Properties of Samples With Segregating Polymerase Chain Reaction (PCR) Dropout Mutations Within a Species

Evolutionary Bioinformatics
Volume 15: 1–15
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934319883612

Cortland K Griswold

Department of Integrative Biology, University of Guelph, Guelph, ON, Canada.

ABSTRACT: In polymerase chain reaction (PCR)-based DNA sequencing studies, there is the possibility that mutations at the binding sites of primers result in no primer binding and therefore no amplification. In this article, we call such mutations PCR dropouts and present a coalescent-based theory of the distribution of segregating PCR dropout mutations within a species. We show that dropout mutations typically occur along branch sections that are at or near the base of a coalescent tree, if at all. Given that a dropout mutation occurs along a branch section near the base of a tree, there is a good chance that it causes the alleles of a large fraction of a species to go unamplified, which distorts the tree shape. Expected coalescence times and distributions of pairwise sequence differences in the presence of PCR dropout mutations are derived under the assumptions of both neutrality and background selection. These expectations differ from when PCR dropout mutations are absent and may form the basis of inferential approaches to detect the presence of dropout mutations, as well as the development of unbiased estimators of statistics associated with population-level genetic variation.

KEYWORDS: Background selection, coalescent, eDNA, metagenomics, pairwise differences, tree shape

RECEIVED: September 6, 2019. **ACCEPTED:** September 25, 2019.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the NSERC (Canada) - Discovery program.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Cortland K Griswold, Department of Integrative Biology, University of Guelph, Guelph, ON N1G 2W1, Canada.
Email: cgriswol@uoguelph.ca

Introduction

DNA sequencing is either primer based, whereby a specific genetic locus is amplified by polymerase chain reaction (PCR) and sequenced[1] or by shotgun sequencing, whereby genomes are randomly broken into small segments and sequenced directly with no amplification.[2,3] In primer-based sequencing, 2 primers that together flank a locus are used. Primers are typically about 18 to 24 nucleotides in length and bind to DNA. Each primer initiates replication on separate strands, with complementary effects such that after several rounds of replication and strand separation a specific locus is amplified.[4]

Substitutions at the binding sites of primers can lower the probability a primer binds and initiates replication. Not all substitutions at the binding site of a primer are equal. For example, substitutions within the last 3 to 4 nucleotides of the 3′ end can significantly reduce PCR replication.[5,6] Primer coverage corresponds to the proportion of species or taxa in the sample that are, or are expected to be, amplified and sequenced for a primer or set of primers.[7] Studies of primer coverage that compare primers to DNA sequence databases indicate that coverage for a specific primer varies depending on taxon, ranging from 0% to nearly 100%,[7,8] where the taxon level is at the domain or phylum. For example, in the bacterial phylum, *Nitrospirae* coverage is about 97% for the commonly used 16S rRNA primer 519F if a single mismatch is allowed within the last 4 nucleotides of a primer binding site, but only about 32% if a single mismatch is not allowed.[7] For the bacterial phylum *Lentisphaerae* and the commonly used 16S rRNA primer 338F, coverage is 97% if a single mismatch is allowed within the last 4 nucleotides of a primer binding site, but 0% if a single mismatch is not allowed.[7] High coverage may occur at the phylum level for a specific primer, but is typically lower at the domain level due to broader sequence diversity.

A lack of primer coverage is straightforward to detect in studies where it is sought to sequence DNA from a single individual because amplification and sequencing will fail, which is directly detectable. In contrast, it is becoming increasingly common to pool, amplify, and sequence DNA from multiple individuals across one or more species. For example, in primer-based metagenomic and eDNA studies, a sample is taken from the environment that contains individuals and/or DNA from potentially thousands of species or OTUs (operational taxonomic units) and multiple individuals per species or OTU. In this approach, PCR amplification and sequencing is often non-targeted and used to discover or detect the species, OTUs or higher taxonomic groups present, as well as targeted and used to detect a specific species, OTU or taxonomic group. In non-targeted applications, there is the possibility that a taxonomic group or subset of individuals within a taxonomic group go undiscovered or undetected because of a mismatch between primers and binding sites. Furthermore, even with targeted sequencing, there is the possibility of low primer coverage.[7,8]

As context, many datasets deposited in the European Bioinformatics Institute metagenomics database (https://www.ebi.ac.uk/metagenomics/)[9] involve primer/amplicon-based and non-targeted sequencing from nature. Although there has been progress in developing reference databases for primer/amplicon-based metagenomic studies to match sequences to species or higher taxonomic groups in a targeted manner,[10] there is still a need for assessing whether a taxonomic group goes partially or fully unsequenced in non-targeted sequencing studies, as well as assess the degree of coverage of targeted taxa.

A challenge for primer-based metagenomic studies is assessing coverage for a sample from nature. A rarefaction curve may be used to assess coverage by plotting the number of taxa delineated versus the number of sequences recorded.[11] If this curve is asymptotic, then nearly all taxa have been delineated. Nevertheless, this only applies to taxa that are amplified by the primer or primer set. Even if primer coverage was less than 100%, an asymptotic rarefaction curve can occur with more sequences because eventually all amplified species can eventually be sequenced and delineated to a taxon or as coming from an unknown taxon. Generally, it would be helpful for a metagenomic study to have available approaches to assess primer coverage for their specific sample from nature. In this article, we present theory to help the development of approaches to assess coverage.

In particular, we use coalescent theory to derive expectations for the pattern of nucleotide variation within a species when a mutation or a set of mutations at primer site(s) block the chain of events during a primer-based sequencing study, such that a set of sequences are not recorded as being present. This mutation or a set of mutations could either completely block a primer from binding to its binding site or reduce binding, such that amplification is so low that the locus is not sequenced for a set of individuals descendent from the mutation or a set of mutations. We call such mutations "PCR dropouts."

A PCR dropout mutation or a set of PCR dropout mutations may occur ancestrally to all individuals within a species or later (forward in time), such that only a subset of individuals from a species have one or more of the dropouts. Dropout mutations that occur ancestrally to all individuals and are sufficient to block amplification of all individuals within the species are *not* the focus of this study. Instead, we focus on dropouts that are segregating within a population, such that these dropouts occur later (forward in time) than the most recent common ancestor (MRCA) to the species. Using coalescent theory,[12,13] we show that segregating dropout mutations give rise to distinct patterns of DNA sequence variation. This pattern may be used to assess whether a species (or OTU) is prone to PCR dropouts and therefore reduced coverage. Furthermore, if several species (or OTUs) within a taxon have a signal of dropouts, this may indicate a larger problem of coverage for the entire taxon.

Although this article focuses on dropouts within a species, similar principles apply at higher levels of biological organization. We choose to first focus on within-species variation because DNA sequence evolution is well characterized by the coalescent process and this process is universal among species and taxa. In contrast, at higher levels, there is no standard theoretical framework of DNA sequence evolution, although Hey's[14] approach may be promising in the context PCR dropouts and metagenomics (see section "Discussion").

Expectations for DNA sequence variation at the population level and in the presence of segregating dropout mutations will also be of potential use in other contexts besides assessing primer coverage. For example, there has been the development of estimators of nucleotide diversity that account for the higher rate of sequencing errors associated with next-generation sequencing technology.[15,16] Yet, methods are not available that account for segregating PCR dropouts. Furthermore, methods have been developed to use population-level variation to infer the phylogenetic structure of ecological communities using metagenomic data.[17] Accordingly, it would be helpful to consider the effect of PCR dropouts on population-level variation and whether they may affect inferences of phylogenetic structure.

The organization of this article is as follows: First, we present theory assessing the probability that dropout mutations occur along branch sections that are deep in the coalescent tree of a species or along branch section that occur more recently, as well as whether we expect a species to segregate a single dropout mutation or multiple dropout mutations. Generally, we find that dropout mutations typically occur along branch sections that are deep in the coalescent tree of a species. Furthermore, given rates of mutation at primer sites, a single dropout mutation typically segregates in a sample. Focusing on dropouts that completely block amplification, we then derive expectations for the distribution of coalescence times and pairwise differences of sequenced alleles at a locus. This analysis indicates that the distribution of coalescence times and pairwise differences is affected by the presence of a dropout and may therefore give rise to a detectable signal of a dropout. The theory accounts for the process of background selection,[18,19] which is occurring at loci used in primer-based metagenomic studies, such as 16S/18S RNA.[1]

Theory

It is assumed that a sample is taken from an environment, be it water, soil, feces, etc, and DNA is purified from the sample and subject to PCR and sequencing. Although it may be that within a sample there are individuals across the tree of life, only a single genetic locus is amplified by PCR. For simplicity, we assume that the PCR primers target bacteria and archaea such that species are haploid and effective population sizes are potentially large. Although this article focuses on haploid microbes, the theory applies directly to diploid microbes with a change in timescale (see below). For a given species in the sample, there is a sample size of n. This sample size is not known. In the theory presented, we nevertheless assume a sample size of n and then develop expectations for how many individuals within the sample are expected to be unamplified given a sample of size n. The theory demonstrates that n is a nuisance parameter and detectable effects of dropout mutations can be inferred without the knowledge of n.

For PCR sequencing of a species or OTU to occur, DNA primers need to amplify the DNA of a species or OTU to a sufficient level that allows for sequencing. Here we define μ to be the rate of mutation that leads to a PCR dropout. For simplicity, we assume that a mutation either causes a dropout or does not, ie, the effects of mutations are discrete. An alternative model is quantitative, such that a mutation may decrease amplification by a certain fraction that is between, but not equal to, 0 and 1. We assume that the nucleotide sites that affect PCR amplification do not affect the fitness of an organism and consequently evolve neutrally. Amplified loci in metagenomic and environmental DNA studies are typically under purifying selection, such that we expect the amplified PCR locus to be affected by background selection.[18] Initially, we assume neutrality and no background selection, but later consider the consequences of background selection. Furthermore, when developing expectations for polymorphism at an amplified locus, we focus on synonymous sites and/or assume neutrality and no background selection, but later consider the consequences of background selection.

