A New Formulation of Random Genetic Drift and Its Application to the Evolution of Cell Populations

Yuxin Chen,[1] Ding Tong,[1,2] and Chung-I Wu*[,1,3,4]

[1]State Key Laboratory of Bio-control, School of Life Sciences, Sun Yat-Sen University, Guangzhou, Guangdong, China

[2]Department of Biostatistics, School of Public Health, Yale University, New Haven, CT

[3]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[4]Department of Ecology and Evolution, University of Chicago, Chicago, IL

*Corresponding author: E-mail: ciwu@uchicago.edu.

Associate editor: John True

Abstract

Random genetic drift, or stochastic change in gene frequency, is a fundamental evolutionary force that is usually defined within the ideal Wright–Fisher (WF) population. However, as the theory is increasingly applied to populations that deviate strongly from the ideal model, a paradox of random drift has emerged. When drift is defined by the WF model, it becomes stronger as the population size, *N*, decreases. However, the intensity of competition decreases when *N* decreases and, hence, drift might become weaker. To resolve the paradox, we propose that random drift be defined by the variance of "individual output", V(*k*) [*k* being the progeny number of each individual with the mean of E(*k*)], rather than by the WF sampling. If the distribution of *k* is known for any population, its strength of drift relative to a WF population of the same size, *N*, can be calculated. Generally, E(*k*) and V(*k*) should be density dependent but their relationships are different with or without competition, leading to opposite predictions on the efficiency of random drift as *N* changes. We apply the "individual output" model to asexual cell populations that are either unregulated (such as tumors) or negatively density-dependent (e.g., bacteria). In such populations, the efficiency of drift could be as low as <10% of that in WF populations. Interestingly, when *N* is below the carrying capacity, random drift could in fact increase as *N* increases. Growing asexual populations, especially tumors, may therefore be genetically even more heterogeneous than the high diversity estimated by some conventional models.

Key words: **random genetic drift, population genetics, cancer evolution, Wright–Fisher model, density dependence.**

Significance Statement

Random genetic drift, or random evolutionary change, is a basic concept in population genetic theory. When the theory is applied to unconventional populations, many limitations have become apparent. We thus propose a new approach to random drift—at the level of individuals' progeny output in lieu of the conventional scheme. The new approach reveals a paradox of random drift. In the conventional view, drift becomes stronger when the population becomes smaller. However, from the ecological perspective, the reduced competition would make drift weaker. The paradox becomes most apparent in tumor cell populations but is far more general. The new approach to random drift, by resolving the paradox, broadens the application of evolutionary theory to medical phenomena.

Introduction

Random genetic drift (random drift or drift, for short) refers to stochastic changes in gene frequency that do not result from the influence of deterministic forces such as mutation, selection, or migration. Random drift is most commonly defined by the Wright–Fisher (WF) model (Crow and Kimura 1970; Ewens 2004; Hartl and Clark 2007), in which a population consists of N individuals. In the generation transition, the progeny of the N individuals would compete for "slots" in the new population of size N'. A neutral variant with frequency x will change by δx in one generation with $E(\delta x)=0$ and $V(\delta x)=x(1-x)/N'$, the latter being the mathematical definition of random drift (Crow and Kimura 1970). The results derived from the WF model are valid for other population genetic models (Moran 1958; Charlesworth and Charlesworth 2010) that share common assumptions about the mechanism of population regulation.

The significance of random drift in molecular evolution can be highlighted as follows. Let f designate the strength of random drift, often expressed as $1/N$ for haploid populations in the WF model. With u designating the mutation rate, the neutral nucleotide evolutionary rate is $N\,u\,f$ ($=u$, when $f=1/N$) and the expected neutral diversity, H, can be written as $(1-H)(f+2u)\sim f$ (Crow and Kimura 1970). Therefore, how random drift operates would critically determine the long-term molecular evolutionary changes as well as the short-term population genetic patterns. Despite its centrality in molecular evolution, random drift has been defined mainly within the WF model. We will show that random drift in nonWF populations may often operate very differently.

Article

In the WF model, populations are clearly delineated and "well-mixed" units such that sampling for the next generation is random. This may be referred to WF sampling under spatial panmixia. Most natural populations, however, cannot be so delineated; for example, when there is a cline of neutral variants that is wider than the migration distance. It would, therefore, be desirable to define random drift without the rigid definition of spatially panmictic populations.

As population-genetic theories are being extended to unconventional populations, generality of the classical theories is being tested to the limit. For example, tumors are populations of cells that expand from a single progenitor. While their evolution has been increasingly subjected to population genetic analyzes (Maley et al. 2006; Gerlinger 2012; Marusyk et al. 2012; Ling et al. 2015; Sottoriva et al. 2015; Williams et al. 2016), they have distinct spatial structures, growth patterns and mechanisms of population regulation(Ling et al. 2015; Sottoriva et al. 2015). Indeed, there is a disagreement over whether random drift becomes stronger or weaker as tumors grow (Gerlinger et al. 2014; Wu et al. 2016), referred to as the paradox of random drift.

The Paradox of Random Drift

When population size is not constant, strength of drift may depend on both how large the population size N is and how N is regulated. In the WF model, random drift becomes weaker as N increases. However, if we see an increase in drift as a manifestation of competition for a limited number of "slots" in the environment, an opposite conclusion may be reached. This is because competition would become keener as the population density increases (Lande et al. 2003). Therefore, with density-dependence, drift may become stronger in larger populations. The paradox of random drift stems from divergent assumptions about mechanisms of population size control.

