The Annals of Applied Probability
2010, Vol. 20, No. 5, 1567–1606
DOI: 10.1214/09-AAP657

THE DIVERSITY OF A DISTRIBUTED GENOME IN BACTERIAL POPULATIONS[1]

BY F. BAUMDICKER, W. R. HESS AND P. PFAFFELHUBER

Albert-Ludwigs University

The distributed genome hypothesis states that the set of genes in a population of bacteria is distributed over all individuals that belong to the specific taxon. It implies that certain genes can be gained and lost from generation to generation. We use the random genealogy given by a Kingman coalescent in order to superimpose events of gene gain and loss along ancestral lines. Gene gains occur at a constant rate along ancestral lines. We assume that gained genes have never been present in the population before. Gene losses occur at a rate proportional to the number of genes present along the ancestral line. In this *infinitely many genes model* we derive moments for several statistics within a sample: the average number of genes per individual, the average number of genes differing between individuals, the number of incongruent pairs of genes, the total number of different genes in the sample and the gene frequency spectrum. We demonstrate that the model gives a reasonable fit with gene frequency data from marine cyanobacteria.

1. Introduction. Population genetics is dealing with biological diversity of species. Concepts developed in this area include models for genetic drift, mutation, selection, recombination and population structure. These models are applied frequently to eukaryotic species to analyze their evolutionary history. For prokaryotes these concepts are applied less frequently and Maynard-Smith (1995) even asked, "*Do bacteria have population genetics*?"

Usually (by the *biological species concept*), a species is a reproductively isolated set of individuals. This definition can hardly be applied to prokaryotes, that is, bacteria and archea. In microbiology, researchers have developed other approaches, mostly defining a species via genomic similarity. This similarity is either based on hybridization of DNA, or on DNA sequences of specific molecules such as ribosomal RNA or suitable housekeeping genes, known to be highly conserved (and which are identified by genome sequencing or by a technique called *multilocus sequence typing* [Maiden et al. (1998)]). The definition of bacterial species is complicated by the fact that the similarity of bacteria depends on the considered genomic region, which can be explained by transfer of genetic material be-

Received July 2009; revised October 2009.

[1] Supported by the BMBF through the Freiburg Initiative for Systems Biology Grant 0313921.

AMS 2000 subject classifications. Primary 92D15, 60J70, 92D20; secondary 60K35.

Key words and phrases. Kingman's coalescent, infinitely many genes model, infinitely many sites model, gene content.

tween these bacteria [Dykhuizen and Green (1991)]. Even more extreme, individuals from the same species carry different genes. For example, one quarter of the genome of a pathogenic variant of *E. coli* was found to be absent in a laboratory strain [Perna et al. (2001)]. Such findings lead to several new hypotheses: the *core genome hypothesis* argues that the set of genes common to all bacteria of a species is responsible for maintaining species-specific phenotypic properties [e.g., Riley and Lizotte-Waniewski (2009)]. The *distributed genome hypothesis* predicts that no single individual comprises the full set of genes of the bacterial population [e.g., Ehrlich et al. (2005)].

The distributed genome hypothesis is similar to the idea of bacterial *pangenomes*. Taking the different gene content of individuals of a population into account, the pangenome consists of all different genes carried by all individuals. The pangenome can be split into the *core genome*, that is, the set of genes carried by every individual of the population, and the *dispensable* (also: *auxiliary* or *flexible* or *contingency*) *genome* [Medini et al. (2005)]. The pangenome was first analyzed for pathogenic strains of *Streptococcus agalactiae* [Tettelin et al. (2005)]. It was shown that around 80% of a single genome (i.e., the genome of a single individual) forms the core genome. However, each fully sequenced genome carries genes which do not occur in other strains, suggesting that the core genome represents only a small fraction of the pangenome. The situation is even more extreme in *Prochlorococcus* and *Synechococcus*, which are marine cyanobacteria, where the core genome consists of around 60% of the genes found in a single genome [Kettler et al. (2007), Dufresne et al. (2008)]. In contrast, Medini et al. (2005) show that a set of four genomes of *Bacillus anthracis* contain all genes found in the complete sample of 8 individuals, showing that the core genome is the biggest part of the pangenome of this species. Recently, the pangenome of all bacteria was considered, using a dataset of 573 completely sequenced genomes showing that only 250 genes (which are 8% of a bacterial genome on average) were common to almost all bacterial species [Lapierre and Gogarten (2009), Bentley (2009)].

The bacterial *supragenome* makes the split into the core and dispensable genome more precise: each gene present in a population (or in a sample) has a frequency (for core genes this is 100%) such that the pangenome gives rise to a gene frequency spectrum. The first analysis of Hogg et al. (2007) on a sample of 13 genomes of *Haemophilus influenzae* shows that the largest class (19% of the pangenome) in the dispensable genome are genes only present in a single genome. In addition, every pair of genomes differs by around 300 genes on average. Similar findings were obtained for *Streptococcus pneumoniae* [Hiller et al. (2007)].

The pan- and supragenome suggest that genes can be gained and lost along lineages of bacteria, leading to diversity of genomes. It is well known that genes can be gained in bacteria by three different mechanisms: (i) The uptake of genetic material from the environment is referred to as *transformation*. (ii) Bacteria can be infected by lysogenic phages which provide additional genetic material that can be built in the bacterial genome. This process is known as *transduction*. (iii) A direct link between two bacterial cells of the same species leads to exchange of genetic material, known as *conjugation*. These three mechanisms are usually referred to as *horizontal gene flow*. Events of gene loss occur by mutations resulting in pseudogenization or deletion of genes.

The aim of the present paper is to model the bacterial pangenome. We focus on two different aspects: the genealogical relationships between individuals and the mutational mechanism. Using the diffusion limit of a standard neutral model (with finite offspring variance) leads to a random genealogy, usually referred to as the Kingman coalescent [Kingman (1982), Wakeley (2008)]. Gene gain and loss is the basis of our mutational model, as introduced by Huson and Steel (2004) in the phylogenetics literature. Here, new genes are taken up from the environment at constant rate along ancestral lines. We assume that all genes taken up are different. In addition, present genes can be lost at constant rate. In analogy to standard population genetic models we refer to this as the *infinitely many genes model*.

2. Model.

The dynamics of our model consist of two parts. Reproduction follows the (diffusion limit of a) neutral Wright–Fisher model (or some other exchangeable population genetic model with finite offspring variance). The mutation model we use is borrowed from the phylogenetics literature [Huson and Steel (2004)] and describes gene gains and losses along ancestral lines. After introducing the model in Sections 2.1 and 2.2, we discuss connections to other mutation models in Section 2.3.

2.1. *Reproduction dynamics.*

We will use the neutral Wright–Fisher model: a panmictic population of size N reproduces neutrally and clonally, that is, asexually. In this model, individuals in generation $t+1$ choose a unique parent from generation t purely at random and independent of other individuals. It is well known that the genealogy of a sample of size n taken from the Wright–Fisher model converges for $N \to \infty$ after a time rescaling by N to the Kingman coalescent started with n lines [e.g., Durrett (2008), Wakeley (2008)]. In this process, starting with n lines:

- if there are k lines left, draw an exponential time with rate $\binom{k}{2}$ (which equals the number of pairs in the k lines), which is the time to the next coalescence event;
- at the next coalescence event pick two lines at random from the k lines and merge these into one line.

If there is one line left, the sample has found its most recent common ancestor which we trace back into the past for an infinite amount of time.

DEFINITION 2.1 (Kingman coalescent). We denote the random tree resulting from the above mechanism—the Kingman coalescent—by $\mathcal{T}$. We consider $\mathcal{T}$ as a partially ordered metric space with order relation $\preceq$ and metric $d_{\mathcal{T}}$ where the distance of two points in $\mathcal{T}$ is given by the sum of the times to their most recent common ancestor. We make the convention that $s \preceq t$ for $s, t \in \mathcal{T}$ if s is an ancestor of t.

We note that our starting point, the Wright–Fisher model, can be replaced by other models. For continuous, overlapping generations, the Moran model is the most canonical choice. Generally, every exchangeable model with genealogy—under a suitable time rescaling—given by the Kingman coalescent leads to the same results as those obtained in the present paper; the genealogy of a sequence of exchangeable models converges to the Kingman coalescent if and only if the sequence of offspring distributions of a single individual has bounded finite variance and fulfills a condition regarding their third moments [Möhle and Sagitov (2001)].