Neutrality and no background selection at the amplified locus

A natural theoretical framework to model dropouts assuming neutrality is Kingman's[12] coalescent (see also Tajima[13]). In the Kingman coalescent, the genealogical history of a sample of size n is divided into $n - 1$ sections in which coalescent events define the sections, whereby descending from the top to the bottom of the coalescent tree, the first section consists of n lineages, the second $n - 1$ lineages, etc, with the last section consisting of 2 lineages. Following Fu,[20] we call the section of the tree consisting of k branches "level k" (Figure 1). Mutations causing a dropout can occur along branches within each of these sections. A dropout mutation along a branch section is "unique" if no other dropout mutation occurs along its corresponding ancestral lineage from the point of mutation back to the MRCA of the sample (see Figure 2, section "Probability Y_k of the X_k dropouts at level k are unique along their corresponding lines of descent"). Unique dropout mutations are of interest because all of the descendent lineages of the dropout will not be amplified by PCR, including lineages that have incurred subsequent dropout mutations (hence why we focus on unique dropouts). Next, we derive expectations for the distribution of the number of dropouts when there are k lineages remaining from a sample of size n. We then derive the conditional distribution of the number of dropouts that are unique at level k given a certain number of dropout mutations. For the set of unique dropouts, we derive the distribution of the number of lineages descendent from each dropout. With information about the number of descendant lineages, we then derive the conditional distribution of pairwise differences in the sample.

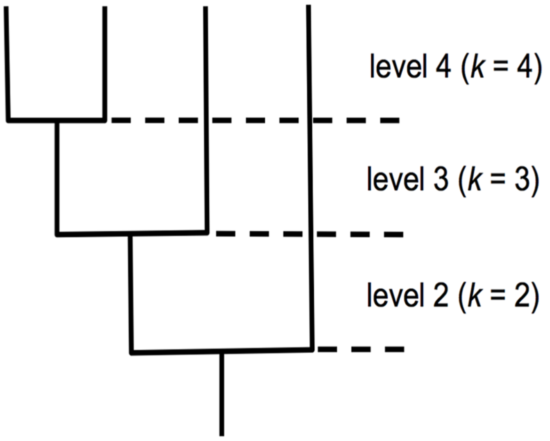

Figure 1. The use of the term "level." A level corresponds to the number of branch sections in the coalescent tree. Branch sections are not drawn to scale. Lower levels in the tree have longer branch sections according to coalescent theory.[12]

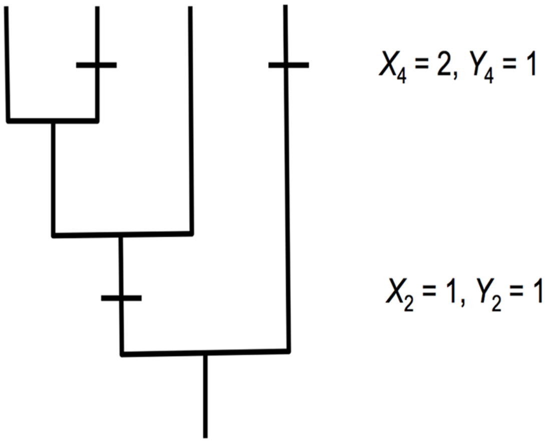

Figure 2. The use of the term "unique." In the coalescent tree, 2 branch sections at level 4 have a dropout mutation, such that $X_4 = 2$ and 1 branch section at level 2 has a dropout mutation, such that $X_2 = 1$. Although there are 2 dropout mutations at level $k = 4$, only 1 is unique ($Y_4 = 1$), such that there are no other dropout mutations along its line of descent to the most recent common ancestor (MRCA) of the sample. At level $k = 2$, there is 1 unique dropout ($Y_2 = 1$) because no other dropout mutation occurs along its line of descent to the MRCA.

This distribution is conditioned on the number of unique dropouts and numbers of descendent lineages from these dropouts, noting that by determining what individuals are not amplified by PCR we can then focus attention on the properties of individuals that are amplified by PCR.

Probability of X_k branches with dropouts at level k, $p(x_k)$. Define T_k to be a random variable giving the length of branch sections at level k in the coalescent tree. In continuous time, the distribution of T_k ($f(T_k = t_k)$) is approximately exponential[12] such that

$$f(t_k) = \binom{k}{2} e^{-\binom{k}{2} t_k}$$

where to simplify notation we let $f(T_k = t_k) = f(t_k)$. The approximation is in the diffusion limit as the effective size of a population (N_e) tends to infinity. In the Wright-Fisher model, 1 unit of time equals N_e generations and in the Moran model 1 unit of time equals $N_e/2$ generations.[12] For simplicity, we use 1 effective population size in our analysis, keeping in mind that the effective size of a Moran population is one half that of a Wright-Fisher population. Furthermore, note that for diploid microbes the baseline timescale is $2N_e$ generations.

The expected number of dropout mutations per branch section at level k is $\mu E(T_k)$, where the mutation rate (μ) is measured per generation and T_k is measured on a generation timescale. The distribution of the number of mutations along a branch section is Poisson distributed and the probability of at least 1 dropout along a branch section at level k (P_k) is

$$P_k = 1 - e^{-\mu E(T_k)}$$

noting that $e^{-\mu E(T_k)}$ is the probability of no dropout mutations. Define X_k to be a random variable for the number of branch sections at level k with at least 1 dropout, then the probability $X_k = x_k$ ($p(X_k = x_k)$) is

$$\begin{aligned} p(x_k) &= \int_0^\infty \binom{k}{x_k} P_k^{x_k}(1 - P_k)^{k-x_k} f(t_k)dt_k \\ &= \frac{2\binom{k}{2}\binom{k}{x_k}\Gamma\left(\frac{k(\theta + k-1)}{\theta} - x_k\right)\Gamma(x_k + 1)}{\theta\Gamma\left(\frac{k(k-1)}{\theta} + k + 1\right)} \end{aligned} \tag{1}$$

where $\theta = 2N_e\mu$, $\Gamma(\cdot)$ is the gamma function, and, to simplify notation, we let $p(X_k = x_k) = p(x_k)$. A straightforward calculation from $p(x_k)$ is the probability $X_k > 0$ ($p(X_k > 0)$), which is

$$p(X_k > 0) = \frac{\theta}{k - 1 + \theta} \tag{2}$$

and corresponds to the probability of at least 1 mutation prior to a coalescent event at level k (cf. Wakeley[21]).

Probability Y_k of the X_k dropouts at level k are unique along their corresponding lines of descent. Define Y_k to be a random variable for the number of lineages at level k that have a dropout mutation at level k and no prior dropout mutations from level k to the MRCA of the sample, such that Y_k is the number of unique dropout mutations at level k and $Y_k \leq X_k$. Figure 2 illustrates the notation, whereby at level $k = 4$ in the tree, there are 2 dropout mutations ($X_4 = 2$), but a second dropout mutation occurs along the line of descent to the MRCA for one of the dropouts, such that $Y_4 = 1$; at level $k = 2$, $X_2 = 1$ and $Y_2 = 1$.

To calculate the distribution for Y_k, we note that moving down the coalescent tree from level k to the MRCA of the sample, we can define a set of numbers $\{x'_{k,i}\}$ for i from 1 to X_k which consists of the number of lineages with i descendants that have a unique dropout mutation at level k. Combining the set $\{x'_{k,i}\}$ with the number k', which is the number of lineages out of the k lineages remaining in the sample, forms a state space $\{x'_{k,1}, x'_{k,2}, \ldots, x'_{k,X_k}, k'\}$ that characterizes the genealogical and dropout mutation process from level k to the MRCA of the sample. At the MRCA, $k' = 1$ and either all of the $x'_{k,i}$ equal zero, such that there are no unique dropout mutations at level k, or one and only one of the $x'_{k,i}$ equals 1, such that if $x'_{k,i} = 1$ and $k' = 1$ then $Y_k = i$.

Backward in time, 2 events can occur that are relevant to determining the distribution of Y_k. One event is a coalescence and the other event is a dropout mutation along a descendent lineage. If a coalescent event occurs, it could be between the 2 lineages with a dropout mutation, between a lineage with a dropout mutation and without a dropout mutation or between 2 lineages without a dropout mutation. For example, 2 lineages within the set with i descendants may coalesce resulting in the transition $\{\ldots, x'_{k,i}, \ldots, x'_{k,i+i}, \ldots, k'\} \rightarrow \{\ldots, x'_{k,i} - 2, \ldots, x'_{k,i+i} + 1, \ldots, k' - 1\}$ and this occurs at rate $\binom{x'_{k,i}}{2}$ on an appropriate timescale. Or, 2 lineages, one with i descendants and the other with j descendants, may coalesce resulting in the transition

$$\begin{gathered} \{\ldots, x'_{k,i}, \ldots, x'_{k,j}, \ldots x'_{k,i+j}, \ldots, k'\} \rightarrow \\ \{\ldots, x'_{k,i} - 1, \ldots, x'_{k,j} - 1, \ldots, x'_{k,i+j} + 1, \ldots, k' - 1\} \end{gathered}$$

for $j > i$ at rate $\binom{x'_{k,i} + x'_{k,j}}{2} - \binom{x'_{k,i}}{2} - \binom{x'_{k,j}}{2}$. A coalescent event could occur between lineages that are not descendants of dropout mutations or between a lineage with a descendent dropout mutation and one without a descendent dropout, such that $\{x'_{k,1}, x'_{k,2}, \ldots, x'_{k,X_k}, k'\} \rightarrow \{x'_{k,1}, x'_{k,2}, \ldots, x'_{k,X_k}, k' - 1\}$ at rate $\binom{k' - \sum_i x'_{k,i}}{2} + \sum_i x'_{k,i}(k' - \sum_j x'_{k,j})$. Last, if a dropout mutation occurs along a descendent lineage below level k, then a dropout mutation at level k is no longer unique. For each element $x'_{k,i}$ in $\{x'_{k,1}, x'_{k,2}, \ldots, x'_{k,X_k}, k'\}$, this event occurs at rate $\theta x'_{k,i}/2$.