To see the underlying causes of the paradox, we consider populations that follow logistic growth:

$$\frac{dN_t}{dt} = rN_t\frac{C - N_t}{C} \quad (1)$$

where C is the limit of the population size, or the carrying capacity. When $N_t \sim C$, the population is producing far more progeny than the carrying capacity can accommodate. As a result, the progeny would be competing for the C "slots" and success is stochastic. However, when $N \ll C$ but increases to N' in the next generation, the WF model still assumes that the N individuals' progeny would compete for the N' "slots", as if N' is the carrying capacity. Since these progenies should be competing for the C slots, the WF model may severely overestimate competition and, hence, drift (see Discussion and fig. 1).

The Formulation of Random Drift by "Individual Output"

An alternative definition of random drift is the variation in individual progeny production. Kimura and Crow (1963) and others (Wright 1969; Gillespie 1975; Eldon and Wakeley 2006) have suggested the enumeration of the progeny number, k, of each individual. In the WF model where N is not very small (i.e., > 50), the probability a gene will not be chosen ($k = 0$) between generation t-1 and t is $(1 - 1/N_{t-1})^{N_t} = e^{-\lambda}$ where $\lambda = N_t/N_{t-1} = E(k)$ is the mean growth rate. More generally, k follows the Poisson distribution with mean E(k) and variance V(k) = E(k) when N is > 50. Note that V(k)/E(k) is density-independent.

Since k can be measured (sometimes with accuracy as in human populations), observations often show its distribution to be nonPoisson. In fact, males and females have the same mean number of progeny but they would likely have different variances in progeny production (Hartl and Clark 2007). Therefore, random sampling from a WF population may not be a suitable model for random drift. In this study, we propose that random drift be redefined by the distribution of k, equivalent to a branching process (Billingsley 1995) where each individual would produce k progeny with the probability, P_k. The probabilities may not follow well-defined statistical distributions but can often be empirically measured. In this nonWF model, the population is simply the sum of "individual outputs" and random drift would be determined by the variation in such output. As long as N is given, random drift can be quantified and the population is merely a collection of individuals in a spatial area. For simplicity, we shall only consider populations with asexual reproduction.

Given any arbitrary distribution of progeny number (k), an equivalent WF population governed by random drift of the same strength can be identified. In a WF population with two allelic forms with frequency x and 1-x, the change in x, *designated as* δx, can be approximated by the diffusion process. The Kolmogorov diffusion equations (Crow and Kimura 1970) only require that we know the mean and variance of δx, E(δx) and V(δx). If variation is neutral, E(δx)=0 and V(δx) would solely determine the speed of frequency change due to random drift.

Let $V^*(\delta x) = x(1-x)/N$ designate specifically such variance in a WF population of size N. Also, let V(δx) be such variance in nonWF populations. We now define the relative efficiency of drift as

$$e_f = \frac{V(\delta x)}{V^*(\delta x)} \quad (2)$$

where e_f is the strength of drift in the branching process relative to that in a WF population of the same size N. It is convenient to define the effective population size (N_e) that generates V(δx) by

$$N_e = \frac{x(1-x)}{V(\delta x)} \quad (3)$$

In other words, a nonWF population experiences random drift as if it is a WF population with size N_e. Because the population size is the sum of N_t independent random variables (k), V(δx) is ultimately determined by the distribution of k. As long as V(k) > 0 random drift would operate.

V(δx) in the Individual Output Model

To obtain V(δx) when the distribution of k is independent of N, we consider i and j copies of allele A and a, respectively, with $N = i + j$. Let x be the frequency of allele A in the population and $x = i/N = i/(i + j)$.

In the next generation, each individual leaves k copies with the probability P_k, giving rise to new quantities i', j', N' and x'. Here,

$$i' = \sum_1^i k_l,$$

$$j' = \sum_{i+1}^N k_l,$$

$$N' = i' + j' \text{ and } x' = \frac{i'}{N'} = \frac{i'}{i' + j'}$$

$$\mathrm{E}(x') = \exp\left(\frac{i'}{i' + j'}\right) = \sum_{i,j}\left[\left(\frac{i'}{i' + j'}\right)\mathrm{P}(i')\mathrm{P}(j')\right] = x$$

$$\mathrm{V}(x') = \mathrm{V}\left(\frac{i'}{i' + j'}\right) = \mathrm{E}\left(x'^2\right) - [\mathrm{E}(x')]^2 \quad (4)$$

where $\mathrm{E}\left(x'^2\right) = \sum_{i,j}\left[\left(\frac{i'}{i'+j'}\right)^2 \mathrm{P}(i')\mathrm{P}(j')\right]$

Also, $\mathrm{V}(\delta x) = \mathrm{V}(x' - x) = \mathrm{V}(x')$.

P(i') and P(j') can be easily obtained by numerical means (see Supplement II, Supplementary Material online).