2.2. *Mutation dynamics.* We model individuals whose genomes consist of sets of genes. Every individual has a set of genes $\mathcal{G}_c$, $g_c := |\mathcal{G}_c|$ which are absolutely necessary to survive and hence are conserved, that is, must be passed from ancestor to offspring. The genes $\mathcal{G}_c$ constitute the *core genome*. In addition, we model an infinite gene pool by a set of genes $I = [0, 1]$ with $\mathcal{G}_c \cap I = \varnothing$. The genome of individual i in the sample, $1 \le i \le N$, contains genes $\mathcal{G}_i \subseteq I$ which are not necessary for the individuals to survive. This set of genes is called the *dispensable genome* of individual i.

During the lifetime of every individual or at every reproduction event, mutations may happen. In our mutation model, the *infinitely many genes model*, we assume the following two mechanisms (in a Wright–Fisher population of size N) which changes the dispensable genome from parent to offspring:

- *gene gain*: before reproduction of individual i, there is a probability μ that a new gene $u \in I$ is taken up from the environment. We assume that gene u has never been present in any genome of the population;
- *gene loss*: every gene of the dispensable genome $u \in \mathcal{G}_i$ of individual i is lost with probability ν before reproduction of individual i.

We take an extreme point of view here in that the core genes are absolutely necessary for an individual to survive and the genes in the dispensable genome evolve completely neutral. Using this view, ancestry is not affected by mutations, that is, all gene gains and losses seen in the population are assumed to be neutral.

Using the same time-scaling as for the genealogies, we assume that $\mu = \mu_N$ and $\nu = \nu_N$ are such that $\theta = \lim_{N\to\infty} 2N\mu_N$ and $\rho = \lim_{N\to\infty} 2N\nu_N$. After this rescaling of the parameters, new genes are gained at rate $\frac{\theta}{2}$ and present genes are lost at rate $\frac{\rho}{2}$.

DEFINITION 2.2 (Tree-indexed Markov chain for gene gain and loss). Let $\mathcal{T}$ be the Kingman coalescent. We either assume that $\mathcal{T}$ is rooted at the most recent common ancestor of the sample or that $\mathcal{T}$ has a single infinite line. Given $\mathcal{T}$, we define a Markov chain $\Gamma_{\mathcal{T}} = (\mathcal{G}_t)_{t\in\mathcal{T}}$, indexed by $\mathcal{T}$, with state space $\mathcal{N}_f(I)$, the space of counting measures on I. [The Markov property for the tree-indexed Markov chain $\Gamma_{\mathcal{T}}$ states that for all $t \in \mathcal{T}$, $(\mathcal{G}_s)_{t\preceq s}$ depends on $(\mathcal{G}_s)_{s\preceq t}$ only through $\mathcal{G}_t$.] Denoting by λ_I the Lebesgue measure on I, $\Gamma_{\mathcal{T}}$ makes transitions

$$\begin{aligned} &\text{from } g \text{ to } g+\delta_u \text{ at rate } \frac{\theta}{2}\lambda_I(du),\\ &\text{from } g \text{ to } g-\delta_u \text{ at rate } \frac{\rho}{2}g(du) \end{aligned} \tag{2.1}$$

along $\mathcal{T}$. Taking into account that the tree $\mathcal{T}$ has n leaves, one for each individual of the sample, we denote these leaves by $1,\dots,n \in \mathcal{T}$. In this setting, $\mathcal{G}_1,\dots,\mathcal{G}_n$ describe the genes present in individuals $1,\dots,n$.

An illustration of the tree-indexed Markov chain is shown in Figure 1.

REMARK 2.3 (Notation). Note that all gained genes are almost surely different, so $\mathcal{G}_t$ does not have double points, that is, $\mathcal{G}_t(\{u\}) \in \{0,1\}$, for all $u \in I$

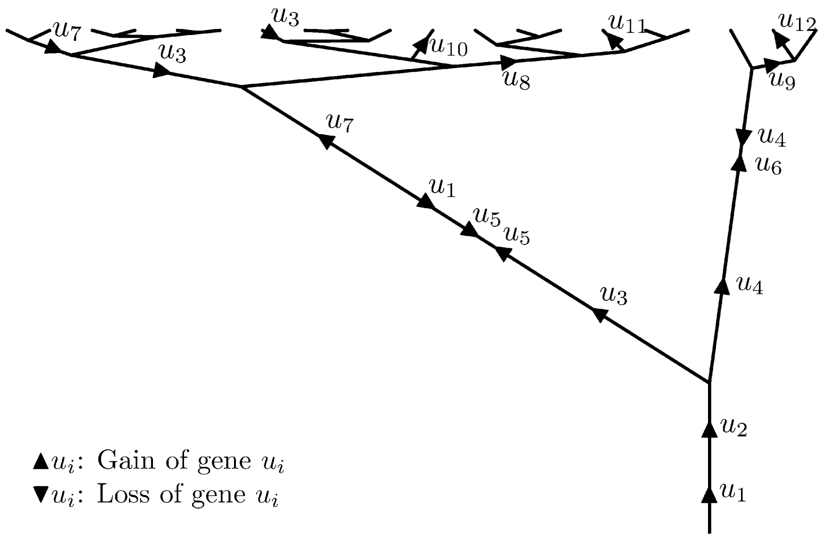

FIG. 1. *An illustration of the infinitely many genes model along a Kingman coalescent. If a gene is gained along a line* (*indicated by the* ▲*-sign*) *it can be lost again* (*indicated by a* ▼*-sign*). *An individual of the sample* (*i.e., a leaf of the coalescent tree*) *carries the set of genes which were gained along its ancestral lines and did not get lost again. Here are some examples*: *the gene* u_2 *is present in all individuals*, u_3 *is only present in* 10 *individuals of the left branch due to two gene losses. The genes* u_4 *and* u_5 *were lost in all ancestral lines and do not occur in any individual. The gene* u_7 *is missing in the* 2 *individuals on the left-hand side due to a gene loss and in the right branch as the gene gain was in the left branch.*

and $t \in \mathcal{T}$, almost surely. We will use the following notation, equating counting measures without double points with their support: consider $g \in \mathcal{N}_f(I)$ without double points. There is $m \in \mathbb{N}$ and $u_1, \dots, u_m$ with $g = \sum_{i=1}^m \delta_{u_i}$. We will refer to $u_1, \dots, u_m$ as the points in g and also write $g = \{u_1, \dots, u_m\}$. Moreover, we define

$$|g| := \int 1\,dg, \qquad g_s \cap g_t := g_s \wedge g_t,$$
$$g_s \setminus g_t := (g_s - g_t)^+$$

for $g, g_s, g_t \in \mathcal{N}_f(I)$.

Our aim is to describe patterns of the dispensable genomes $\mathcal{G} := (\mathcal{G}_1, \dots, \mathcal{G}_n)$ or whole genomes $(\mathcal{G}_1 \cup \mathcal{G}_c, \dots, \mathcal{G}_n \cup \mathcal{G}_c)$. These results can then be compared to genomic data of a sample of bacteria which gives the genes (or gene families) carried by individuals in the sample.

2.3. *Comparison to other mutation models.* In mathematical population genetics, there are several standard mutation models, for example, the *infinitely many alleles model* and the *infinitely many sites model* [see, e.g., Durrett (2008) or Ewens (2004)]. In the former, every mutation (along some random tree) leads to a new type, also called a new allele. It is assumed that mutated alleles have never been present in the population before. The latter is a refinement of the former: the allele of an individual in the sample is modeled as an infinite stretch on DNA. Every mutation is assumed to change a single site on this genome (hence leading to a new allele), and it is assumed that every mutation hits a site that has never been hit before. The last assumption is relaxed in the *finite sites model* where sites can be hit several times changing their state between several possibilities.

The *infinitely many genes model* as described above is conceptually different from these standard models: the infinitely many sites model (along some random tree $\mathcal{T}$) can be described, when a genome is given as the linear set I as above using that events, occurring at rate $\frac{\theta}{2}\lambda(du)$ along the tree, changes the state from the ancestral to a derived state at position u in the genome. However, loss events do not have a correspondence in the infinitely many sites model. In the finite sites model, a site can change from the ancestral state to a derived state and back; however, in the infinitely many genes model, once a gene has changed from the ancestral state (not present) to the derived state (present) and back (not present) no further changes of the state are possible.