Generally, for $k > 2$, the following transitions are possible and their corresponding rate of transition, where for efficiency we only show elements in the state space set that change

$$
\begin{array}{ll}
\text{Transition} & \text{Rate} \\
\{x'_{k,i}, x'_{k,i+i}, k'\} \rightarrow \{x'_{k,i} - 2, x'_{k,i+i} + 1, k' - 1\} & \binom{x'_{k,i}}{2} \\
\{x'_{k,i}, x'_{k,j}, x'_{k,i+j}, k'\} \rightarrow \{x'_{k,i} - 1, x'_{k,j} - 1, x'_{k,i+j} + 1, k' - 1\} & \binom{x'_{k,i} + x'_{k,j}}{2} - \binom{x'_{k,i}}{2} - \binom{x'_{k,j}}{2} \\
\{k'\} \rightarrow \{k' - 1\} & \binom{k' - \sum_i x'_{k,i}}{2} + \sum_i x'_{k,i}\left(k' - \sum_j x'_{k,j}\right) \\
\{x'_{k,i}, k'\} \rightarrow \{x'_{k,i} - 1, k' - 1\} & \begin{cases} \theta x'_{k,i}/2 & k' < k \\ 0 & k' = k \end{cases}
\end{array}
\qquad (3)
$$

For the set $\{x'_{k,1}, x'_{k,2}, \ldots, x'_{k,X_k}, k'\}$, there will be a corresponding state space $\mathcal{X}_k$ from level k to the MRCA. Define $X(t)$ to be a random variable for the state of the system at time t into the past starting at some point in level k. $X(t)$ is a finite-state and continuous-time Markov process with multiple absorbing states when $k' = 1$. We can define an $|\mathcal{X}_k| \times |\mathcal{X}_k|$ matrix $\mathbf{A}$ to be the corresponding finite-state and continuous-time Markovian transition rate matrix, where $|\mathcal{X}_k|$ is the size of the set $\mathcal{X}_k$. If the current state of the system is i, then the probability of being in state j at time t generations into the past ($P_{ij}(t)$) satisfies the differential equation

$$\frac{dP(t)}{dt} = \mathbf{A}P(t) \qquad (4)$$

where $P(t)$ is the full $|\mathcal{X}_k| \times |\mathcal{X}_k|$ matrix consisting of elements $P_{ij}(t)$ and $\mathbf{A}$ is defined by the transition rates in equation (3). The solution to the differential equation is

$$P(t) = \sum_{n=0}^{\infty} \frac{\mathbf{A}^n t^n}{n!} \qquad (5)$$

For our purposes, we are interested in the probability at time t that the system is in a state in which $k' = 1$ starting from the initial state $\{X_k, 0, \ldots, 0, k\}$ and furthermore whether one of the $x'_{k,i}$ equals 1. Label the initial state $\{X_k, 0, \ldots, 0, k\}$ as $\varnothing$ and states in which $x'_{k,i} = 1$ with an integer j corresponding to the number of descendent lineages with a unique dropout mutation at level k. For example, for $k = 4$ and $X_k = 2$, the possibilities are $\{0, 0, 1\} \Rightarrow j = 0$, $\{1, 0, 1\} \Rightarrow j = 1$, and $\{0, 1, 1\} \Rightarrow j = 2$.

Define $p(Y_k = j|X_k, t)$ to be the probability $Y_k = j$ at time t given X_k, then

$$p(Y_k = j|X_k, t) = P_{\varnothing j}(t) \qquad (6)$$

Furthermore, in the limit as $t \rightarrow \infty$, then

$$p(Y_k = j|X_k) = \lim_{t \rightarrow \infty} P_{\varnothing j}(t) \qquad (7)$$

As $\mathbf{A}$ has a finite dimension and satisfies properties of instantaneous transition rates, the limit is finite and defined.[21]

For the case $X_k = 1$, it is straightforward to directly write down $p(Y_k = 1|X_k = 1)$ for $k > 2$, which is

$$
\begin{aligned}
p(Y_k = 1|X_k = 1) &= \prod_{i=2}^{k-1} \frac{\binom{i}{2}}{\binom{i}{2} + \frac{\theta}{2}} \\
&= \frac{\theta\Gamma(k-1)\Gamma(k)\Gamma(a_-)\Gamma(a_+)}{\Gamma(a_- + k - 1)\Gamma(a_+ + k - 1)}
\end{aligned}
\qquad (8)
$$

where $a_- = (1/2) - (1/2)\sqrt{1 - 4\theta}$ and $a_+ = (1/2) + (1/2)\sqrt{1 - 4\theta}$, as well as recognizing that for the dropout to be unique no subsequent dropout mutations can occur along its lineage back to the MRCA of the sample.

Probability of the set of $\{i_1, i_2 \ldots, i_{Y_k}\}$ descendants given Y_k unique dropouts at level k. Define $\{i_1, i_2, \ldots, i_{Y_k}\}$ to be a set consisting of Y_k elements, in which an element is an integer greater than 0 equaling the number of descendants of one of the Y_k lineages, respectively, and noting that for $Y_k = 1$ the set is $\{i_1\}$. Enumeration of $\{i_1, i_2, \ldots, i_{Y_k}\}$ is a direct application of the Pólya urn scheme and nicely applied in a similar population genetic context by Fu.[19] For a sample of size n and level k in a coalescent tree, there are

$$\binom{n-1}{k-1}$$

ways the n individuals can be assigned to k ancestors. Of the k ancestors, we are interested in the number of descendants of each of the Y_k unique dropout mutations. The number of ways of drawing a set $\{i_1, i_2, \ldots, i_{Y_k}\}$ from n individuals under the constraint that $k - Y_k - 1 > 0$ at level k is

$$\binom{n - \sum_{j=1}^{Y_k} i_j - 1}{k - Y_k - 1}$$

such that the probability of the set $\{i_1, i_2, \ldots, i_{Y_k}\}$ given Y_k ($p(\{i_1, i_2, \ldots, i_{Y_k}\}|Y_k)$) is

$$p(\{i_1, i_2, \ldots, i_{Y_k}\}|Y_k) = \frac{\binom{n - \sum_{j=1}^{Y_k} i_j - 1}{k - Y_k - 1}}{\binom{n-1}{k-1}} \qquad (9)$$

Probability of the set of $\{i_1, i_2, \ldots, i_{Y_k}\}$ descendants of unique dropouts at level k. From the law of conditional probability, the probability of the set of $\{i_1, i_2, \ldots, i_{Y_k}\}$ descendants of unique dropouts at level k ($p_k(\{i_1, i_2, \ldots, i_{Y_k}\})$ is

$$p_k(\{i_1, i_2, \ldots, i_{Y_k}\}) \\ = \sum_{X_k}\sum_{Y_k} p(\{i_1, i_2, \ldots, i_{Y_k}\}|Y_k)p(Y_k|X_k)p(X_k) \tag{10}$$

For sufficiently small θ, it is most likely that at a level there is a single dropout, given a dropout occurs, and we can focus our attention on the case

$$p_k(\{i_1\}) = p(\{i_k\}|Y_k = 1)p(Y_k = 1|X_k = 1)p(X_k = 1) \\ = \frac{\theta\Gamma(k-1)\Gamma(k)\Gamma(a_-)\Gamma(a_+)}{\Gamma(a_- + k - 1)\Gamma(a_+ + k - 1)} \\ \frac{\binom{n-i_1-1}{k-1-1}}{\binom{n-1}{k-1}} \frac{k^2(k-1)\Gamma\left(\frac{k(\theta+k-1)}{\theta}\right)}{\theta\Gamma\left(k+1+\frac{k(k-1)}{\theta}\right)} \tag{11}$$

For the purposes of metagenomics or eDNA studies, we may be interested in the probability $i_1 \geqslant d$ for level k ($p_k(i_1 > d)$), which is the probability of the loss of d or more individuals in the sample due to a unique dropout at level k, and is calculated as

$$p_k(i_1 \geqslant d) = \sum_{i_1 = d}^{n-k-1} p_k(\{i_1\}) \\ = \frac{\theta\Gamma(k)\Gamma(a_-)\Gamma(a_+)\left(\frac{\Gamma(n-d+1)}{\Gamma(n-d-k+2)} - \Gamma(k-1)\right)}{(k-1)\binom{n-1}{k-1}\Gamma(a_- + k - 1)\Gamma(a_+ + k - 1)} \\ \frac{k^2(k-1)\Gamma\left(\frac{k(\theta+k-1)}{\theta}\right)}{\theta\Gamma\left(k+1+\frac{k(k-1)}{\theta}\right)} \tag{12}$$

Distribution of pairwise differences with dropouts. The presence of dropouts may affect the distribution of pairwise differences for a given species. For sufficiently small θ, unique dropouts tend to occur along branch sections near the MRCA of a sample (see section "Analysis and Results"). We therefore focus on the derivation of expectations for the distribution of pairwise difference when $Y_2 = 1$ and $Y_3 = 1$, noting that, for unique dropouts at higher levels, the calculation of pairwise differences follows similar principles.

Single unique dropout at level 2 (Y_2 = 1). A dropout generates a subtree of $n - i_1$ sequenced individuals from a sample of n individuals in total. A subtree has levels just like the coalescent tree for an entire sample. To calculate the distribution of pairwise differences between 2 random and sequenced individuals involves determining (1) whether the 2 sequenced individuals coalesce at level j in the subtree, (2) whether the entire sample is at level k given the subtree is at level j, and (3) the time to coalesce given the level in the entire sample.

For a subtree of size $n - i_1$, the probability that 2 randomly chosen individuals at the tips of a subtree coalesce at level j ($p_{n-i_1}(j)$) for $j \geqslant 1$ is

$$p_{n-i_1}(j) = \frac{1}{\binom{j+1}{2}} \prod_{\ell = j+2}^{n-i_1} \left(1 - \frac{1}{\binom{\ell}{2}}\right) \\ = \frac{2(n-i_1+1)}{(n-i_1-1)(j+1)(j+2)} \tag{13}$$

Wiuf and Donnelly[22] derived the probability that, given a subtree first enters level j, the entire sample is at level k ($\phi(k|j, n, n-i_1)$), such that

$$\phi(k|j, n, n-i_1) = \frac{\binom{n-k-1}{n-i_1-1}\binom{k}{j+1}}{\binom{n}{n-i_1-1}}$$

and where here $j > 0$ and $k > j$.