An approximation of equation (4) has been given as

$$\mathrm{V}\left(x'\right) = \frac{\mathrm{V}(k)}{\mathrm{E}^2(k)}\frac{x(1-x)}{N} \quad (5)$$

by Kimura and Crow (1963) if we set $\alpha = 1$ and $N_e = 1/2N$ (since cell populations are asexual) in their equation (18). In this approximation, $\mathrm{E}(i' + j') = \mathrm{E}(N')$ is substituted for $i' + j' = N'$ (see below). Equation (5) can also be obtained by extending equation (4) using the approximation formula for the variance of the ratio of two random variables (Elandt-Johnson and Johnson 1980; Kendall et al. 1994).

By substituting E(N') for N',equation (5) assumes V(N') = 0 as in the WF model. In that case, the distribution of k should be Poisson with V(k) = E(k). Therefore, equation (5), which permits V(k) ≠ E(k), cannot be used in the same model that assumes V(N') = 0. The two conditions mutually contradict. Since Kimura and Crow (1963), population genetic theories have largely kept the assumption of V(N') = 0 of the WF model. In this study, we suggest abandoning V(N') = 0, thus allowing V(k) ≠ E(k) but necessitating the regulation of N' (see eqs. 6 and 7 below). Further details are presented in Supplement, Supplementary Material online.

N_t—*Changes in* N *as a Function of Time*

So far, the distribution of k is independent of N. Since the progeny production is almost always density-dependent, we first allow N, or N_t, to change with time. By the branching process, the mean and variance of N_t are:

$$\mathrm{E}(N_t) = N_0\mathrm{E}(k)^t \quad (6)$$

$$\mathrm{V}(N_t) = \begin{cases} N_0\mathrm{V}(k)t, & \text{if } \mathrm{E}(k) = 1 \\ N_0\mathrm{V}(k)\mathrm{E}(k)^{t-1}\dfrac{\mathrm{E}(k)^t - 1}{\mathrm{E}(k) - 1}, & \text{if } E(k) > 1 \end{cases} \quad (7)$$

It is clear from the equations that without a mechanism of density regulation many populations will eventually reach $N_t = 0$ or infinity. The WF model assumes that populations are regulated and are near the carrying capacity, C. Therefore, when the WF model permits N_t to vary, it assumes that C is changing and N keeps close track of C. Violation of this assumption would lead to incorrect estimates of random drift.

Given the variation in N_t for any t, we should note that equation (5) is only an approximation. In this approximation, E(N_t) is substituted for N_t with the assumption that N_t = E(N_t) and V(N_t) = 0. As V(N_t) can be very large (see eq. 7), the accuracy of equation (5) will need to be evaluated against the numerical calculation of equation (4). Such comparisons are carried out for a specific application of the individual-output model in supplementary figure S1, Supplementary Material online.

Density-Dependent Progeny Production (k)—An Example

The distribution of k should be a function of population density. At a higher density, each individual is expected to have fewer progeny and E(k) should be smaller. While density-dependent population growth is an active field of research, we shall consider the simplest model possible, which provides a different perspective on random drift = vis-à-vis that of the WF model. Consider the case where each individual can produce a maximum of κ progeny. The population size is N and the population's reproductive potential is $N\kappa$. Given C, the expected survival probability of each progeny would be w where

$$w = \frac{C}{N\kappa} \quad (8)$$

when $N\kappa \geq C$. When $N\kappa < C$, $w = 1$. The mean and variance of progeny production would then be binomially distributed with

$$\mathrm{E}(k) = \kappa w$$

$$\mathrm{V}(k) = \kappa w(1 - w) \quad (9)$$

Figure 1 shows that, as N increases, E(k) would decrease until E(k) = 1. In contrast, V(k) is very low when N is small due to the reduction in competition. (If $N\kappa < C$, all progeny survive and V(k) = 0.) V(k) increases rapidly until $N = 2C/\kappa$ and then declines to $1 - 1/\kappa$ when $N \sim C$. The pattern is opposite of the dynamics in WF populations where V(k) = E(k) for

all N's. The strength of drift [i.e., variance normalized by $x(1-x)$] is

$$\frac{V(x')}{x(1-x)} = \frac{V(k)}{E^2(k)N} = \frac{1-\kappa}{\kappa w N} = \frac{1}{C} - \frac{1}{N\kappa} \quad (10)$$

when $N\kappa > C$. Note that $V(x')/[x(1-x)]$ increases as N increases in this simple model whereas the equivalent term in the WF model is $1/N$ which decreases as N increases.

Random Drift in Local Populations

An advantage in tracking drift by $V(k)$ is the ability to formulate local drift without the need to delineate a WF population. The approach is particularly useful when populations are spatially structured and one wishes to measure random drift in local populations. In that case, one would only have to know the number of individuals being tracked. Since k is sensitive to local density, rather than the total size of the species, the individual-output approach should be more accurate in tracking local frequency changes. Random drift in the larger meta-population can then be considered reiterations of many smaller local populations.

Application of the Individual-Output Model to Cell Populations

The "individual output" model is applicable to any population if the distribution of k is known. In this study, we focus on cell populations of yeast, bacteria, and tumors. At the generation transition, each cell assumes any of the following three states with corresponding probabilities: P_0 ($k = 0$, cell death), P_1 ($k = 1$, no change) or P_2 ($k = 2$, cell division); $P_0 + P_1 + P_2 = 1$. We first assume that the three probabilities do not change as N changes but will consider density-dependent regulation later.