Although all mutation models are conceptually different, the infinite sites model can be seen as the infinitely many genes model for $\rho = 0$. To understand this, consider a random tree with gene gain events (and no losses due to $\rho = 0$). Reinterpreting these gene gains as point mutations along a chromosome, each mutation hitting a new site, leads to the infinitely many sites model. However, there are still differences between the infinitely many genes model for $\rho = 0$ and the infinitely many sites model. On the one hand, $\rho = 0$ implies that genes cannot get lost which leads to an infinite genome for all individuals. On the other hand, most interesting quantities concentrate on mutations segregating (i.e., showing both the ancestral and the mutated—or derived—state) in the sample. In summary, the infinitely many genes model for $\rho = 0$ and the infinitely many sites model are the same with respect to properties of segregating genes/sites. However, we will see in our results below (Theorems 2, 4 and 5) that the infinitely many genes model is not continuous in $\rho = 0$ for certain aspects of segregating mutations.

3. Results.

In the rest of the paper, we fix a sample of size $n \in \mathbb{N}$ and $\theta, \rho > 0$. We describe expectations and variances of several quantities of interest. If we want to stress the dependence on the model parameters we will use subscripts, for example, $\mathbb{E}_{\theta,\rho}[\cdot]$, in order to make this clear. Since the core genome is conserved for all individuals of the population, we focus on the dispensable genome first. We provide results for the average number of genes in the sample (Section 3.1, Theorem 1), the average number of pairwise differences (Section 3.2, Theorem 2), incongruent pairs of genes (Section 3.3, Theorem 3), the size of the dispensable genome of the sample (Section 3.4, Theorem 4) and the gene frequency spectrum (Section 3.5, Theorem 5). We then extend these results to the complete pangenome, that is, the union of the dispensable and core genome (Section 3.6) and describe the application of our model to a dataset from Prochlorococcus, a marine cyanobacterium (Section 3.7). Finally, we discuss biologically realistic extensions of our model (Section 3.8).

3.1. *Average number of genes.*

The simplest statistics in the infinitely many genes model is based on counting the number of genes for all individuals in the sample. The *average number of genes* (*in the dispensable genome*) is given by

$$A := \frac{1}{n}\sum_{i=1}^{n} |\mathcal{G}_i|. \tag{3.1}$$

Our first results provide the first and second moment for A.

THEOREM 1 (Average number of genes). *For A as above*

$$\mathbb{E}[A] = \frac{\theta}{\rho},$$

$$\mathbb{V}[A] = \frac{1}{n}\frac{\theta}{1+\rho} + \frac{\theta}{\rho(1+\rho)}.$$

REMARK 3.1. Note that the result for $\mathbb{E}[A]$ is robust against changes of the reproduction mechanism or nonequilibrium situations. Consider any model of reproduction which has not gone extinct by time t. As long as the mutation mechanism is independent of reproduction, picking an individual at time t from the population gives a single ancestral line along which genes accumulate by the same distribution. In particular, the results for $\mathbb{E}[A]$ remain unaltered under population size changes or population subdivision.

The fact that $\mathbb{V}[A]$ does not converge to 0 as $n \to \infty$ does not come as a big surprise: the sets $\mathcal{G}_1, \dots, \mathcal{G}_n$ are dependent through the joint genealogy—given through the Kingman coalescent—relating the sample.

3.2. *Average number of pairwise differences.* The *average number of pairwise differences* is given by

$$D := \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} |\mathcal{G}_i \setminus \mathcal{G}_j|. \tag{3.2}$$

THEOREM 2 (Average number of pairwise differences). *For D as above,*

$$\begin{aligned}
\mathbb{E}[D] &= \frac{\theta}{1+\rho},\\
\mathbb{V}[D] &= \theta\biggl(\frac{(3+14\rho+23\rho^2+16\rho^3+4\rho^4+4\theta+2\rho\theta)}{(1+\rho)^2(2+\rho)(3+\rho)(1+2\rho)(3+2\rho)}\\
&\qquad + \frac{6+19\rho+19\rho^2+12\rho^3+4\rho^4+8\theta+4\rho\theta}{(1+\rho)(2+\rho)(3+\rho)(1+2\rho)(3+2\rho)}\frac{1}{n}\\
&\qquad + \frac{3+11\rho+12\rho^2+4\rho^3+10\theta+9\rho\theta+2\rho^2\theta}{(1+\rho)(2+\rho)(3+\rho)(1+2\rho)(3+2\rho)}\frac{2}{n(n-1)}\biggr).
\end{aligned}$$

REMARK 3.2. The quantity D is only based on genes segregating in the sample. Hence, as explained in Section 2.3, the infinitely many genes model for $\rho = 0$ is equivalent to the infinitely many sites model with respect to D. As the theorem shows, the expected number of differences between individuals i and j is $\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j| + |\mathcal{G}_j \setminus \mathcal{G}_i|] = 2\frac{\theta}{1+\rho}$. Hence, $\mathbb{E}_\rho[D]$ is not continuous in $\rho = 0$ since the comparable quantity in the infinite sites model, the average number of segregating sites in a sample of size two, is θ. The reason is that for small ρ, every individual carries a lot of genes, all of which can get lost at the small rate ρ. These loss events lead to differences between individuals as well as events of gene gain. A similar argument shows that the variance is not continuous [see, e.g., Wakeley (2008), (4.15)] for the variance in the infinite sites model.

Note that $\mathbb{V}[D]$ does not converge to 0 as $n \to \infty$. Again—the reason is that the differences $(\mathcal{G}_i \setminus \mathcal{G}_j)_{1 \le i \ne j \le n}$ are dependent through the underlying common genealogy.

3.3. *Incongruent pairs of genes.* Assume the following situation: for a pair of genes there are four individuals in which all four possible states of presence/absence of the two genes are observed. This means that the following situation is found:

	gene 1	gene 2
Individual 1	present	present
Individual 2	present	absent
Individual 3	absent	present
Individual 4	absent	absent

If genes cannot be lost ($\rho = 0$) this situation cannot occur in our model. The reason is that gene 1 would indicate that individuals 1 and 2 have a common ancestor before 1 and 3 have (otherwise individual 3 would also carry gene 1), while gene 2 indicates that individuals 1 and 3 have a common ancestor before 1 and 2 have. This is the reason why we call pairs of genes for which the above situation appears *incongruent*. If $\rho > 0$, incongruent pairs can arise by gene loss; see Figure 2 for an example. We will now state how many incongruent pairs we can expect to see in our sample.

The *average number of incongruent pairs of genes* (*in four genomes*) is given by

$$P := \frac{1}{n(n-1)(n-2)(n-3)} \sum_{i,j,k,l=1}^{n} D_{ij,kl} \cdot D_{ik,jl},$$

where

$$D_{ij,kl} := |(\mathcal{G}_i \cap \mathcal{G}_j) \setminus (\mathcal{G}_k \cup \mathcal{G}_l)|, \qquad 1 \le i, j, k, l \le n. \tag{3.3}$$

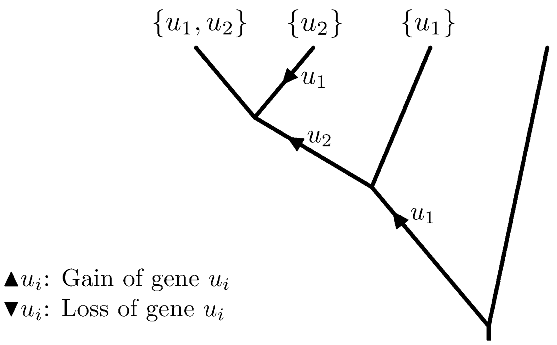

FIG. 2. *If genes can be lost* ($\rho > 0$) *it is possible that all four possible configurations of presence/absence of a pair of genes are seen in a sample of four individuals. We call such pairs incongruent. The figure shows one possible history of the two genes which leads to incongruence.*

THEOREM 3 (Incongruent pairs of genes). *For P as above*,

$$\mathbb{E}[P] = \frac{\theta^2\rho}{4}\left(18 + 117\frac{\rho}{2} + 203\frac{\rho^2}{4} + 105\frac{\rho^3}{8}\right) \\ \times \left(\left(1+\frac{\rho}{2}\right)^2\left(1+2\frac{\rho}{2}\right)\left(1+4\frac{\rho}{2}\right) \\ \times \left(3+4\frac{\rho}{2}\right)\left(3+5\frac{\rho}{2}\right)\left(6+5\frac{\rho}{2}\right)\left(6+7\frac{\rho}{2}\right)\right)^{-1}.$$

REMARK 3.3. In the proof of Theorem 3, we have to consider all possible genealogies of four individuals. In order to obtain the variance of P, one would have to take into account all possible genealogies relating eight individuals.