Given k, the distribution of coalescence times $f_k(t)$ is

$$f_k(t) = \sum_{i = k+1}^{n} \\ \frac{(-1)^{k+1-i}(2i-1)\Gamma(i+k)\Gamma(n)\Gamma(n+1)}{i(i-1)\Gamma(i-k)\Gamma(k)\Gamma(k+1)\Gamma(n-i+1)\Gamma(n+i)} \\ \binom{i}{2} e^{-\binom{i}{2}t}$$

and can be calculated directed from the sum of exponentially distributed random variables.[21,24] Note further that i is indexed initially with $k + 1$ because the transition in $\phi(k|j, n, n-i_1)$ was from $k + 1 \rightarrow k$. If Z is a random variable for the number of pairwise differences between the 2 sequences, then its distribution $p_k(Z = z) = p_k(z)$ is

$$p_k(z) = \int_0^\infty \frac{2^z e^{-2\nu t}(2\nu t)^z}{z!} f_k(t)dt$$

assuming that mutations are Poisson distributed along branch sections and the rate of mutation causing pairwise differences is ν. Over the distributions for j and k, the probability of z pairwise differences ($p(z)$) is

$$p(z) = \sum_{j=1}^{i_1}\sum_{k=2}^{n} p_k(z)\phi(k|j, n, n-i_1)p_{n-i_1}(j) \tag{14}$$

Single unique dropout at level 3 (Y_3=1). Here we assume that the dropout mutation occurred such that the 2 clades of sequenced individuals coalesce when $k = 2$ for the entire sample (Figure 3). We label these 2 clades A and B, such that the number of sequenced individuals in clade A is a and the

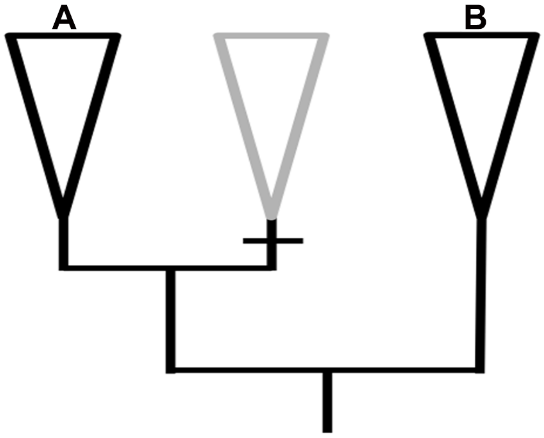

Figure 3. A coalescent tree when $Y_3 = I$. Triangles indicate clades of individuals in the sample. Clades A and B are sequenced, whereas the gray clade descends from a dropout mutation.

number in clade B is b. Two lineages of sequenced individuals can both occur in clade A with probability $(a/(a+b))((a-1)/(a+b-1))$, both in clade B with probability $(b/(a+b))((b-1)/(a+b-1))$, or 1 in clade A and 1 in clade B $2(ab/(a+b)(a+b-1))$.

Let $p(z|A)$ be the probability of z pairwise differences given that 2 sequences are in clade A. As we are finding the marginal probability of z conditioned on a single clade, the result from Wiuf and Donnelly[22] given above applies, such that

$$p(z|A) = \sum_{j=1}^{a} \sum_{k=3}^{n} p_k(z)\phi(k|j, n, a)p_a(j) \tag{15}$$

which differs from $p(z)$ because for clade A the minimum value of k for which $j = 1$ is $k = 3$, noting that a is substituted for $n - i_1$. For clade B, the 2 lineages can coalesce up to $k = 3$ also such that $p(z|B)$ is equal to $p(z|A)$, except for b substituted for a.

When one sequenced individual is in clade A and the other in clade B (A/B), the time to coalesce is distributed as $f_2(t)$, such that

$$p(z|A/B) = p_2(z) \tag{16}$$

and together $p(z)$ for $Y_3 = 1$ is

$$p(z) = \frac{a}{a+b}\frac{a-1}{a+b-1}p(z|A) + \frac{b}{a+b}\frac{b-1}{a+b-1} p(z|B) + 2\frac{ab}{(a+b)(a+b-1)}p(z|A/B) \tag{17}$$

The effect of dropouts on coalescence times of sequenced individuals. The previous section demonstrated that dropouts affect the distribution of pairwise differences relative to when there are no dropouts. The difference in the distribution of pairwise difference comes about because of differences in coalescence times, which in turn reflects the $j \rightarrow k$ mapping in the $\phi(k|j, n, .)$ function. Here we further explore the effect of the $j \rightarrow k$ mapping on coalescence times to see if there is additional information that may be used to detect dropouts.

Single unique dropout at level 2 ($Y_2 = 1$). For $Y_2 = 1$, there is a subtree of $n - i_1$ sequenced individuals that can be divided into j levels in a similar manner as for the entire sample n that is divided into k levels. The mapping $j \rightarrow k$ $(j \in \{1, 2, \ldots, n - i_1\}, k \in \{1, 2, \ldots, n\})$ is itself a Markov process from $j = n - i_1$ to $j = 1$, with the probability of a $j \rightarrow k$ mapping depending on the k level in the previous mapping. We leave the study of this Markov process for later and here investigate the unconditional expectation for the $j \rightarrow k$ mapping, which is defined as the expected value of k at level j in the subtree $(E_j(k))$ or

$$E_j(k) = \sum_{k>1} k\phi(k|j, n, n - i_1) \tag{18}$$

The expected time between the jth and $(j+1)$th levels in the subtree $(E(t_{j+1}))$ is

$$E(t_{j+1}) = \sum_{\ell = \underline{E}_j(k)+1}^{m} \frac{N_e}{\binom{\ell}{2}} \qquad \begin{cases} m = \underline{E}_{j+1}(k) + 1 & j + 1 < n \\ m = n & j + 1 = n \end{cases} \tag{19}$$

where $\underline{E}_j(k)$ is the largest integer less than or equal to $E_j(k)$. In contrast, the expected time between the jth and $(j+1)$th levels without drops is

$$\frac{N_e}{\binom{j+1}{2}}$$

from the standard coalescent.

Single unique dropout at level 3 ($Y_3 = 1$). As before, we assume that the sequenced individuals are in subclades A and B, as in Figure 3. The numbers of sequenced individuals in subclades A and B are a and b, respectively. We could use a similar approach as in section "Distribution of pairwise differences with dropouts" and condition on each subclade. This conditioning is justified when considering 2 sequenced lineages because either the 2 sequenced lineages occur within the same subclade with $j \rightarrow k$ mapping according to $\phi(k|j, n, n - i_1)$, or in separate subclades and by definition coalesce at $k = 2$. In contrast, when seeking to derive expectations for the properties of coalescence times for the full set of $n - i_1$ sequenced individuals, we need to account for the joint set of coalescence times in both subtrees. The Wiuf and Donnelly[23] theory does not include this case.

A simple argument indicates that the coalescence times are expected to be distorted with a dropout mutation versus no dropouts. For example, consider a period in the history of the sample in which there are a lineages in clade A and b lineages in clade B, and correspondingly $n - a - b$ unsequenced lineages. In the absence of recognizing that a dropout mutation has generated a structured coalescent process, the observed sample size is $a + b$, as opposed to n. Accordingly, the expected coalescence rate, given the observed sample, is $\binom{a+b}{2}$ on a timescale of N_e generations. Yet, the dropout mutation causes the coalescence process to be structured, such that the actual coalescence rate among sequenced individuals is

$$\binom{a}{2} + \binom{b}{2} \tag{20}$$

for $k \geqslant 3$ in the coalescent tree. Together, the coalescence rate is reduced by a factor of

$$\frac{\binom{a}{2} + \binom{b}{2}}{\binom{a+b}{2}} = \frac{a(a-1) + b(b-1)}{(a+b)(a+b-1)} \tag{21}$$

compared with what is expected for an observed sample size of $a + b$.

Neutrality at the primer sites and background selection at the amplified locus

If deleterious mutations occur at rate U and have multiplicative effects on fitness, each with effect s, then the expected frequency of individuals with j deleterious mutations (f_j) is Poisson distributed with mean $U/(2s)$.[25] For a sample of size n, the expected number of sampled individuals in the jth mutational class is nf_j, where the jth mutational class has j deleterious mutations. Following the structured coalescent modeling framework of Nordborg,[26] the class structure of deleterious mutations can be represented abstractly as a set $\{n_0, n_1, \ldots, n_{max(j:n_j>0)}\}$, where n_j is the number of individuals in the jth class and $max(j : n_j > 0)$ is the maximum value of j for which $n_j > 0$. Assuming no back mutation and a continuous-time model, transitions corresponding to deleterious mutations, such that $n_j \rightarrow n_j - 1$ and $n_{j-1} \rightarrow n_{j-1} + 1$, occur at rate U and transitions corresponding to coalescent events, such that $n_j \rightarrow n_j - 1$, occur at rate $(N_e f_j)^{-1}$. For N_e sufficiently large $U > > (N_e f_j)^{-1}$, mutational transitions occur on a fast timescale, whereas coalescent transitions occur on a slow timescale, such that the sample is expected to quickly transition to the 0-class. Once in the 0-class, the sample behaves in a neutral manner with coalescence rate $(N_e f_0)^{-1}$. Accordingly, background selection is expected to shorten coalescence times proportionally along a genealogical tree.[18]

In principle, we could just scale time as $(N_e f_0)^{-1}$ and measure dropout properties leading to an expectation for pairwise differences, but this would be somewhat inaccurate because it would miss coalescent events that occur during the transition from the initial sample to the 0-class. Although these events are rare, they lead to a lack of differences between sequences, allowing for average pairwise difference to be less than S/a_1, where S is the number of segregating sites and $a_1 = \sum_{k=1}^{n-1} \frac{1}{k}$.[27]

For a genotype in the jth class, the expected time to reach the 0th class is j/U units of time. During this time, lineages starting in the jth class may coalesce with other lineages, such that the probability that ℓ of the n_j sequences in the jth mutational class coalesce at or before reaching the 0-class ($p(\ell|n_j)$) is approximately

$$p(\ell|n_j) = \prod_{i=1}^{\ell} \frac{\binom{n_j - (i-1)}{2}}{\binom{n_j - (i-1)}{2} + \frac{N_e f_j U}{j}} \tag{22}$$

where $p(\ell|n_j)$ is an approximation because it assumes that only sequences starting within the same mutational class can coalesce. This assumption is relaxed later when the distribution of pairwise differences is derived. Allowing for sequences starting within different mutational classes to coalesce during the fast period from the initial sample to the 0-class could be modeled as a nested Markovian process. Nevertheless, $p(\ell|n_j)$ is an accurate approximation given that coalescent events are rare relative to mutation. The expected number of sequences that coalesce during the period of collapse to the 0th mutational class (n_c) is

$$n_c = \sum_j \sum_\ell \ell p(\ell|n_j) p(n_j) \tag{23}$$

where $p(n_j)$ is the probability of the jth mutational class. With n_c, $n_0 = n - n_c$.