All probability combinations of (P_0, P_1, P_2) are evaluated at intervals of 0.05 in figure 2. For each distribution, $V(\delta x)$, $V^*(\delta x)$ and e_f [$= V(\delta x)/V^*(\delta x)$] are calculated. Figure 2 shows that random drift under the WF model is stronger than most (P_0, P_1, P_2) combinations, as indicated by the prevalence of blue spots. Notably, when the population is growing (the right half of the ternary plot), e_f is always < 1; hence, $N_e > N$. This property is quite unusual as we typically think of $N_e \ll N$. The reason for $N_e > N$ is the limited range of $V(k)$ when k is restricted to 0, 1, and 2. Imagine that, at a regular time interval, a cell may have *0*, *1*, or *n* (≥ 2) progeny with the probabilities P_0, P_1 and P_n; $P_0 + P_1 + P_n = 1$. The inset of figure 2 presents the results when $n = 5$. With the greatly increased $V(k)$'s, e_f would be mostly >1 (red dots), suggesting that the strength of drift is strongly dependent on $V(k)$.

The rate of population growth in figure 2 is presented as the distance to the midline which indicates $P_0 = P_2$ (birth = death) and no growth. The salient feature of figure 2 is that e_f is mainly governed by P_0, as shown by the double-headed blue arrow. When P_0 is close to 0 (i.e., little cell death), drift efficiency is often $<20\%$ of the strength in a WF population of equal size (fig. 2). Drift efficiency is affected by the growth of population *indirectly* because rapidly growing populations usually have smaller P_0 values. Populations at the two apexes with $P_1 = 1$ or $P_2 = 1$ have no random drift.

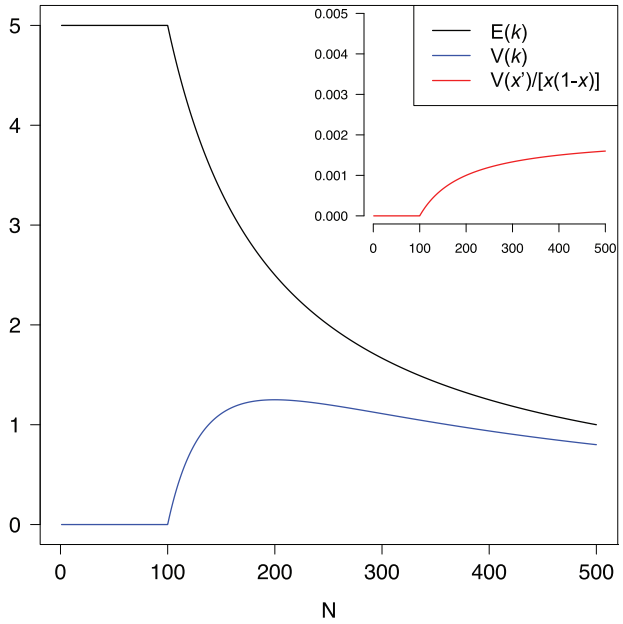

Fig. 1. $E(k)$ and $V(k)$ as a function of N when N grows logistically. $E(k)$ (black line) and $V(k)$ (blue line) are the mean and variance of k as given in equation (9) with $C = 500$ and $\kappa = 5$. $E(k)$ decreases as N increases according to the logistic equation. $V(k)$, in contrast, rises quickly and then decreases slowly as the intensity of competition changes. Inset—the strength of random drift, $V(x')/x(1-x)$, according to equation (10) is given by the red line. The drift strength increases as N increases in this model. This trend is opposite of the conventional thinking due to the density-dependent regulation of $V(k)/E(k)$, which is 1 in the standard model.

Random drift in the lower right corner, where P_2 approaches one, is particularly inefficient. When P_2 approaches 1, N is growing rapidly and drift is weak even in WF populations. In the branching process e_f in this corner is weaker by $>90\%$ than the weak drift in WF populations. In contrast, drift in the lower left corner is stronger than in WF populations when $P_0 \gg P_2$. Although reducing P_0 would reduce drift and reduce the loss of genetic diversity, less cell death would also permit population growth with fewer rounds of cell divisions. As the generation of new mutations depends on active cell divisions, reducing P_0 would retard mutation accumulation. In actual tumors of equal size when mutations are taken into account, smaller P_0 may in fact give rise to higher genetic diversity(Tao et al. 2011; Ling et al. 2015; Sottoriva et al. 2015; Williams et al. 2016).

Lastly, we verify the accuracy of the calculated N_e's in predicting the speed of drift. For example, a population with $e_f = 0.1$ should be equivalent to a WF population that is 10 times larger ($1/e_f$) with respect to the speed of random drift. Simulation results presented in figure 3 show three cases of decline in heterozygosity [$= 2x(1-x)$] as a measure of random drift. The three sets of probabilities all permit the population to grow by 1% per generation ($P_2 - P_0 = 0.01$) but the e_f is set at 0.10, 0.50 and 0.90, respectively, for $P_1 = 0.9$, 0.5 and 0.1. In both the simulations and predictions, random drift is slower than in the

