A., 2003. Mutation–selection networks of cancer initiation: tumor suppressor genes and chromosomal instability. J. Theoret. Biol. 223 (4), 433–450.
Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., et al., 2015. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. Proc. Natl. Acad. Sci. USA 112 (47), E6496–E6505.
Luria, S.E., Delbrück, M., 1943. Mutation of bacteria from virus sensitivity to virus resistance. Genetics 28, 491–511.
Marusyk, A., Tabassum, D.P., Altrock, P.M., Almendro, V., Michor, F., Polyak, K., 2014. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature 514 (7520), 54–58.
Merlo, L.M., Pepper, J.W., Reid, B.J., Maley, C.C., 2006. Cancer as an evolutionary and ecological process. Nat. Rev. Cancer 6 (12), 924–935.
Michor, F., Iwasa, Y., Nowak, M.A., 2004. Dynamics of cancer progression. Nat. Rev. Cancer 4 (3), 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. Sci. 38 (1), 33.
Navin, N.E., 2015. The first five years of single-cell cancer genomics and beyond. Genome Res. 25 (10), 1499–1507.
Nielsen, R., 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154 (2), 931–942.
Nielsen, R., Slatkin, M., 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 (4260), 23–28.
Polanski, A., Bobrowski, A., Kimmel, M., 2003. A note on distributions of times to coalescence, under time-dependent population size. Theor. Popul. Biol. 63 (1), 33–40.
Polanski, A., Kimmel, M., 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165 (1), 427–436.
Sidow, A., Spies, N., 2015. Concepts in solid tumor evolution. Trends Genet. 31 (4), 208–214.
Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., et al., 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314 (5797), 268–274.
Slatkin, M., Hudson, R.R., 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129, 555–562.
Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., et al., 2015. A big bang model of human colorectal tumor growth. Nat. Genet. 47 (3), 209–216.
Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105, 437–460.
Tajima, F., 1990a. Relationship between DNA polymorphism and fixation time. Genetics 125, 447–454.
Tajima, F., 1990b. Relationship between migration and DNA polymorphism in a local population. Genetics 126, 231–234.
Tavaré, S., Balding, D.J., Griffiths, R.C., Donnelly, P., 1997. Inferring coalescence times from dna sequence data. Genetics 145 (2), 505–518.
The Cancer Genome Atlas Research Network, 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455 (7216), 1061–1068.
The Cancer Genome Atlas Research Network, 2012. Comprehensive genomic characterization of squamous cell lung cancers. Nature 489 (7417), 519–525.
The Cancer Genome Atlas Research Network, 2014. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513 (7517), 202–209.
Uchi, R., Takahashi, Y., Niida, A., Shimamura, T., Hirata, H., Sugimachi, K., Sawada, G., Iwaya, T., Kurashige, J., Shinden, Y., et al., 2016. Integrated multiregional analysis proposing a new model of colorectal cancer evolution. PLoS Genet. 12 (2), e1005778.
Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Kinzler, K.W., 2013. Cancer genome landscapes. Science 339 (6127), 1546–1558.
Waclaw, B., Bozic, I., Pittman, M.E., Hruban, R.H., Vogelstein, B., Nowak, M.A., 2015. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. Nature 525 (7568), 261–264.
Wakeley, J., 2009. Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village.
Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A., Sottoriva, A., 2016. Identification of neutral tumor evolution across cancer types. Nat. Genet. 48, 238–244.
Williamson, S.H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., Bustamante, C.D., 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA 102 (22), 7882–7887.
Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al., 2007. The genomic landscapes of human breast and colorectal cancers. Science 318 (5853), 1108–1113.
Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.