Once we have n_0, we can use the theory for the neutral case to derive expectations for the number of descendants of dropouts ($p(\{i_1, i_2, \ldots, i_{Y_k}\}|Y_k)$, $p_k(\{i_1, i_2, \ldots, i_{Y_k}\})$, etc), as well as the distribution of pairwise differences. Expectations for the probability of a dropout along a branch section are a function of $N_e f_0$ with background selection compared with N_e without background selection.

Conditional distribution of pairwise differences with dropouts.

Single unique dropout at level 2 ($Y_2 = 1$). With background selection, the corresponding expressions for $p_{n-i_k}(j)$, $\phi(k|j, n, n - i_1)$, and $f_k(t)$ are the same, except for the replacement of n with n_0. We need to consider that a pair of sequences coalesce during the fast process in which sequences collapse to the 0th mutational class. For a sample of size n, the probability that 2 random sequences are in deleterious mutation classes j_1 and j_2 is $f_{j_1}^2$ for $j_1 = j_2$ and $2f_{j_1}f_{j_2}$ for $j_1 \neq j_2$.

For 2 sequences in the same class (j), the average time to reach the 0th class is $j/(U)$ units of time and the probability that they coalesce before reaching the 0-class is

$$\frac{(Nf_j)^{-1}}{(Nf_j)^{-1} + U/j} \tag{24}$$

For 2 sequences in different classes (j_1 and j_2), the probability that they come together in the ℓth mutational class is

$$\frac{(U/2)^{j_1-\ell+j_2-\ell}}{\sum_{i=0}^{j_2-1}(U/2)^{j_1+j_2-2i}} \tag{25}$$

and once in the ℓth class they can coalesce, such that the overall probability they coalesce is

$$\sum_{\ell=0}^{j_2-1}\frac{(U/2)^{j_1-\ell+j_2-\ell}}{\sum_{i=0}^{j_2-1}(U/2)^{j_1+j_2-2i}}\frac{(Nf_\ell)^{-1}}{(Nf_\ell)^{-1}+U/\ell} \tag{26}$$

Together, the probability of no pairwise difference ($z = 0$) due to a coalescent event during the fast process of collapse to the 0th mutational class $p_{n\to n_0}(z = 0)$ is

$$p_{n\to n_0}(z=0) = \sum_j f_j^2\frac{(Nf_j)^{-1}}{(Nf_j)^{-1}+U/j} + \sum_{j_2>j_1} 2f_{j_1}f_{j_2}\sum_{\ell=0}^{j_2-1} \frac{(U/2)^{j_1-\ell+j_2-\ell}}{\sum_{\ell=0}^{j_2-1}(U/2)^{j_1+j_2-2i}}\frac{(Nf_\ell)^{-1}}{(Nf_\ell)^{-1}+U/\ell} \tag{27}$$

After entering the 0-class, pairwise differences are expected to behave as $p(z)$ with n being replaced by n_0 and N_e by N_ef_0 such that $p(z)$ becomes $p'(z)$, and with background selection the distribution of pairwise differences ($p_b(z)$) is

$$p_b(z) = \begin{cases} p_{n\to n_0}(z=0) + (1-p_{n\to n_0}(z=0))p'(0), & z=0 \\ (1-p_{n\to n_0}(z=0))p'(z), & z>0 \end{cases} \tag{28}$$

Single unique dropout at level 3 (Y_3 = 1). Following similar logic as for $Y_2 = 1$ and assuming the topology in Figure 3, with background selection $p(x|A)$ becomes

$$p_b(z|A) = \begin{cases} p_{n\to n_0}(z=0) + (1-p_{n\to n_0}(z=0))p'(0|A), & z=0 \\ (1-p_{n\to n_0}(z=0))p'(z|A), & z>0 \end{cases} \tag{29}$$

where again $p'(z|A)$ is $p(z|A)$ with n substituted for n_0 and N_e for N_ef_0. Likewise, $p(z|B)$ becomes

$$p_b(z|B) = \begin{cases} p_{n\to n_0}(z=0) + (1-p_{n\to n_0}(z=0))p'(0|B), & z=0 \\ (1-p_{n\to n_0}(z=0))p'(z|B), & z>0 \end{cases} \tag{30}$$

and $p(z|A/B)$ becomes

$$p_b(z|A/B) = p'(z|A/B) \tag{31}$$

noting that the 2 sequences cannot coalesce during the fast time period when sampled in clades A and B.

The effect of dropouts on coalescence times of sequenced individuals. Once the sample reaches the 0th class mutationally, we expect the effects of dropouts on coalescence times to be the same as the neutral case, except for N_e being replaced by N_ef_0. In the neutral case, a signal of the effect of dropouts on coalesce times comes about by taking the ratio of adjacent coalescence times, which is independent of N_e. We therefore expect a similar effect of dropouts on adjacent coalescence times when N_e is replaced by N_ef_0 in the presence of background selection.

Analysis and Results

Expected values for θ

In the context of dropout mutations, θ is not expected to be large. As presented in the introduction, primer sites are typically around 20 nucleotides, with the highest probability that a mutation leads to a dropout occurring within 3 to 4 nucleotides of the 3′ end.[7] Accordingly, regarding mutations with a high probability of directly leading to a dropout, the mutation rate is near that of the baseline rate for a single nucleotide and at most about 10 times the baseline rate per nucleotide site. Estimates of per-site mutation rates in bacteria are typically 10^{-10} to 10^{-9}.[28] Nevertheless, a species or OTU may be predisposed to a dropout mutation due to prior mutational history within primer sites, leading to a higher dropout rate of mutation. θ is also a function of N_e. Less is known about N_e for microbes. Smith[29] estimated an effective size of 4×10^8 or less in *Neisseria meningitis*, based on electrophoretic data and noting that this is many orders of magnitude less than our understanding of census population size. By comparison in the eukaryotic microbial species *Saccharomyces paradoxus*, an estimate of its effective population size is of the order 10^6.[29] Focusing on effective population sizes in the range of 10^5 to 10^9 and mutation rates in the range of 10^{-10} to 10^{-8} to capture uncertainty leads to θ in the range of approximately 2×10^{-5} to 20. In the following section, we focus the analysis on θ values in the middle of this range, namely, of the order 10^{-2} to 10^{-1}.

Under neutrality

Probability of X_k branches with dropouts at level k, $p(x_k)$. Figure 4A plots $p(x_k)$ and indicates that the probability of a lineage with 1 dropout along branch sections that are at lower levels in the coalescent tree is not too uncommon in the context of studies involving thousands of species or OTUs and assuming 2 values of θ (0.1, 0.01). Two lineages with dropouts are rarer, but still appreciable at lower levels of the coalescent tree, provided that θ is large enough. Three or more lineages with dropout mutations are fairly rare in the context of studies involving thousands of species or OTUs. Although higher levels in a coalescent tree have more branch sections, these sections are short in length, such that the total probability of a dropout mutation along levels 500 to 1000 for a sample of size $n = 1000$ is less than that for level $k = 2$ by itself and levels 4 through 50 together (Figure 4B). The θ values of the order 10^{-2} to 10^{-1} lead to the $p(X_k > 0)$ values between 0.010 and 0.091 at level $k = 2$, 0.005 and 0.047 at level $k = 3$, and 0.003 and 0.032 at level $k = 4$. In total, for $k \geqslant 500$, $p(X_k > 0)$ equals 0.007 and 0.070 for $\theta = 0.01$ and $\theta = 0.1$, respectively.

Probability Y_k of the X_k dropouts at level k are unique along their corresponding lines of descent, $p(Y_k = j|X_k, t)$ and $p(Y_k = j|X_k)$. Expressions for $p(Y_k = j|X_k, t)$ and $p(Y_k = j|X_k)$ (see section "Probability Y_k of the X_k dropouts at level k are unique along their corresponding lines of descent") work for any X_k, but given a low dropout mutation rate it is likely that only 1 dropout occurs at level k, if at all. Multiplying $p(Y_k = 1|X_k = 1)$ by $p(X_k = 1)$ gives the joint probability of $Y_k = 1$ and $X_k = 1$ in the limit as $t \to \infty$. Figure 5 indicates that lower levels of a coalescent tree are expected to have a unique dropout mutation with the highest probability, and that for $k \geqslant 500$ the probability that at least 1 level has a unique mutation is still less than level $k = 2$.