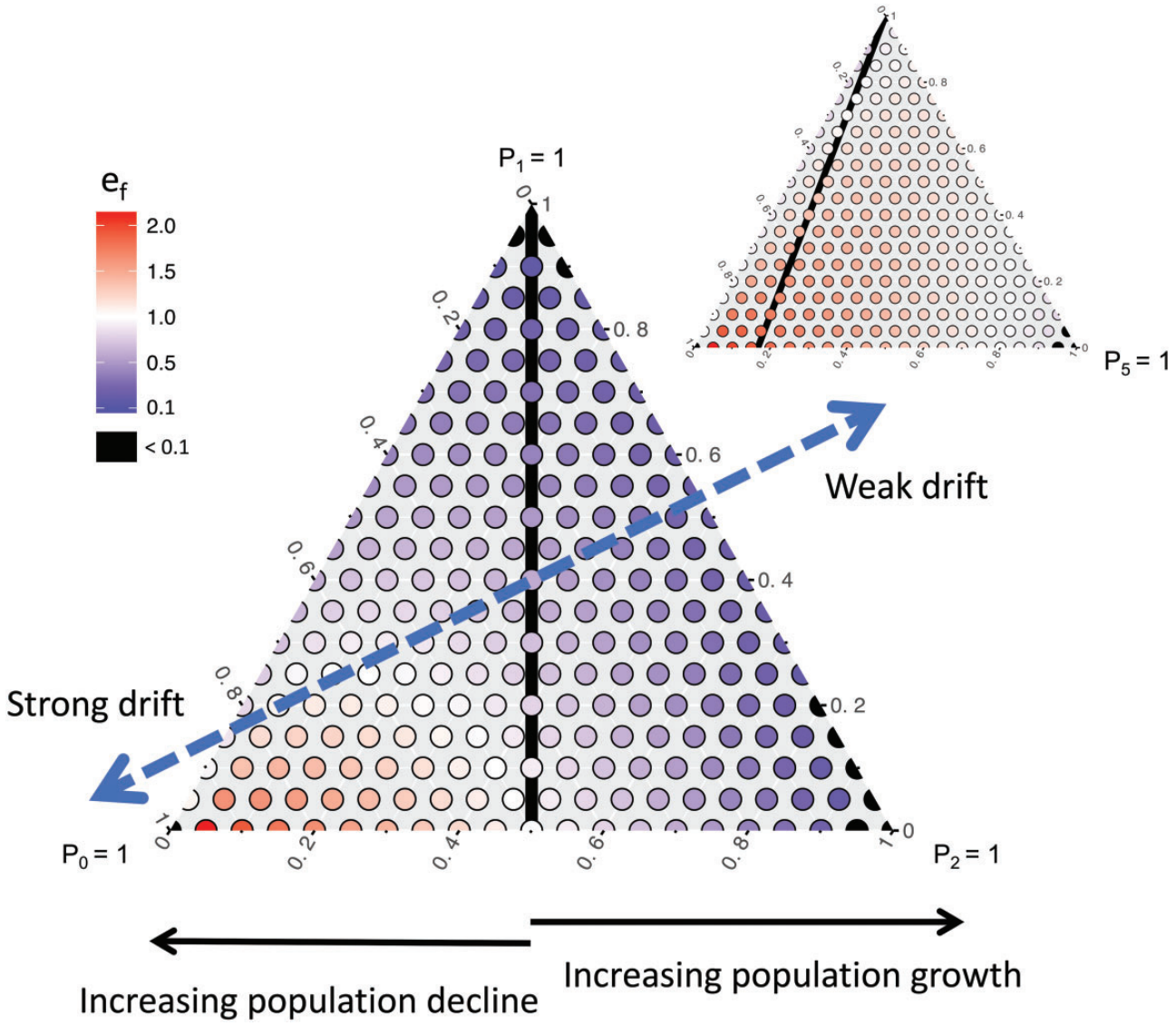

Fig. 2. The efficiency of random drift (e_f) as a function of (P_0, P_1, P_2). P_0: probability of $k = 0$ (cell death), P_1: $k = 1$ (no change), and P_2: $k = 2$ (cell division). The efficiency, ranging from 0 to 2.15 (five for the inset), is shown in the heat map. White dots mark drift efficiency comparable with that of the WF model and red/blue dots indicate relative increases/decreases in the strength of drift. The apex corresponds to $P_0 = 1$, $P_1 = 1$, and $P_2 = 1$, as shown. The interior spots cover all (P_0, P_1, P_2) combinations with $P_0 + P_1 + P_2 = 1$ at intervals of 0.05. The thick midline goes through dots with no population growth ($P_0 = P_2$). The trend of e_f is mainly dependent on P_0 (cell death) indicated by the blue double-headed line. The plots are done under $N = 500$ and $x = 0.5$ but the pattern is nearly identical as long as N is not close to 1 and x is not close to 0. Inset—The same as the main figure except that P_5 is substituted for P_2. As $V(k)$ increases with $k = 5$, the efficiency of drift becomes higher than that in the WF populations although the general trend is the same.

WF population of the same size. Predicted values based on the N_e calculations using equation (2) agree well with simulations in the short term.

Random Drift with Density-Dependent k

Expected random drift starts to deviate from simulations by $t = 50$ (see arrows, fig. 3). Predicted speed of drift gradually becomes lower than the simulated value. The discrepancy is because the prediction is based on WF populations with a fixed size of $N_t = E(N_t)$ whereas the simulated populations have variable sizes with the same mean of $E(N_t)$ but $V(N_t) > 0$.