Note that

$$\mathbb{E}[D_{ij,kl}] = \frac{1}{6}\frac{\theta}{2}\frac{4\cdot 3}{(3+\rho)(2+\rho)} = \frac{\theta}{(3+\rho)(2+\rho)} \tag{3.4}$$

by the gene frequency spectrum (Theorem 5). Using this results it can be shown that $\mathbb{COV}[D_{ij,kl}, D_{ik,jl}] < 0$ in all cases.

For $\rho = 0$ we find that either $D_{ij,kl} = 0$ or $D_{ik,jl} = 0$ implying that $P = 0$ (hence $\mathbb{E}_{\rho=0}[P] = 0$), almost surely. The theorem implies that $\mathbb{E}_\rho[D_{ij,kl} \cdot D_{ik,jl}] \xrightarrow{\rho\to 0} 0$, that is, $\mathbb{E}_\rho[P]$ is continuous in $\rho = 0$. This observation is not obvious since small ρ implies that all individuals carry many genes. Using this fact, one could argue that the chance to observe a pair of genes giving rise to $D_{ij,kl} \cdot D_{ik,jl}$ grows with decreasing ρ. However, although the number of genes grows for small ρ, pairs of genes giving rise to $D_{ij,kl} \cdot D_{ik,jl}$ are most probably created by two gene gains and one gene loss, as shown in Figure 2, for small ρ.

As we will discuss in Section 3.8, the possibility of horizontal gene transfer (by bacterial conjugation) would be a biologically realistic extension of our model. Under such a mechanism, new genes are not only taken from the environment, but also from other individuals of the population. As a result, the genealogical tree would be different for different genes. Hence, the order of coalescence can be different, and so there is an increased number of incongruent pairs of genes. Hence, the theorem is valuable for determining the possibility of horizontal gene transfer in real populations.

3.4. *Size of the dispensable genome.* Now we come to properties involving the whole sample (in contrast to pairs and quartets of individuals in the last theorems). The simplest statistics involving all individuals of the sample is the total number of genes, that is, the size of the dispensable genome.

The *size of the dispensable genome* is given by

$$G := \left|\bigcup_{i=1}^{n} \mathcal{G}_i\right|. \tag{3.5}$$

We need the following definition in order to give the variance of the number of genes in the dispensable genome in our next theorem.

DEFINITION 3.4 (The functions h_k and $g_{\underline{k}}$). For $k \geq 0$ we define

$$h_k := 2\sum_{i=0}^{k-1} \frac{1}{\rho + i}. \tag{3.6}$$

Moreover, for $\underline{k} = (k_1, k_2, k_3)$ with $k_1, k_2, k_3 \geq 0$, we set

$$\begin{aligned}
&\underline{k}'_1 = (k_1 - 1, k_2, k_3), \qquad &&\underline{k}'_2 = (k_1, k_2 - 1, k_3),\\
&\underline{k}'_3 = (k_1, k_2, k_3 - 1), \qquad &&\underline{k}'_4 = (k_1 + 1, k_2 - 1, k_3 - 1),\\
&\underline{k}'_5 = (k_1 - 1, k_2 + 1, k_3), \qquad &&\underline{k}'_6 = (k_1 - 1, k_2, k_3 + 1),\\
&\lambda_1 = \binom{k_1}{2}, \qquad &&\lambda_2 = \binom{k_2}{2} + k_1 k_2 + \frac{\rho}{2} k_2,\\
&\lambda_3 = \binom{k_3}{2} + k_1 k_3 + \frac{\rho}{2} k_3, \qquad &&\lambda_4 = k_2 k_3, \qquad \lambda_5 = \lambda_6 = \frac{\rho}{2} k_1
\end{aligned}$$

and $\overline{\lambda} = \sum_{i=1}^{6} \lambda_i$. We define recursively

$$g_{\underline{k}} = \begin{cases} \dfrac{2}{\rho} h_{k_1 + k_2}, & \text{if } k_1 + k_3 = 1,\\ \dfrac{2}{\rho} h_{k_1 + k_3}, & \text{if } k_1 + k_2 = 1, \end{cases} \tag{3.7}$$

and

$$g_{\underline{k}} = (k_1 + k_2)(k_1 + k_3) \frac{2}{\overline{\lambda}^2} \tag{3.8}$$

$$+ \sum_{i=1}^{6} \frac{\lambda_i}{\overline{\lambda}} \left(\frac{1}{\overline{\lambda}} \big((k_1 + k_2) h_{k'_1 + k'_3} + (k_1 + k_3) h_{k'_1 + k'_2} \big) + g_{\underline{k}'_i} \right) \tag{3.9}$$

in all other cases.

THEOREM 4 (Size of the dispensable genome). *For G as above*,

$$\mathbb{E}[G] = \theta \sum_{i=0}^{n-1} \frac{1}{\rho + i}. \tag{3.10}$$

In addition, with $g_{(k_1, k_2, k_3)}$ given in Definition 3.4,

$$\mathbb{V}[G] = \theta \sum_{i=0}^{n-1} \frac{1}{\rho + i} - \theta^2 \left(\sum_{i=0}^{n-1} \frac{1}{\rho + i} \right)^2 + \frac{\theta^2}{4} g_{(n,0,0)}. \tag{3.11}$$

REMARK 3.5. An estimate for the size of the pangenome (dispensable plus core genome) in real bacterial populations has attained much interest [e.g., Tettelin et al. (2005), Lapierre and Gogarten (2009)]. Most interestingly, some species like *Bacillus anthracis* seem to have a *closed genome*, that is, only a limited number of genes in the pangenome, since no new gene was found after sequencing the fourth out of eight strains [Tettelin et al. (2005)]. Other species like *Prochlorococcus* and *Synechococcus*, have an *open genome* since estimates based on 22 different strains show that every newly sequenced genome exhibits 277 new genes on average (Baumdicker, unpublished observation). For open genomes, a model based on some linguistic insights, Heap's law, has been considered [Tettelin et al. (2008)]. As a result, a power law for the total number of genes is found and it is estimated that a total of $n^{0.43\pm 0.02}$ genes are found in a sample of n individuals from *Bacillus cereus*. This finding is in stark contrast to our theorem, which implies that the number of genes grows only logarithmically in n. However, in order to decide which is the correct asymptotics certainly requires a lot more data, since $n = 14$ strains are not enough to specify asymptotic behavior.

We conjecture that $\mathbb{V}[G]$ grows like $\mathbb{E}[G]$ for $n \to \infty$. (The corresponding statement is true in the infinite sites model [see Wakeley (2008), (4.8)].) The reason is that for given $\mathcal{T}$, G is Poisson distributed with a parameter increasing with the tree length. In addition, for large n the length of the Kingman coalescent is largest near the leaves and the coalescent almost becomes deterministic near the leaves. For example, it has been shown that the sum of *external branch* lengths (i.e., branches connecting a leaf to the next node in the tree) converges to 2 in L^2 [Fu (1995)].

We give an example for the computation of $g_{\underline{k}}$ in the case $\underline{k} = (2, 0, 0)$. For the calculation, we observe that $\lambda_1 = 1, \lambda_2 = \lambda_3 = \lambda_4 = 0, \lambda_5 = \lambda_6 = \rho$, $\overline{\lambda} = 1 + 2\rho$ and, from (3.6) and (3.7),

$$h_1 = \frac{2}{\rho}, \qquad h_2 = \frac{2}{\rho} + \frac{2}{1+\rho} = \frac{2(1+2\rho)}{\rho(1+\rho)},$$

$$g_{(1,0,0)} = \frac{4}{\rho^2}, \qquad g_{(1,1,0)} = g_{(1,0,1)} = \frac{4}{\rho}\left(\frac{1}{\rho} + \frac{1}{\rho+1}\right) = \frac{4(1+2\rho)}{\rho^2(1+\rho)}.$$

The recursion (3.8) then gives

$$\begin{aligned} g_{(2,0,0)} &= \frac{8}{(1+2\rho)^2} + \frac{1}{1+2\rho}\left(\frac{4h_1}{1+2\rho} + g_{(1,0,0)}\right) \\ &\quad + \frac{2\rho}{1+2\rho}\left(\frac{1}{1+2\rho}(2h_1 + 2h_2) + g_{(1,0,1)}\right) \\ &= 4\left(\frac{2}{(1+2\rho)^2} + \frac{2}{\rho(1+2\rho)^2} + \frac{1}{\rho^2(1+2\rho)} + \frac{2}{(1+2\rho)^2}\right. \\ &\qquad \left. + \frac{2}{(1+2\rho)(1+\rho)} + \frac{2}{\rho(1+\rho)}\right) \end{aligned} \tag{3.12}$$

$$= 4\left(\frac{1}{\rho^2} + \frac{2}{\rho(1+\rho)} + \frac{2}{(1+2\rho)(1+\rho)}\right)$$

$$= 4\left(\frac{1}{\rho} + \frac{1}{(1+\rho)}\right)^2 + \frac{4}{(1+\rho)^2(1+2\rho)}.$$