Although single dropouts occur with the highest probability, to illustrate the utility of $p(Y_k = j|X_k, t)$ and $p(Y_k = j|X_k)$, we consider the case when a dropout occurs at level $k = 4$ and assume 2 dropouts at that level ($X_k = 2$). For this case, the state space corresponding to $\{x'_{k,1}, x'_{k,2}, \ldots x'_{k,X_k}, k'\}$ (see section "Probability Y_k of the X_k dropouts at level k are unique along their corresponding lines of descent") is the set $\{\{2,0,4\}, \{2,0,3\}, \{1,0,3\}, \{0,1,3\}, \{2,0,2\}, \{1,0,2\}, \{0,1,2\}, \{0,0,2\}, \{0,1,1\}, \{1,0,1\}, \{0,0,1\}\}$. The corresponding $\mathbf{A}$ matrix is

$$\mathbf{A} = \begin{pmatrix} -6 & 5 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -(3+\theta) & 0 & 0 & 2 & \theta & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -(3+\frac{\theta}{2}) & 0 & 0 & 3 & 0 & \frac{\theta}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & -(3+\frac{\theta}{2}) & 0 & 0 & 3 & \frac{\theta}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -(1+\theta) & 0 & 0 & 0 & 1 & \theta & 0 \\ 0 & 0 & 0 & 0 & 0 & -(1+\frac{\theta}{2}) & 0 & 0 & 0 & 1 & \dfrac{\theta}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & -(1+\frac{\theta}{2}) & 0 & 1 & 0 & \dfrac{\theta}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and whereas $p(Y_k = j|X_k, t)$ is a somewhat complicated function of θ and t, $p(Y_k = j|X_k)$ is more easily written down, such that

$$p(Y_4 = 2|X_4 = 2) = \frac{108 + 99\theta + 16\theta^2}{3(1+\theta)(2+\theta)(3+\theta)(6+\theta)}$$
$$p(Y_4 = 1|X_4 = 2) = \frac{5\theta(3 + 2\theta)}{3(1+\theta)(2+\theta)(3+\theta)}$$
$$p(Y_4 = 0|X_4 = 2) = \frac{\theta(27 + 23\theta + 3\theta^2)}{3(2+\theta)(3+\theta)(6+\theta)^2}$$

and, numerically, for $\theta = 0.1$

$$p(Y_4 = 2|X_4 = 2) = 0.901$$
$$p(Y_4 = 1|X_4 = 2) = 0.074$$
$$p(Y_4 = 0|Y_4 = 2) = 0.025$$

indicating that there is a high probability given 2 dropout mutations at level $k = 4$ that they are both unique.

Probability of d or more descendent lineages of a unique dropout mutation at level k, $p_k(i_1 \geqslant d)$. There is a fairly good chance of a unique dropout at level 2 and that it leads to the loss of at least 100 individuals (~8%; Figure 6) for $\theta = 0.1$ and $n = 1000$. There is a smaller, but appreciable, chance under the same conditions at level 4 (~2%). Figure 6 assumes a single dropout mutation at a given level. To illustrate calculations for the case of 2 dropout mutations at level $k = 4$

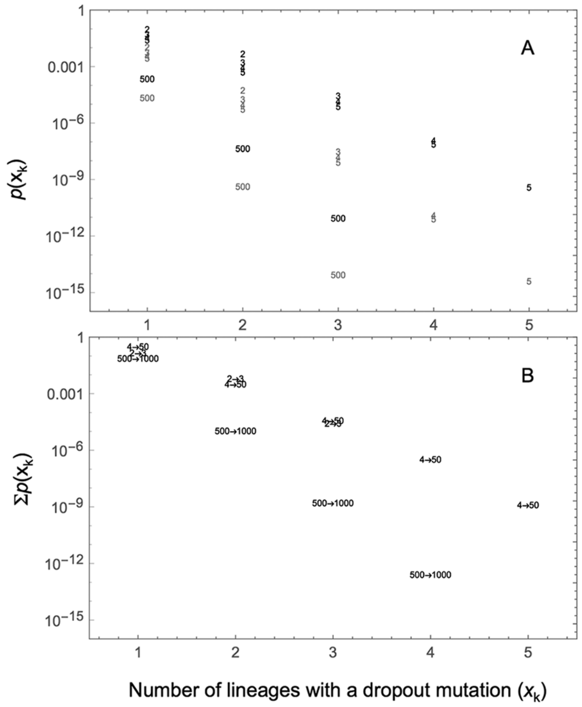

Figure 4. (A) The probability of the number of lineages with a dropout mutation ($p(x_k)$) at different levels in a coalescent tree. Points in the plot are represented by numbers in which the number corresponds to the k level in the coalescent tree of a sample. Black numbers correspond to $\theta = 0.1$ and gray numbers to $\theta = 0.01$. (B) The sum of $p(x_k)$ across levels for $\theta = 0.1$. Points in the plot indicate the levels that are summed. For example, 500 → 1000 sums $p(x_k)$ for $k = 500$ to $k = 1000$.

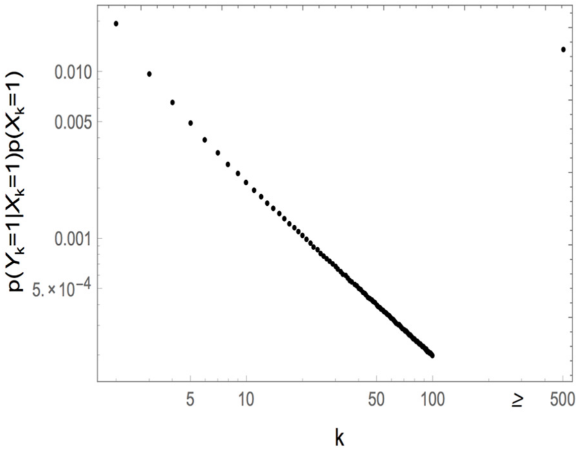

Figure 5. The combined probability of 1 dropout mutation at level k and it being unique ($p(Y_k = 1|X_k = 1)p(X_k = 1)$). For levels greater than or equal to 500, the point is the sum of at least 1 dropout at each level up to $k = 1000$ and it being unique $\theta = 0.02$.

$$p_4(\{i_1\}) = p(\{i_1\}|Y_4 = 1)p(Y_4 = 1|X_4 = 2)p(X_4 = 2) = \frac{30\theta^3(3+2\theta)(n-i_1-2)(n-i_1-1)}{(n-3)(n-2)(n-1)(1+\theta)(2+\theta)(3+\theta)^2(4+\theta)(6+\theta)}$$

$$p_4(\{i_1, i_2\}) = p(\{i_1, i_2\}|Y_4 = 2)p(Y_4 = 2|X_4 = 2)p(X_4 = 2) = \frac{12\theta^2(108+99\theta+16\theta^2)(n-i_1-i_2-1)}{(n-3)(n-2)(n-1)(1+\theta)(2+\theta)(3+\theta)^2(4+\theta)(6+\theta)^2}$$

$$p_4(i_1 \geqslant d) = \frac{10\theta^3(3+2\theta)(n-4-d)(6+n+n^2-2nd+d(d-1))}{(n-3)(n-2)(n-1)(1+\theta)(2+\theta)(3+\theta)^2(4+\theta)(6+\theta)}$$

$$p_4(i_1 + i_2 \geqslant d) = \frac{6\theta^2(108+99\theta+16\theta^2)(n-4-d)(n+3-d)}{(n-3)(n-2)(n-1)(1+\theta)(2+\theta)(3+\theta)^2(4+\theta)(6+\theta)^2}$$

As $Y_4 = 2$ is more likely than $Y_4 = 1$ when $X_4 = 2$, the probability of d or more dropout individuals is greater for $Y_4 = 2$ versus $Y_4 = 1$ (Figure 7). For $\theta = 0.1$, it is rare (~0.004%) to simultaneously have 2 unique dropout mutations at level 4 and the loss of at least 100 individuals in a sample of size 1000. Here, the rarity primarily comes about from the rarity of the occurrence of 2 dropout mutations at level 4 (cf. Figure 4A).

Distribution of pairwise differences with dropouts and ratio of coalescent times between adjacent nodes. The distribution of pairwise differences is distorted when there is a single dropout at level 2 (Figure 8A). When the number of sequenced individuals is small relative to the entire sample, the distribution of pairwise difference is shifted toward smaller values because sequenced individuals must coalesce before the dropout, which shortens coalesce times. For a moderate to large number of sequenced individuals relative to the entire sample, the distribution can have a mode, unlike the case without dropouts.

With a dropout leading to 2 clades of sequenced individuals, such that $Y_3 = 1$ (ie, Figure 3), the distribution of pairwise differences is distorted further relative to the $Y_2 = 1$ case, such that coalescences restricted to be within A or B result in relatively few pairwise differences and coalescences restricted to be between clades A and B result in relatively large pairwise differences (Figure 8B).

The ratio of coalescent times between adjacent nodes indicates that it is lower at levels near the base of the tree in the presence of a dropout (Figure 9).

With background selection

Expected values for θ. Background selection is expected to decrease the effective value of θ by a factor of f_0, which is the expected proportion of sequences free of deleterious mutation as a result of background selection.[17] Furthermore, Charlesworth et al[18] showed that the average pairwise

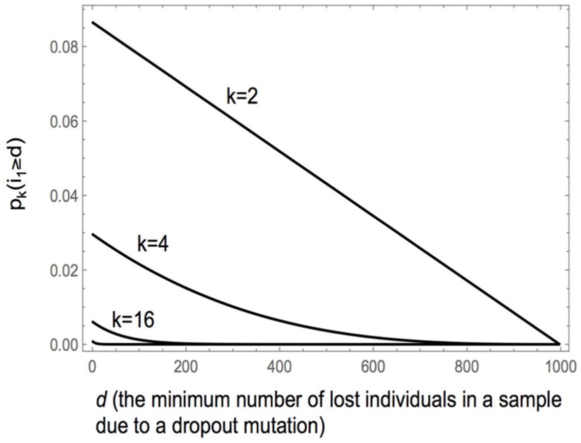

Figure 6. The probability of a dropout out at level k leading to the loss of d or more individuals in a sample ($p_k(i_1 \geqslant d)$) assuming $\theta = 0.1$ and $n = 1000$. The bottom curve corresponds to $k = 128$.