The distribution of N_t becomes broad and highly skewed (see supplementary fig. S2, Supplementary Material online) as t increases. Because there is no upper bound and the lower bound is 0, the skew is expected. The variation in N_t explains why the mean effective population size will not yield unbiased results for the mean speed of random drift. After all, there are far more populations with smaller values of N that speed up the average drift than there are populations with larger N that slow it down. This asymmetric effect of N_t variation on drift should lead to its underestimation, which is particularly pronounced when t/N_0 is large (see supplementary fig. S1, Supplementary Material online) or when $E(N_t)$ is small (< 100, see supplementary fig. S2, Supplementary Material online).

When regulation is density-dependent, cells should divide less often and/or die more frequently as N increase. When N is

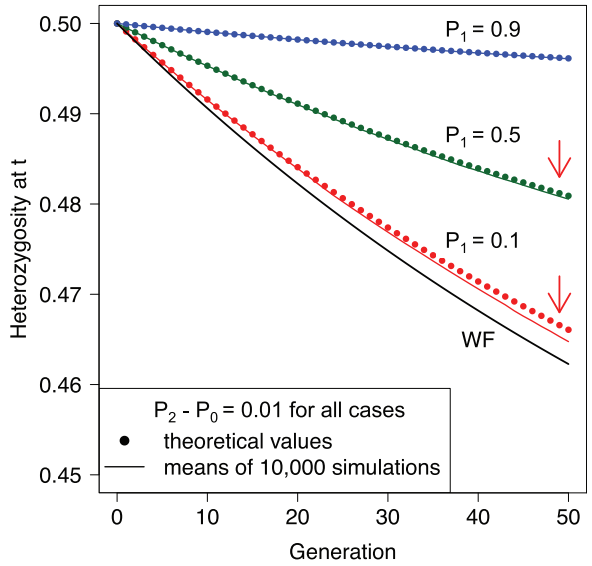

Fig. 3. The speed of random drift under different modes of progeny production (k), all with P_2–$P_0 = 0.01$ (= growth rate). Drift, measured by $H_t = 2x_t(1 - x_t)$ where x_t is gene frequency at time t, is slower in all three modes of progeny production than in the WF population of the same size. The simulations (averaged over 10,000 repetitions) and expectations based on N_e of equation (3) are generally in good agreement. However, as t becomes large, the theory tends to underestimate drift (indicated by arrows) due to the skewed distribution of N_t's around $E(N_t)$ (see text). Mechanisms of population size regulation are presented in figure 3.

very small, all divisions should lead to two viable cells and $P_2 \sim 1$. As the population grows, P_2 should decrease and $P_1 + P_0$ would increase until $P_0 = P_2$, when birth equals death and the population stops growing. P_1, P_2 and P_3 are functions of t. Given that $P_2(t) - P_0(t)$ is the growth rate (see Supplement, Supplementary Material online), the population can take various routes to evolve from $P_2(0) \sim 1$ to $P_0(t) = P_2(t)$. Figure 4 shows the dynamics of random drift along three routes of population growth. Random drift in the WF population with a comparable growth trajectory is shown for comparison. More routes are presented in see supplementary figure S2, Supplementary Material online. Additional possibilities are discussed in the supplementary text, Supplementary Material online.

The Y-axis of figure 4*a* and *b* portrays the change in heterozygosity in each generation ($|\Delta H_t/\Delta t|$) as a measure of random drift, plotted against either time or N_t. There are two notable features. First, random drift in these growing populations is much weaker than in WF populations growing at the same rate. Second, during the course of population growth, the strength of random drift actually increases. Although the increases are modest, the contrast with random drift in WF populations is very significant. The patterns show that the conventional wisdom about random drift depends critically on the specific model of N_t regulation in the WF model.

It should be noted that the conclusion above is robust even when the cell populations grow in arbitrary patterns. Since all combinations of (P_0, P_1, P_2) have been evaluated for