Using (3.11) this then gives for $n = 2$

$$\mathbb{V}_{n=2}[G] = \theta \frac{1+2\rho}{\rho(1+\rho)} + \theta^2 \frac{1}{(1+\rho)^2(1+2\rho)}.$$

For $n = 3$, the computation is more involved[2] and leads to

$$\mathbb{V}_{n=3}[G] = \frac{\theta}{\rho} + \frac{\theta}{1+\rho} + \frac{\theta}{2+\rho} + \theta^2 \frac{90 + 249\rho + 275\rho^2 + 145\rho^3 + 30\rho^4}{(1+\rho)^2(2+\rho)^2(1+2\rho)(3+2\rho)(6+5\rho)}.$$

3.5. *Gene frequency spectrum.*[3] By definition, core genes are present in all individuals of the sample. In contrast, genes from the dispensable genome can be present at any frequency. These possibilities give rise to the gene frequency spectrum.

The *gene frequency spectrum* (*of the dispensable genome*) is given by $G_1, \dots, G_n$, where

$$G_k^{(n)} := G_k := |\{u \in I : u \in \mathcal{G}_i \text{ for exactly } k \text{ different } i\}|. \tag{3.13}$$

THEOREM 5 (Gene frequency spectrum). *For $G_1, \dots, G_n$ as above*,

$$\mathbb{E}[G_k] = \frac{\theta}{k} \frac{n \cdots (n-k+1)}{(n-1+\rho)\cdots(n-k+\rho)}, \qquad k = 1, \dots, n.$$

REMARK 3.6. In the case $\rho = 0$, genes cannot get lost and consequently $G_n = \infty$. However, the classes $k = 1, \dots, n-1$ consist of genes segregating in the sample (since both states—presence and absence of the gene—are observed). Hence, as discussed in Section 2.3, these classes follow predictions for the infinite sites model. In this model, it is implicit in results already obtained by Wright (1938) [and later were refined by Kimura (1964), Griffiths (2003), Evans, Shvets and Slatkin (2007)] that

$$\mathbb{E}_{\rho=0}[G_k] = \frac{\theta}{k}.$$

[2] Several computations in the paper are most easily done using a program like MATHEMATICA. Therefore, a MATHEMATICA-notebook with all relevant computations can be downloaded from the homepage of the corresponding author.

[3] The term *gene frequency spectrum* was used by Kimura (1964) to denote the frequency of alleles in the infinite sites model. Later, the term changed to *site frequency spectrum* since single sites on the chromosome could be sequenced [e.g., Durrett (2008)]. Here, we reintroduce the term for gene frequencies in the infinitely many genes model.

On the other hand, by the theorem,

$$\mathbb{E}_\rho[G_k] \xrightarrow{\rho\downarrow 0} \frac{\theta n}{k(n-k)}$$

such that the gene frequency spectrum is not continuous at $\rho = 0$.

The model for the bacterial supragenome, introduced in Tettelin et al. (2005) takes population frequencies of genes into account, that is, the gene frequency spectrum. While the supragenome model assumes several different frequency classes to begin with, we derive the gene frequency spectrum from first principles, that is, from gene gain and loss events along the genealogy.

3.6. *Union of core and dispensable genome.* Until now we only derived results for the dispensable genome. In data obtained from bacterial species, the union of the core and dispensable genome is of primary interest. It is straightforward to extend our results to this union:

If we replace $\mathcal{G}_i$ by $\mathcal{G}_i \cup \mathcal{G}_c$, $1 \le i \le n$, in (3.1)–(3.3), (3.5) and (3.13), recall $g_c := |\mathcal{G}_c|$, and denote the resulting quantities by $\widetilde{A}$, $\widetilde{D}$, $\widetilde{P}$, $\widetilde{D}_{ij,kl}$, $\widetilde{G}$, $\widetilde{G}_k$, we obtain

$$\widetilde{A} = A + g_c, \qquad \widetilde{D} = D, \qquad \widetilde{D}_{ij,kl} = D_{ij,kl}, \qquad \widetilde{P} = P, \qquad \widetilde{G} = G + g_c$$

and

$$\widetilde{G}_k = \begin{cases} G_k, & k = 1, \dots, n-1, \\ G_k + g_c, & k = n. \end{cases}$$

Hence, properties of $\widetilde{A}$, $\widetilde{D}$, $\widetilde{P}$, $\widetilde{G}$, $\widetilde{G}_k$ follow immediately from Theorems 1–5.

3.7. *Application*: *A dataset from* Prochlorococcus. Data from complete genomes of a population sample of bacteria have been available only for a few years. Because the infinitely many genes model we propose is new in the population genetic context, we show some data in order to see if the model as studied above could be realistic.

Here we chose a set of $n = 9$ strains of *Prochlorococcus* which appear to be closely related. *Prochlorococcus* is a marine picocyanobacterium (length $\sim$ 0.6 μm, genome size $\sim$ 2 Mbp) living in the ocean at depth up to 200 m. Their population size can be as large as 10^6 individuals (i.e., cells) per ml. In total, 22 complete genome sequences of these cyanobacteria are available in GenBank at the moment [Kettler et al. (2007), Dufresne et al. (2008)]. The $n = 9$ chosen *Prochlorococcus* genomes are similar to each other in terms of GC-content and share a similar physiology.

We estimate the model parameters θ, ρ and g_c based on the gene frequency spectrum $\widetilde{G}_1, \dots, \widetilde{G}_9$ which we compare with our results from Theorem 5. The number of genes present in all individuals is 1282, forming the largest class in the observed gene frequency spectrum (see Figure 3). Genes occurring in only a single

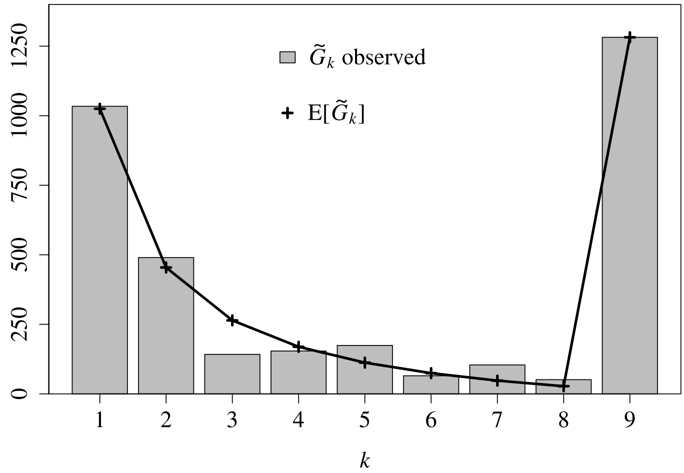

FIG. 3. *The fit of observed data from nine closely related strains of* Prochlorococcus *with the expectations for the gene frequency spectrum. Estimates were as given in* (3.14).

individual were the second largest class with 1034 genes. By a least squares fit of $\widetilde{G}_k$ and $\mathbb{E}[\widetilde{G}_k]$ for $k = 1, \ldots, n$ we obtain the estimates

$$\widehat{\theta} = 1142.17, \qquad \widehat{\rho} = 2.03, \qquad \widehat{g}_c = 1270. \tag{3.14}$$

Note that the estimate for g_c means that we expect that 14 genes which are carried by all individuals belong to the dispensable genome. As shown in Figure 3, these estimates produce a reasonably good fit with the data. Of course, a statistical test which is able to reject our model for gene content in general, and the assumption that all genes in the dispensable genome evolve neutrally in particular, would be desirable.

3.8. *Outlook.* We introduce the infinitely many genes model on a Kingman coalescent as a simple null-model of genome evolution in bacterial species. However, both the reproduction and the mutation dynamics can be extended to become biologically more realistic. For the reproduction dynamics, several extensions have been considered in the literature, for example, structured populations and populations of varying size [see, e.g., Durrett (2008)].