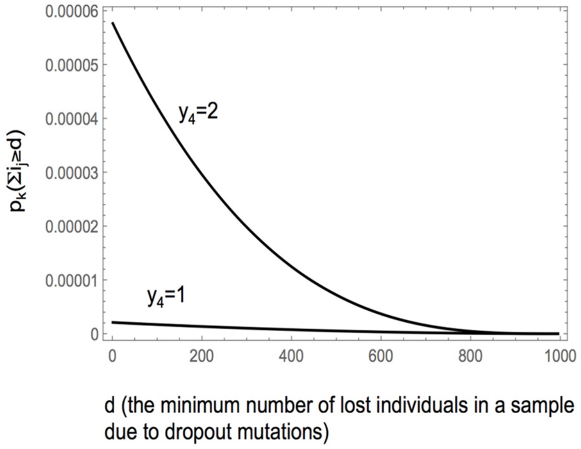

Figure 7. The probability of 2 dropout mutations at level $k = 4$ leading to the loss of d or more individuals in a sample assuming $\theta = 0.1$ and $n = 1000$.

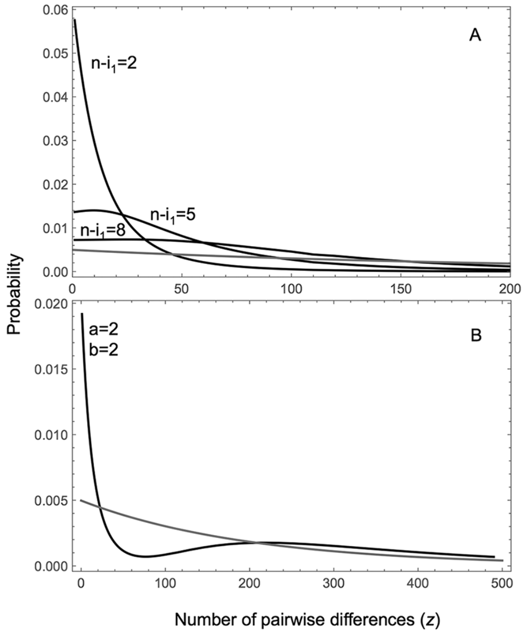

Figure 8. The distribution of pairwise differences with (A) 1 dropout at level 2 leading to a single clade of sequenced individuals and (B) 1 dropout mutation at level 3, leading to 2 clades of sequenced individuals with the topology shown in Figure 3. The gray curve shows the distribution in the absence of dropouts. The parameters are $2N_e\nu = 100$ and $n = 10$ for both plots. The number of sequenced individuals is shown in the plots. Note that the shape of the distribution is qualitatively independent of sample size, such that we expect a similar pattern for $n > 10$ and the corresponding increases in a and b.

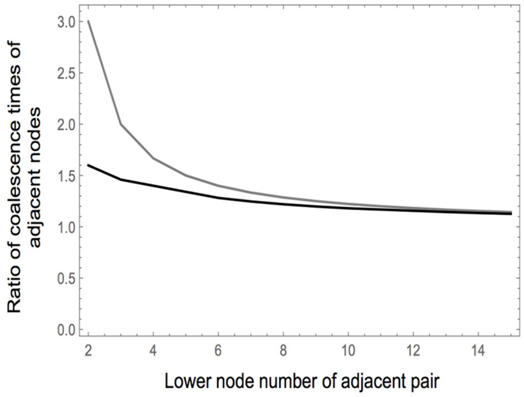

Figure 9. The ratio of coalescence times between adjacent nodes in the absence of a dropout mutation (gray) and in the presence of a dropout at level $k = 2$ (black). The parameters are $n = 200$ and $n - i_1 = 50$. Note that the ratios are independent of N_e.

difference is accurately approximated by a neutral model with θ multiplied by a factor of f_0, where f_0 is of the order 10^{-2} to 10^{-1} for strong and moderate background selection.

Properties of samples with dropout mutations and background selection. Let us first revisit $p(x_k)$ in the context of background selection for a species with effective population size 10^8. With background selection, such that $U = 0.1$ and $s = 0.02$ (following Charlesworth et al[18]), $f_0 = 0.08$ and $N_e f_0 = 8 \times 10^6$. For mutation rates in the range of 10^{-10} to 10^{-8}, the corresponding values of θ ($0.001 \leqslant \theta \leqslant 0.1$) encompass the range of values explored in Figure 4, which were $\theta = 0.01$ and $\theta = 0.1$. As an additional point of comparison, for a smaller value of θ ($\theta = 0.001$), $p(X_2 = 1) = 0.001$, which indicates that segregating dropouts would be rare for metagenomic studies with about 1000 species under these conditions. Nevertheless, for smaller values of U and/or s, $p(X_2 = 1)$ is expected to increase.

Next, consider n_0 for $n = 1000$, $N_e = 10^8$, $U = 0.1$, and $s = 0.02$. The expected number of sequences that coalesce is $n_c = 0.12$ prior to collapse of the sample to the 0th mutational class, such that we expect $n_0 \approx n$. As a point of comparison when N_e is smaller ($N_e = 10^6$), and n, U, and s remain the same, the number of sequences that coalesce is $n_c = 36$. Overall, this suggests that $n_0 \approx n$ with background selection and large effective population sizes.

The analysis involving n_0 suggests that, for large N_e relative to $1/U$, the probability of no pairwise difference due to a coalescent event during the fast process of collapse to the 0th mutational class ($p_{n\rightarrow n_0}(z=0)$) is small because coalescence events are rare relative to the mutational collapse process to the 0th class. Numerically, for the conditions $U=0.1$ and $s=0.02$ and for a large sample size, $p_{n\rightarrow n_0}(z=0)=5\times10^{-6}$ for $N_e=10^8$ and $p_{n\rightarrow n_0}(z=0)=0.003$ for $N_e=10^5$.

As for large N_e, $p_{n\rightarrow n_0}(z=0)\approx 0$, we expect only a small difference between the distribution of pairwise difference with background selection ($p_b(z)$) versus without background selection ($p(z)$), as well as between $p_b(z|A)$, $p_b(z|B)$, and $p_b(z|A/B)$ and $p(z|A)$, $p(z|B)$, and $p(z|A/B)$, respectively. Therefore, Figure 8 presents an accurate approximation of the distribution of pairwise differences under background selection for $2N_ef_0\nu=100$.

Discussion

Bacterial and other microbial species are associated with populations of large effective size. This in combination with metagenomic and/or eDNA studies of microbial communities that include thousands of species or OTUs[9] brings about the possibility that one or more of the sampled species or OTUs may be segregating mutations that cause dropouts at PCR primer sites. This article quantified the probability of a segregating dropout mutation at a primer site, as well as the effects of dropouts on the distribution of pairwise differences and coalescence times within a species or OTU.

The analysis indicates that segregating PCR dropout mutations are reasonably common provided that the effective population size of a species is large and/or the dropout mutation rate is reasonably high. Furthermore, if dropout mutations occur, they are expected to occur along basal branch sections in the coalescent tree of a sample. That dropouts are expected to occur along the basal branch section brings about the possibility that a large fraction of sampled individuals go unsequenced due to a dropout mutation within a species. This fraction of individuals that go unsequenced leaves behind a signal in the pattern of pairwise differences, as well as the coalescence times among sequenced individuals. In particular, there is a greater tendency for the distribution of pairwise differences to have a non-zero mode in its distribution, with a mode being particularly pronounced when a dropout mutation gives rise to 2 subclades of sequenced individuals as depicted in Figure 3. Furthermore, a dropout mutation is expected to distort coalescence times near the base of a coalescent tree, such that coalescence times are more equal in value relative to when dropouts are absent. The distortion of coalescence times is independent of n (the true sample size) in the context of equation (21). In other contexts, such as pairwise differences, the effect of dropouts is qualitatively the same across n. Here, inference approaches could integrate across possible values of n to assess whether a distribution is consistent with the presence of one or more dropout mutations.

When a dropout mutation gives rise to 2 clades of sequenced individuals as in Figure 3, the coalescent process for these 2 clades is structured such that individuals between clades cannot coalesce until after the dropout mutation (backwards in time). Qualitatively, the distribution of pairwise differences is similar to that when there is population structure,[25] in which the distribution is a combination of small pairwise differences between individuals within the same clade and large pairwise differences between individuals in different clades. When a single dropout mutation occurs at level 2 in a coalescent tree, its effect on the distribution of pairwise differences is less pronounced than a level 3 dropout mutation that gives rise to 2 clades, such that there is a tendency for the distribution to have a non-zero mode.

A future direction for research is to develop statistical techniques to assess whether a metagenomic or eDNA sample has segregating dropout mutations. Assessing whether a dropout mutation occurs in a single species will be challenging because the pattern that is generated is similar to population structure[26] or changes in population size.[31] Furthermore, the coalescence process in a panmictic population of constant size gives rise to a highly diverse set of distributions of pairwise differences and coalesce times for a given sample size. It may be more promising to first develop methods that seek to assess whether there is a signal of dropouts at the population level across a clade of species or OTUs in a metagenomic study. A consistent signal of dropouts within a clade may lead to greater confidence in the presence of dropouts. A clade of species or OTUs may be predisposed to dropout mutations due to prior substitutions at the primer sites in the ancestor to the clade. We think that the EMBL Metagenomics database[9] provides an ideal resource to begin to assess the potential occurrence of PCR dropout mutations. The database consists of thousands of studies and often within a study there are multiple samples from a site, each of which is sequenced using an amplicon approach. Having multiple samples from a site would allow for the assessment of whether a signature of PCR dropouts, perhaps across species or OTUs within a clade, is replicated across samples. If it is, then this would be evidence against the random occurrence of a pattern that is similar to predictions made in this study.

A consistent signal across samples could nevertheless be due to population structure or changes in population size, instead of PCR dropouts. Although qualitatively there are similarities in patterns of DNA sequence variation, quantitatively there are likely differences that can distinguish PCR dropouts from the population structure and changes in population size. For example, polymorphisms within clades of a coalescent tree are expected to be at different frequencies geographically because of partial isolation, drift, and the history of mutation. In contrast at a single geographic site and with no population structure at a larger spatial scale, we would expect more equal frequencies of polymorphisms, on average, across clades of a coalescent tree in the presence of PCR dropouts. Population growth may be distinguished from PCR dropouts in that recent growth is expected to increase the number of rare polymorphisms in a population, in contrast to PCR dropouts. Overall, the detection of PCR dropouts will likely be quantitative and involve the development of a model-based inference method with the processes of PCR dropout mutations, population structure, and changes in population size co-occurring. Assessment of the occurrence of PCR dropouts would then be probabilistic, whereby the likelihood or posterior probability of occurrence of a set of PCR dropouts would be weighed in combination with parameters that model population structure and demography.