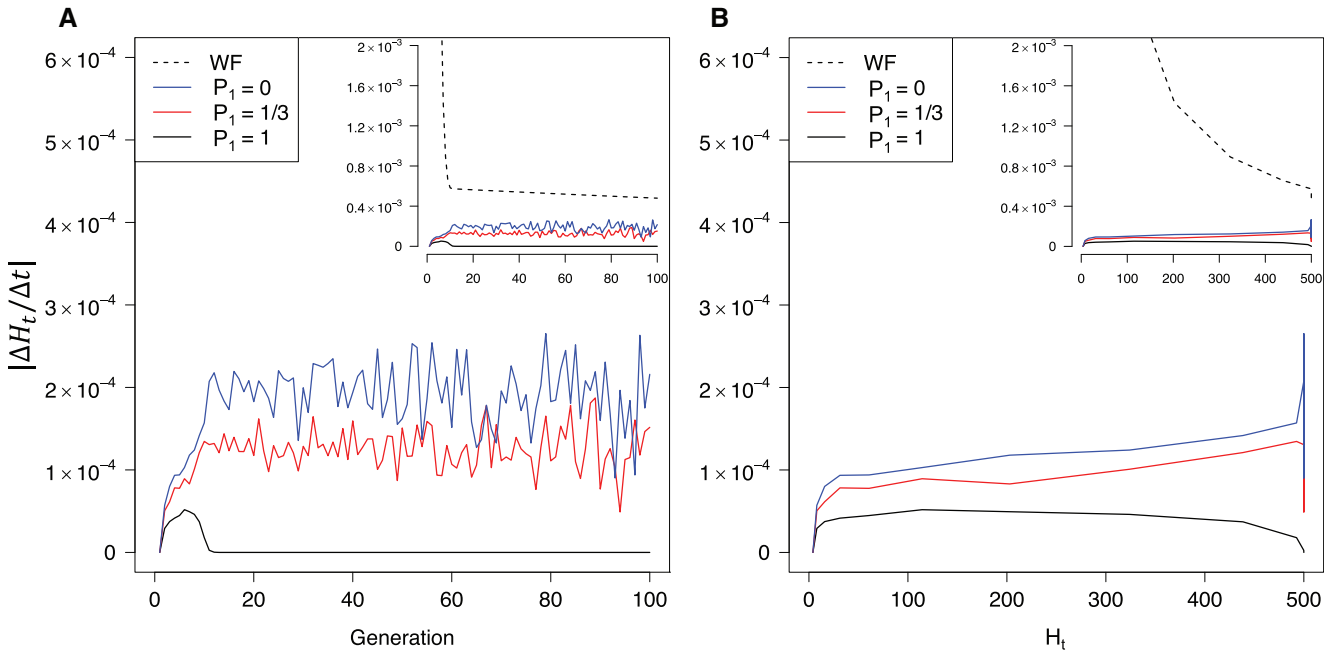

Fig. 4. The decrease in heterozygosity in each generation ($|\Delta H_t/\Delta t|$) as a measure of random drift when cell populations grow from $N_0 = 2$ to the carrying capacity of 500. (a) ($|\Delta H_t/\Delta t|$) versus t. (b) ($|\Delta H_t/\Delta t|$) versus N_t. By the three modes of growth, the population starts at the lower right corner of figure 1 ($P_2 = 1$) and evolves toward the midline of no growth ($P_2 = P_0$; birth = death), arriving at three different endpoints ($P_1 = 0$, $P_1 = 1/3$ and $P_1 = 1$). In the inset shows that ($|\Delta H_t/\Delta t|$) decreases much more drastically in the WF model than in the three modes of growth. Importantly, the strength of drift, shown in higher resolution in the main figure, increases as N increases, opposite of the trend in WF populations. (The black line is a special case with $P_2 = 1$ and $P_1 = 1$ in the beginning and at the end, where there is no drift due to $V(k) = 0$.).

their effect on random drift, it is straightforward to trace any route of population growth by "connecting the dots" of figure 1. Additional discussions on population regulation are presented in Section 3 of Supplement, Supplementary Material online.

Random Drift in Local Patches of Tumors

So far we set the carrying capacity at $C = 500$ (fig. 4). For cell populations with little migration (bacteria, yeast or tumors), this would be equivalent to drift in local patches (Liu et al. 2011; Ling et al. 2015; Sottoriva et al. 2015). The larger population would then consist of many such patches that grow in size as a function of time. Such patterns have indeed been reported in bacterial and yeast colonies (Hallatschek et al. 2007).

In solid tumors, the mass in the interior is either nongrowing or declining due to cell death or necrosis. If these cells in the interior continue to divide ($P_1 < 1$), drift would be stronger inside the mass than on the periphery where the mass grows rapidly. In contrast, in "surfing" natural populations, the prediction is that random drift is stronger on the expanding fronts than in the interior(Excoffier and Ray 2008). The seemingly counter-intuitive inference on drift in growing tumors thus deserves to be tested empirically. In Supplement, Supplementary Material online, several possibilities, documented or conjectured, are further discussed.

Discussion

The attempt to analyze tumor evolution using existing population genetic theories has forced a re-assessment of their generality. Notably, the conventional definition of random drift by WF-scheme sampling has conceptual as well as operational limitations. In this study, we suggest an alternative model of "individual output" to define random drift. The definition has several advantages. First, random does not require the delineation of a well-mixed spatially panmictic population. Second, the profile of progeny production, mainly $E(k)$ and $V(k)$, is often empirically measurable, thus permitting the estimation of drift strength. Third, it is feasible to incorporate population size regulation into the "individual output" model, as shown in figure 4. All three advantages are important for studying the evolution of tumors.

Among these advantages, population size regulation may be most noteworthy. How N is regulated, in addition to the magnitude of N itself, should affect random drift. The WF model assumes that the population is always near the carrying capacity, C. When the model is extended to growing/declining populations, the assumption of $N \sim C$ is still necessary. However, many populations grow when they are far below C. We show that under plausible ecological conditions (figs. 1 and 4), random drift may become weaker when N becomes smaller. This result contradicts the conventional wisdom that a larger N always reduces random drift. The paradox is the consequence of divergent assumptions on population size regulation.