The mutation dynamics can be extended as well. Our strongest assumption is that genes taken from the environment are completely new. In particular, the model does not allow for genes being transferred between individuals directly. Such a physical exchange of genes between bacteria is known as horizontal gene transfer. The underlying mechanism is *bacterial conjugation*. The donor cell produces a pilus that attaches to the recipient cell and a single strand of DNA is transported from the donor to the recipient. After replication of the DNA, both cells carry the transferred genetic material. The duration of conjugation is long enough in order to transfer several genes. Hence, by events of horizontal gene transfer, the transferred genes do not share the genealogy of the cell line. Thus, building such a mechanism into the above model requires the use of different genealogical trees for different genes. Such a mechanism was already considered in the phylogenetics literature by Kunin and Ouzounis (2003).

In order to add even more biological realism, at least three aspects can be considered:

1. As Lefébure and Stanhope (2007) show there is frequent recombination even within the core genome. Such recombination can also be explained by conjugation and has attained much interest [e.g., Fraser, Hanage and Spratt (2007)] since the amount of recombination is known to be related to sequence similarity [e.g., Vulic et al. (1997)], suggesting that bacterial species can be distinguished by the extent of recombination between strains [Dykhuizen and Green (1991), Maynard-Smith (1995)].
2. As seen in genomic data, several genes are clustered in gene families. This is best explained by events of *gene duplication* with a potential *subfunctionalization* of these genes along ancestral lines [e.g., Durrett (2008), Durrett and Popovic (2009)].
3. There are certainly selective constraints on the number of genes in the dispensable genome. If these genes are evolving neutrally and are not necessarily needed for a bacterium to function properly, selection should act in order to minimize the dispensable genome.

Specifying the set of genes of an individual requires that the whole genome of the individual is sequenced. Finding the different genes in a dataset like the one used in the last section means that open reading frames (ORFs), that is, regions in the genome between start codons and stop codons of all individuals are found. In the dataset we say that two individuals carry the same gene if we find a pair of ORFs in both individuals that are highly similar. However, the DNA sequence of this pair of ORFs is usually not identical. Refined mutation models should extend our approach and describe the genomic diversity of the different genes as well as the variation of DNA sequences within the genes.

4. The one-line-equilibrium and proof of Theorem 1. Consider a sample of size n and recall the sets of genes $\mathcal{G}_1, \dots, \mathcal{G}_n$ from Section 2.2. All results we provide with Theorems 1–5 are dealing with the joint distribution of $\mathcal{G}_1, \dots, \mathcal{G}_n$. We start with properties of one- and two-dimensional marginals of the total masses of this joint distribution. First, we have to obtain a key result for the gene content along a single ancestral line in Section 4.1. The first two moments of the one- and two-dimensional marginals are obtained in Section 4.2 which then lead to a proof of Theorem 1 in Section 4.3.

4.1. *The one-line equilibrium.* We start with some arguments that will appear frequently in the next sections. For $n = 1$, the random tree $\mathcal{T}$ is only a single infinite line. We consider the gene content along a single ancestral line $\mathcal{T} = \mathbb{R}_-$. In this setting, recall the process $\Gamma_{\mathbb{R}_-} = (\mathcal{G}_t)_{t\in\mathbb{R}_-}$ from Definition 2.2. Note that, almost surely, $\mathcal{G}_t$ does not have double points for all $t \in \mathbb{R}_-$. Recall our notation from Remark 2.3.

DEFINITION 4.1 (Poisson random measure and thinning). We denote by $\mathcal{POI}(\alpha)$ the distribution of a Poisson random measure with intensity measure α. We will also write $\mathcal{POI}(\alpha)$ for the Poisson distribution with parameter α if $\alpha \in \mathbb{R}_+$.

For $g \in \mathcal{N}_f(I)$, we denote by $\mathcal{THIN}(g, p)$ the distribution of the random measure arising by keeping any point in g with probability p.

PROPOSITION 4.2 (Distribution of $\Gamma_{\mathbb{R}_-}$). *Let $s < t$.*

Given $\mathcal{G}_s = g \in \mathcal{N}_f(I)$, the two random measures $\mathcal{G}_t \cap \mathcal{G}_s$ and $\mathcal{G}_t \setminus \mathcal{G}_s$ are independent. Their distribution is given by

$$\mathcal{G}_t \cap \mathcal{G}_s \sim \mathcal{THIN}(g, e^{-\rho/2(t-s)}),$$
$$\mathcal{G}_t \setminus \mathcal{G}_s \sim \mathcal{POI}\left(\frac{\theta}{\rho}(1 - e^{-\rho/2(t-s)}) \cdot \lambda_I\right).$$

The distribution $\mathcal{POI}(\frac{\theta}{\rho}\lambda_I)$ is the unique equilibrium for $\Gamma_{\mathbb{R}_-}$ and it is reversible. In equilibrium, $\mathcal{G}_t \cap \mathcal{G}_s, \mathcal{G}_t \setminus \mathcal{G}_s$ and $\mathcal{G}_s \setminus \mathcal{G}_t$ are independent and their distributions are given by

$$\mathcal{G}_t \cap \mathcal{G}_s \sim \mathcal{POI}\left(\frac{\theta}{\rho}e^{-\rho/2(t-s)} \cdot \lambda_I\right),$$
$$\mathcal{G}_s \setminus \mathcal{G}_t \overset{d}{=} \mathcal{G}_t \setminus \mathcal{G}_s \sim \mathcal{POI}\left(\frac{\theta}{\rho}(1 - e^{-\rho/2(t-s)}) \cdot \lambda_I\right).$$

REMARK 4.3. Recall from Remark 2.3 that we identify $\mathcal{G}_s, \mathcal{G}_t$ with the set of genes carried at times s and t. Note that $\mathcal{G}_s \cap \mathcal{G}_t = \mathcal{G}_s \wedge \mathcal{G}_t$ represents the genes present both at time s and at time t. Moreover, $\mathcal{G}_s \setminus \mathcal{G}_t = (\mathcal{G}_s - \mathcal{G}_t)^+$ are the genes present at time s but absent at time t, that is, genes lost during time $(s, t]$. The genes in $\mathcal{G}_t \setminus \mathcal{G}_s = (\mathcal{G}_t - \mathcal{G}_s)^+$ are genes gained during time $(s, t]$. As the proposition shows, all three quantities are independent in equilibrium.

PROOF OF PROPOSITION 4.2. First, recall that all new points in $\mathcal{G}_{s'}$, $s < s' \le t$ are pairwise different and different from points in $\mathcal{G}_s$, almost surely. During $(s, t]$ several points of $\mathcal{G}_{s'}$ are lost. A point in $\mathcal{G}_{s'}$ is not lost with probability $e^{-\rho/2(t-s')}$. Since all points are lost independently, we find that $\mathcal{G}_s \cap \mathcal{G}_t \sim \mathcal{THIN}(g, e^{-\rho/2(t-s)})$. Additionally, several new points in $\mathcal{G}$ arise during $(s, t]$.

Hence, we find that $\mathcal{G}_t \setminus \mathcal{G}_s$ is independent of both, $\mathcal{G}_s$ and $\mathcal{G}_s \cap \mathcal{G}_t$. To obtain the distribution of $\mathcal{G}_t \setminus \mathcal{G}_s$, note that a point in $\mathcal{G}_{s'} \setminus \mathcal{G}_{s'-}$ is lost at rate $\frac{\rho}{2}$ and hence is present in $\mathcal{G}_t$ with probability $e^{-\rho/2(t-s')}$. Since new points arise at rate $\frac{\theta}{2}$ during $(s,t]$ and are lost independently, we find that the number of points in $\mathcal{G}_t \setminus \mathcal{G}_s$ is Poisson distributed with parameter $\frac{\theta}{2}\int_0^{t-s} e^{-\rho/2(t-s')}\,ds' = \frac{\theta}{\rho}(1-e^{-\rho/2(t-s)})$. Since these points must be uniformly distributed on I, we have that $\mathcal{G}_t \setminus \mathcal{G}_s \sim \mathcal{POI}(\frac{\theta}{\rho}(1-e^{-\rho/2(t-s)})\cdot\lambda_I)$. So we have shown the first assertion.