The detection of the presence of dropout mutations may be useful in metagenomic studies that seek to not only identify taxon-level diversity, but also the abundances of taxa in a community.[32] If a taxon has a signature of dropout mutations, then its abundance estimate may be lower than its true value.

Regarding existing methods that use next-generation sequencing approaches to the estimate nucleotide diversity[14,15] and phylogenetic structure,[17] our results indicate that PCR dropouts are a factor of consideration. In the case of nucleotide diversity, dropouts give rise to patterns of DNA sequence variation similar to population structure or changes in population size (see above) and therefore are expected to affect estimates of nucleotide diversity. The process of PCR dropout mutations could be added to coalescent-based estimators of population-level statistics, particularly in the context of the use of next-generation sequencing and environmental samples. Likewise, when inferring the phylogenetic structure of environmental samples that use primer-based sequencing, the addition of the process of mutation at primer sites may be warranted, as these mutations may lead to DNA sequence patterns that suggest demographic processes such as population structure and growth. Similar to the previous discussion of detecting PCR dropouts, the incorporation of dropouts to estimators of statistics and phylogenetic structure will likely involve adding the dropout process to model-based approaches, which would then affect the likelihood or posterior distributions of focal statistics or histories.

This article focused on properties of segregating dropout mutations within a species. The development of more powerful methods to detect dropouts will likely involve combining within-species and between-species approaches. In the context of between-species approaches, a promising theoretical framework is Hey's[14] model of cladogenesis and its potential generalization. In this article, we showed that dropout mutations affect the shape of coalescent trees and it is likely that they also affect the shape of phylogenetic trees. Properties of tree shape at the phylogenetic level under different diversification processes has been a topic of ongoing research, with detectable differences in diversification processes arising from differences in internode lengths in a phylogenetic tree.[33] Differences in internode lengths phylogenetically are analogous to the effect of dropouts on the ratio of coalescence times. A strong signal for the presence of dropouts would be a clade with a distorted tree shape at the species level combined with distorted coalescence times within species within the clade. Of course, this type of inference requires segregating dropouts within species and polymorphism for dropouts at the clade level, whereas a more common pattern may be that an entire clade is not amplified by PCR due to ancestral dropout mutations. Furthermore, it is important to note that expectations about where dropout mutations occur genealogically at the population or species level using the coalescent may not hold at higher taxonomic levels due to differences in the expected distribution of relative internode lengths in a phylogenetic versus coalescent model.

Regarding sequencing methods, a solution to the occurrence of PCR dropout mutations is to use complementary non-overlapping primer sets that both amplify the same genetic locus.[34] Using this approach, a locus in a species' DNA will be amplified even if there are dropout mutations at a primer binding site. Nevertheless, this approach is not always used and there exist a large number of datasets that did not use complementary non-overlapping primer sets (see Mitchell et al).[9] For these studies, approaches to assess coverage would be useful.

This article is a starting point to derive expectations for the effects of PCR efficiencies on the inference of taxon diversity and within-taxon genetic diversity in other contexts. For example, a common approach to biodiversity assessment is DNA metabarcoding.[35] DNA metabarcoding is often applied to organisms with smaller effective population sizes than prokaryotes, such that the population-level rate of mutation at primer sites is lower. Accordingly, PCR dropout mutations are expected to be less common in these types of studies, but other factors such as differential PCR inhibition or amplification bias may result in the dropout of individual sequences.[36] Generally, accounting for PCR efficiency may be important in studies that seek to infer population-level genetic diversity from metabarcode data.[37]

Coalescent theory has been applied in metagenomic studies previously. For example, O'Brien et al[17] derived a coalescent-based approach to infer the phylogenetic relationships of species from a metagenomic sample, and Bittner et al[38] and Liberles et al[39] provide a broad overview for integrating coalescent approaches into metagenomic studies. Furthermore, Johnson and Slatkin[40] presented a coalescent-based approach to detect recombination in a metagenomic sample. Our article suggests that coalescent theory may allow for the detection of PCR primer dropouts in metagenomic studies.

Conclusions

Primer coverage in microbial metagenomic and eDNA studies is a key uncertainty. In this article, we derived coalescent-based expectations for the pattern of DNA sequence variation within a species in the presence of segregating PCR dropout mutations, where a dropout mutation completely blocks PCR amplification. Dropout mutations can alter coalescence times and the distribution of pairwise differences, which may form the basis of statistical techniques to detect or account for reduced primer coverage.

Acknowledgements

I wish to thank C Somerville for her assistance developing this line of research and two reviewers for their comments that improved this work, as well as the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant program for their support.

Author Contributions

CKG conceived and derived the model, performed its analysis and wrote the paper.

ORCID iD

Cortland K Griswold https://orcid.org/0000-0003-2993-7043

REFERENCES

1. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol*. 1998;180:4765-4774.
2. Tyson GW, Chapman J, Hugenholtz P, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. 2004;428:37-43.
3. Venter JC, Remington K, Heidelberg JF, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 2004;304:66-74.
4. Alberts AB, Johnson A, Lewis J, et al. *Molecular Biology of the Cell*. 4th ed. New York, NY: Garland Science; 2002.
5. Bru D, Martin-Laurent F, Philippot L. Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Appl Environ Microb*. 2008;74:1660-1663.
6. Wu JH, Hong PY, Liu WT. Quantitative effects of position and type of single mismatch on single base primer extension. *J Microbiol Meth*. 2009;77:267-275.
7. Mao DP, Zhou Q, Chen CY, Quan ZX. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC Microbiol*. 2012;12:66.
8. Klindworth A, Pruesse E, Schweer, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*. 2012;41:e1.
9. Mitchell AL, Scheremetjew M, Denise H, et al. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Research*. 2018;46:D726-D735.
10. Santamaria M, Fosso B, Consiglio A, et al. Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform*. 2012; 13: 682–695.
11. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol*. 2010;6:e1000667.
12. Kingman JFC. On the genealogy of large populations. *J Appl Prob A*. 1982;19:27-43.
13. Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*. 1983;105:437-460.
14. Hey J. Using phylogenetic trees to study speciation and extinction. *Evolution*. 1992;46:627-640.
15. Liu X, Fu YX, Maxwell TJ, Boerwinkle E. Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. *Genome Res*. 2010;20:101-109.
16. Liu X. jPopGen Suite: population genetic analysis of DNA polymorphism from nucleotide sequences with errors. *Method Ecol Evol*. 2012;3:624-627.
17. O'Brien J, Didelot X, Iqbal Z, Amenga-Etego L, Ahiska B, Falush D. A Bayesian approach to inferring the phylogenetic structure of communities from metagenomic data. *Genetics*. 2014;197:925-937.
18. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral variation. *Genetics*. 1993;134:1289-1303.
19. Charlesworth D, Charlesworth B, Morgan MT. The pattern of neutral molecular variation under the background selection model. *Genetics*. 1995;141:1619-1632.
20. Fu YX. Statistical properties of segregating sites. *Theor Popul Biol*. 1995;48:172-197.
21. Wakeley J. *Coalescent Theory*. Greenwood Village, CO: Roberts; 2008.
22. Bharucha-Reid AT. *Elements of the Theory of Markov Processes and Their Applications*. New York, NY: Dover; 1997.
23. Wiuf C, Donnelly P. Conditional genealogies and the age of a neutral mutant. *Theor Popul Biol*. 1999;56:183-201.
24. Ross SM. *Introduction to Probability Models*. Toronto, ON, Canada: Academic Press; 1997.
25. Kimura M, Maruyama T. The mutational load with epistatic gene interactions in fitness. *Genetics*. 1996;54:1337-1351.
26. Nordborg M. Structured coalescent process on different time scales. *Genetics*. 1997;146:1501-1514.
27. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989;123:585-595.
28. Ford CB, Lin PL, Chase MR, et al. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nature Genet*. 2011;43:482-486.
29. Smith JM. The population genetics of bacteria. *Proceed Roy Soc Biol Sci*. 1991;245:37-41.
30. Masel J, Griswold CK. The strength of selection against the yeast prion [PSI +]. *Genetics*. 2009;181:1057-1063.
31. Rogers AR, Harpending HC. Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol*. 1992;9:552-569.
32. He JZ, Shen JP, Zhang LM, et al. Quantitative analyses of the abundance and composition of ammonia-oxidizing bacteria and ammonia-oxidizing archaea of a Chinese upland red soil under long-term fertilization practices. *Environ Microbiol*. 2007;9:2364-2374.
33. Mooers AØ, Harmon LJ, Blum MGB, Wong DHJ, Heard SB. Some models of phylogenetic tree shape. In: Gascuel O, ed. *Reconstructing Evolution: New Mathematical and Computational Advances*. Oxford, UK: Oxford University Press; 2007:149-170.
34. Blais J, Lavoie SB, Giroux S, et al. Risk of misdiagnosis due to allele dropout and false-positive PCR artifacts in molecular diagnostics: analysis of 30,769 genotypes. *J Mol Diagnost*. 2015;17:505-514.
35. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol Ecol*. 2012;21:2045-2050.
36. Piñol J, Mir G, Gomez-Polo P, Agustí N. Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Mol Ecol Res*. 2015;15:819-830.
37. Elbrecht V, Vamos EE, Steinke D, Leese F. Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*. 2018;6:e4644.
38. Bittner L, Halary S, Payri C, et al. Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. *Biol Direct*. 2010;5:47.
39. Liberles DA, Teufel AI, Liu L, Stadler T. On the need for mechanistic models in computational genomics and metagenomics. *Genome Biol Evol*. 2013;5:2008-2018.
40. Johnson PLF, Slatkin M. Inference of microbial recombination rates from metagenomic data. *PLoS Genet*. 2009;5:e1000674.