The need to relax the highly restrictive regulations in the WF model is apparent (Marusyk et al. 2012; Gerlinger et al. 2014; Shpak and Lu 2016; Wu et al. 2016). For example, species with great fluctuations in population size may be subjected to very different strength of random drift than we have believed. Invasive species may be subjected to population regulation distinct from that of native species. Domesticated plants and animals may also be under regulation not seen in natural systems. Establishing how random drift operates across these population types will be a challenge to evolutionary biologists. Finally, the evolution of DNA sequences is the cumulative effect of random drift across a large span of time. For instance, if drift in larger populations, which tend to produce more mutations, is stronger than the current estimate, the expected rate of DNA sequence evolution could be correspondingly higher. How much the new assessment of random drift will change our understanding of molecular evolution will be most interesting.

This study also suggests that random drift during tumor evolution may be weaker than expected. Since the level of intratumor diversity has been shown to be high (Ling et al. 2015; Sottoriva et al. 2015), weaker drift may mean an even higher level of diversity, which would have medical implications. Random drift may be only one of many aspects of population genetic theories that require fine-tuning for applications to cancer evolution. Selection and migration may be two other challenging subjects.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Author Contributions

All authors contributed equally.

Acknowledgments

We thank Wen-Hsiung Li, Nancy Moran, Dan Hartl, Brian Charlesworth, Tom Nagylaki, Yunxin Fu, Manyuan Long, Marty Kreitman, Trevor Price, Manuel Lerdau, Darryl Shibata, Robert Gatenby, Hurng-Yi Wang, Shou-Hsien Li, Antony Dean, Weiwei Zhai, Hua Chen, Zheng Hu, Xionglei He, Tian Tang, and Ziwen He for helpful comments. We thank Ao Lan and Yongsen Ruan for finding the equivalent formula of equation (5) under different conditions. This work was supported by 985 Project (33000-18821105), the National Key Basic Research Program of China (2014CB542006), the Science Foundation of State Key Laboratory of Biocontrol (SKLBC16A37) and The University of Chicago Comprehensive Cancer Center Pilot project.

References

Billingsley P. 1995. Probability and measure. New York: Wiley.
Charlesworth B, Charlesworth D. 2010. Elements of evolutionary genetics. Greenwood Village (CO): Roberts and Co. Publishers.
Crow JF, Kimura M. 1970. An introduction to population genetics theory. New York: Harper & Row.
Elandt-Johnson RC, Johnson NL. 1980. Survival models and data analysis. New York: Wiley.

Eldon B, Wakeley J. 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* 172:2621–2633.

Ewens WJ. 2004. Mathematical population genetics. New York: Springer-Verlag.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends Ecol Evol*. 23:347–351.

Gerlinger M. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing (vol 366, pg 883, 2012). *N Engl J Med*. 367:976–976.

Gerlinger M, McGranahan N, Dewhurst SM, Burrell RA, Tomlinson I, Swanton C. 2014. Cancer: evolution within a lifetime. *Ann Rev Genet*. 48:215–236.

Gillespie JH. 1975. Natural selection for within-generation variance in offspring number II. Discrite haploid models. *Genetics* 81:403–413.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR. 2007. Genetic drift at expanding frontiers promotes gene segregation. *Proc Natl Acad Sci U S A*. 104:19926–19930.

Hartl DL, Clark AG. 2007. Principles of population genetics. Sunderland (MA): Sinauer Associates.

Kendall MG, Stuart A, Ord JK, Arnold SF, O'Hagan A. 1994. Kendall's advanced theory of statistics. London; New York: Edward Arnold; Halsted Press.

Kimura M, Crow JF. 1963. The measurement of effective population number. *Evolution* 17:279–288.

Lande R, Engen S, Sæther B-E. 2003. Stochastic population dynamics in ecology and conservation. In: Oxford series in ecology and evolution. Oxford: Oxford University Press. p. 1 online resource (x, 212 p.).

Ling SP, Hu Z, Yang ZY, Yang F, Li YW, Lin P, Chen K, Dong LL, Cao LH, 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 U S A*. 113:E663–E663.

Liu C, Fu X, Liu L, Ren X, Chau CK, Li S, Xiang L, Zeng H, Chen G, Tang LH, et al. 2011. Sequential establishment of stripe patterns in an expanding cell population. *Science* 334:238–241.

Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li XH, Sanchez CA, Paulson TG, Blount PL, Risques RA, Rabinovitch PS, et al. 2006. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet*. 38:468–473.

Marusyk A, Almendro V, Polyak K. 2012. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer*. 12:323–334.

Moran PAP. 1958. Random processes in genetics. *Math Proc Camb Philos Soc*. 54:60.

Shpak M, Lu J. 2016. An evolutionary genetic perspective on cancer biology. *Ann Rev Ecol Evol Syst*. 47:25–49.

Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, Marjoram P, Siegmund K, Press MF, Shibata D, et al. 2015. A Big Bang model of human colorectal tumor growth. *Nat Genet*. 47:209–216.

Tao Y, Ruan J, Yeh SH, Lu X, Wang Y, Zhai W, Cai J, Ling S, Gong Q, Chong Z, et al. 2011. Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. *Proc Natl Acad Sci U S A*. 108:12042–12047.

Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. 2016. Identification of neutral tumor evolution across cancer types. *Nat Genet*. 48:238–244.

Wright S. 1969. Evolution and the genetics of populations: Vol. 2. The theory of gene frequencies. Chicago (IL): University of Chicago Press.

Wu CI, Wang HY, Ling S, Lu X. 2016. The ecology and evolution of cancer: the ultra-microevolutionary process. *Annu Rev Genet*. 50:347–369.