To see that $\mathcal{POI}(\frac{\theta}{\rho}\cdot\lambda_I)$ is the unique equilibrium of $\Gamma_{\mathbb{R}}$, note that there can be at most one equilibrium since the Markov process $\Gamma_{\mathbb{R}_-}$ is Harris recurrent. Moreover, if $\mathcal{G}_s \sim \mathcal{POI}(\frac{\theta}{\rho}\cdot\lambda_I)$, then $\mathcal{THIN}(\mathcal{G}_s, e^{-\rho/2(t-s)}) = \mathcal{POI}(\frac{\theta}{\rho} \times e^{-\rho/2(t-s)}\cdot\lambda_I)$ and so

$$\mathcal{G}_t \sim \mathcal{POI}\left(\frac{\theta}{\rho}e^{-\rho/2(t-s)}\cdot\lambda_I\right) * \mathcal{POI}\left(\frac{\theta}{\rho}(1-e^{-\rho/2(t-s)})\cdot\lambda_I\right) = \mathcal{POI}\left(\frac{\theta}{\rho}\cdot\lambda_I\right)$$

(where $*$ denotes convolution). For reversibility, we write $\mathcal{G}_s = (\mathcal{G}_s \cap \mathcal{G}_t) \uplus (\mathcal{G}_s \setminus \mathcal{G}_t)$, $\mathcal{G}_t = (\mathcal{G}_s \cap \mathcal{G}_t) \uplus (\mathcal{G}_t \setminus \mathcal{G}_s)$ where $\mathcal{G}_s \cap \mathcal{G}_t, \mathcal{G}_s \setminus \mathcal{G}_t, \mathcal{G}_t \setminus \mathcal{G}_s$ are independent, such that $\Gamma_{\mathbb{R}_-}$ is in equilibrium at times s and t and

$$\begin{aligned}\mathcal{G}_s \cap \mathcal{G}_t &\sim \mathcal{POI}\left(\frac{\theta}{\rho}e^{-\rho/2(t-s)}\cdot\lambda_I\right),\\ \mathcal{G}_s \setminus \mathcal{G}_t \overset{d}{=} \mathcal{G}_t \setminus \mathcal{G}_s &\sim \mathcal{POI}\left(\frac{\theta}{\rho}(1-e^{-\rho/2(t-s)})\cdot\lambda_I\right).\end{aligned}$$

By this representation, given some continuous functions $f_1, f_2 : I \to \mathbb{R}$, writing $\langle f_i, x\rangle := \int f_i\,dx$, $i = 1,2$,

$$\begin{aligned}\mathbb{E}[e^{-\langle f_1,\mathcal{G}_s\rangle}\cdot e^{-\langle f_2,\mathcal{G}_t\rangle}] &= \mathbb{E}[e^{-\langle f_1+f_2,\mathcal{G}_s\cap\mathcal{G}_t\rangle}]\cdot\mathbb{E}[e^{-\langle f_1,\mathcal{G}_s\setminus\mathcal{G}_t\rangle}]\cdot\mathbb{E}[e^{-\langle f_2,\mathcal{G}_t\setminus\mathcal{G}_s\rangle}]\\ &= \mathbb{E}[e^{-\langle f_1+f_2,\mathcal{G}_s\cap\mathcal{G}_t\rangle}]\cdot\mathbb{E}[e^{-\langle f_1,\mathcal{G}_t\setminus\mathcal{G}_s\rangle}]\cdot\mathbb{E}[e^{-\langle f_2,\mathcal{G}_s\setminus\mathcal{G}_t\rangle}]\\ &= \mathbb{E}[e^{-\langle f_2,\mathcal{G}_s\rangle}\cdot e^{-\langle f_1,\mathcal{G}_t\rangle}].\end{aligned}$$

Hence, since the joint Laplace transforms $\mathbb{E}[e^{-\langle f_1,\mathcal{G}_s\rangle}\cdot e^{-\langle f_2,\mathcal{G}_t\rangle}]$ determine the joint distribution of $(\mathcal{G}_s,\mathcal{G}_t)$ uniquely, we find that $(\mathcal{G}_s,\mathcal{G}_t) \overset{d}{=} (\mathcal{G}_t,\mathcal{G}_s)$ and reversibility is shown. □

4.2. *Gene content in individuals and pairs*. Next we obtain the first two moments of the two-dimensional distribution of $(|\mathcal{G}_1|,\dots,|\mathcal{G}_n|)$.

PROPOSITION 4.4 [Distribution of $(\mathcal{G}_i,\mathcal{G}_j)$]. *For* $i = 1,\dots,n$,

$$\mathcal{G}_i \sim \mathcal{POI}\left(\frac{\theta}{\rho}\cdot\lambda_I\right).$$

In particular,

$$\mathbb{E}[|\mathcal{G}_i|] = \mathbb{V}[|\mathcal{G}_i|] = \frac{\theta}{\rho}.$$

For $1 \le i \neq j \le n$,

$$\mathbb{COV}[|\mathcal{G}_i|, |\mathcal{G}_j|] = \frac{\theta}{\rho(1+\rho)}.$$

REMARK 4.5. In the proof of 2. we use the well-known fact that for random variables X, Y, T

$$\mathbb{COV}[X, Y] = \mathbb{COV}[\mathbb{E}[X|T], \mathbb{E}[Y|T]] + \mathbb{E}[\mathbb{COV}[X, Y|T]]$$

with

$$\mathbb{COV}[X, Y|T] := \mathbb{E}[(X - \mathbb{E}[X|T])(Y - \mathbb{E}[Y|T])|T].$$

PROOF OF PROPOSITION 4.4. Consider the ancestral line of individual i. The process $(\mathcal{G}_t)_{t \preceq i}$ follows the same dynamics as the process $\Gamma_{\mathbb{R}_-}$ studied in Section 4.1. The first claim follows from Proposition 4.2(2), which shows that the unique reversible equilibrium for this dynamics is $\mathcal{POI}(\frac{\theta}{\rho}\lambda_I)$.

We denote the random coalescence time of individuals i and j by T. As above, $\mathcal{G}_i$ and $\mathcal{G}_j$ denote (the finite measures describing) the genes present in both individuals. Recall that we have shown in Proposition 4.2(2) that the equilibrium of the Markov chain $\Gamma_{\mathbb{R}_-} = (\mathcal{G}_t)_{t \in \mathbb{R}_-}$ of Section 4.1 is reversible. Hence, given T, we have that $(\mathcal{G}_i, \mathcal{G}_j)$ and $(\mathcal{G}_{-2T}, \mathcal{G}_0)$ have the same distribution. So we find that $\mathcal{G}_i \cap \mathcal{G}_j, \mathcal{G}_i \setminus \mathcal{G}_j$ and $\mathcal{G}_j \setminus \mathcal{G}_i$ are independent and, by Proposition 4.2,

$$\mathcal{G}_i \cap \mathcal{G}_j \sim \mathcal{POI}\left(\frac{\theta}{\rho} e^{-\rho T} \cdot \lambda_I\right), \qquad \mathcal{G}_i \setminus \mathcal{G}_j \overset{d}{=} \mathcal{G}_i \setminus \mathcal{G}_j \sim \mathcal{POI}\left(\frac{\theta}{\rho}(1 - e^{-\rho T}) \cdot \lambda_I\right).$$

Moreover, both sets of genes, $\mathcal{G}_i$ and $\mathcal{G}_j$, are independent of T. We obtain

$$\begin{aligned}
&\mathbb{COV}[|\mathcal{G}_i|, |\mathcal{G}_j|] \\
&\quad = \mathbb{COV}[\mathbb{E}[|\mathcal{G}_i||T], \mathbb{E}[|\mathcal{G}_j||T]] + \mathbb{E}[\mathbb{COV}[|\mathcal{G}_i|, |\mathcal{G}_j||T]] \\
&\quad = \mathbb{COV}\left[\frac{\theta}{\rho}, \frac{\theta}{\rho}\right] + \mathbb{E}[\mathbb{COV}[|\mathcal{G}_i \cap \mathcal{G}_j| + |\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_i \cap \mathcal{G}_j| + |\mathcal{G}_j \setminus \mathcal{G}_i||T]] \\
&\quad = \mathbb{E}[\mathbb{V}[|\mathcal{G}_i \cap \mathcal{G}_j||T]] = \mathbb{E}\left[\frac{\theta}{\rho} e^{-\rho T}\right] = \frac{\theta}{\rho(1+\rho)}
\end{aligned}$$

as $T \sim \text{Exp}(1)$. □

4.3. *Proof of Theorem* 1. Theorem 1 now follows from Proposition 4.4 and

$$\mathbb{E}[A] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|\mathcal{G}_i|] = \frac{\theta}{\rho},$$

$$\begin{aligned}\mathbb{V}[A] &= \frac{1}{n^2}\Biggl(\sum_{i=1}^{n}\mathbb{V}[|\mathcal{G}_i|] + \sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathbb{COV}[|\mathcal{G}_i|, |\mathcal{G}_j|]\Biggr)\\ &= \frac{1}{n}\frac{\theta}{\rho} + \Bigl(1-\frac{1}{n}\Bigr)\frac{\theta}{\rho(1+\rho)} = \frac{1}{n}\frac{\theta}{1+\rho} + \frac{\theta}{\rho(1+\rho)}.\end{aligned}$$

5. Extension of Proposition 4.2 and proof of Theorem 2.

The one-line equilibrium considered in Proposition 4.2 provides the right setting for computing the one- and two-dimensional marginals of $\mathcal{G}_1, \dots, \mathcal{G}_n$ as shown in the proof of Proposition 4.4. In Section 5.1 we provide a method to compute higher order marginals. We will use this method for second (Section 5.2), third (Section 5.3) and fourth (Section 5.4) order which finally leads to a proof of Theorem 2 in Section 5.5.

5.1. *Extending the one-line equilibrium to a genealogical tree.* Before we introduce the general method, how to obtain all marginals of $\mathcal{G}_1, \dots, \mathcal{G}_n$, we have to set the scene. Consider the genealogical tree $\mathcal{T}$ relating all n individuals and the tree-indexed Markov chain $\Gamma_{\mathcal{T}} = (\mathcal{G}_t)_{t\in\mathcal{T}}$. In equilibrium, we have seen above that $\mathcal{G}_t \sim \mathcal{POI}(\frac{\theta}{\rho}\cdot\lambda_I)$ for all $t \in \mathcal{T}$. Hence, we now consider the case that $\mathcal{T}$ is a rooted tree with root r and $\mathcal{G}_r \sim \mathcal{POI}(\frac{\theta}{\rho}\cdot\lambda_I)$. We need some notation to deal with the genealogical tree $\mathcal{T}$.

DEFINITION 5.1 (Survival function). Let $\mathcal{T}$ be a binary tree with one distinguished point $r \in \mathcal{T}$, referred to as the root of $\mathcal{T}$, a finite set of leaves $\mathcal{L} \subseteq \mathcal{T}$ and internal vertices $\mathcal{V}$. For $s, t \in \mathcal{T}$ we denote by $(s, t]$ the set of points which must be visited on any path between s and t. Moreover, $d_{\mathcal{T}}(s, t)$ is the length of the path between s and t. Define a partial order $\preceq$ on $\mathcal{T}$ by saying that $s \preceq t$ iff $s \in (r, t]$ (such that r is the minimal element). For $s, t \in \mathcal{T}$ the point $s \wedge t$ is given as the maximal element in $\{q : q \preceq s \text{ and } q \preceq t\}$. For an internal node (i.e., a branch point) $t \in \mathcal{T}$ we denote by t_1 and t_2 the two directions in $\mathcal{T}$ leading to bigger (with respect to $\preceq$) elements.

We define the *survival function* $p_{\mathcal{T}} : \mathcal{T} \to [0, 1]$ by

$$\begin{aligned} &p_{\mathcal{T}}(t) = 1 && \text{for } t \in \mathcal{L},\\ &\frac{\partial p_{\mathcal{T}}(t)}{\partial t} = \frac{\rho}{2}p_{\mathcal{T}}(t) && \text{for } t \in \mathcal{T}\setminus(\mathcal{L}\cup\mathcal{V}),\\ &p_{\mathcal{T}}(t) = 1 - (1 - p_{\mathcal{T}}(t_1))(1 - p_{\mathcal{T}}(t_2)) && \text{for } t \in \mathcal{V},\end{aligned} \tag{5.1}$$

where for $f : \mathcal{T} \setminus (\mathcal{L} \cup \mathcal{V}) \to \mathbb{R}$

$$\frac{\partial f(t)}{\partial t} := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\big(f(t+\varepsilon) - f(t)\big)$$

and $t + \varepsilon$ is any point in $\mathcal{T}$ with $d_{\mathcal{T}}(t, t+\varepsilon) = \varepsilon$ and $t \preceq t + \varepsilon$, if the limit exists.

PROPOSITION 5.2 (Probability of no loss along $\mathcal{T}$). *Let $\mathcal{T}$ be a binary tree, rooted at r, $p_{\mathcal{T}}$ as in Definition 5.1 and $\Gamma_{\mathcal{T}} = (\mathcal{G}_t)_{t \in \mathcal{T}}$ be the tree-indexed Markov chain from Definition 2.2 with $\mathcal{G}_r \sim \mathcal{POI}(\frac{\theta}{\rho}\lambda_I)$. Then for $u \in I$ and $t \in \mathcal{T}$*

$$\mathbb{P}\Bigg[u \in \bigcup_{\substack{t \preceq s \\ s \in \mathcal{L}}} \mathcal{G}_s \Bigg| u \in \mathcal{G}_t\Bigg] = p_{\mathcal{T}}(t).$$

PROOF. Denote the probability on the left-hand side by $q(t)$. First note that $q(t) = 1$ if $t \in \mathcal{L}$ since $\{s \in \mathcal{L} : t \preceq s\} = \{t\}$. Moreover, the probability on the left-hand side decreases exponentially at rate $\frac{\rho}{2}$ along branches of $\mathcal{T}$ due to loss events of u. Last, consider the case $t \in \mathcal{V}$. Then, u must not be lost to either t_1 or t_2. This occurs with probability $q(t) = 1 - (1 - q(t_1))(1 - q(t_2))$. In other words, the function q fulfills all defining properties of $p_{\mathcal{T}}$ from (5.1) and we are done. □

We need some more notation for subsets of a finite binary rooted tree $\mathcal{T}$.

DEFINITION 5.3 (Length and subtrees of $\mathcal{T}$). We use the notation of Definition 5.1.

For the binary tree $\mathcal{T}$ we denote by $\ell(\mathcal{T})$ its total length, that is, the sum of lengths of all its branches.

Let $\mathcal{L}', \mathcal{M}' \subseteq \mathcal{L}$ be sets of leaves with $\mathcal{L}' \cap \mathcal{M}' = \varnothing$. We set $r_0 := \bigwedge_{t \in \mathcal{L}'} t$ and denote by $\mathcal{T}^0(\mathcal{L}')$ the minimal connected, binary tree spanning the leaves $\mathcal{L}'$, rooted at r_0. The set $(\mathcal{T}^0(\mathcal{L}' \cup \mathcal{M}')) \setminus \mathcal{T}^0(\mathcal{L}')$ consists of $k \le |\mathcal{M}'|$ different connected subtrees, connected with $\mathcal{T}^0(\mathcal{L}')$ at vertices $r_1, \dots, r_k$. We denote the resulting binary trees by $\mathcal{T}^1(\mathcal{L}', \mathcal{M}'), \dots, \mathcal{T}^k(\mathcal{L}', \mathcal{M}')$, rooted at $r_1, \dots, r_k$, respectively.

REMARK 5.4. For an illustration of the objects introduced in Definition 5.3, see Figure 4.

If $|\mathcal{L}'| = 1$, it is important to note that $\mathcal{T}^0(\mathcal{L}')$ only consists of a single point. Consequently, $\ell(\mathcal{T}^0(\mathcal{L}')) = 0$ in this case.

PROPOSITION 5.5 (Distribution of $\bigcap \mathcal{G}_s \setminus \bigcup \mathcal{G}_t$). *Let $\mathcal{T}$ be a finite binary tree, rooted at $r \in \mathcal{T}$, $\mathcal{L}$ its finite set of leaves and $\mathcal{L}', \mathcal{M}' \subseteq \mathcal{L}$ with $\mathcal{L}' \cap \mathcal{M}' = \varnothing$. Moreover, let $\mathcal{T}^0(\mathcal{L}'), \mathcal{T}^1(\mathcal{L}', \mathcal{M}'), \dots, \mathcal{T}^k(\mathcal{L}', \mathcal{M}')$ be as in Definition 5.3. Let*

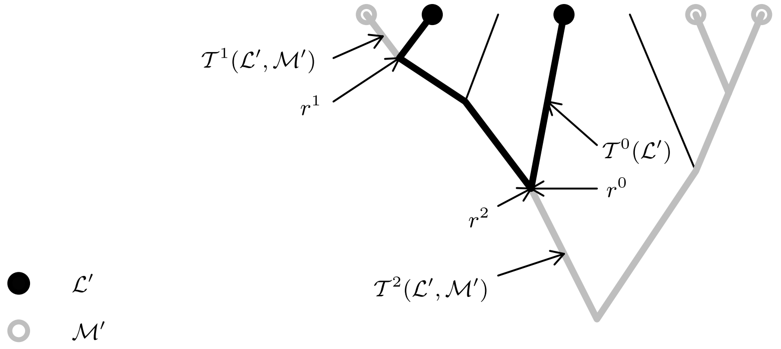
$\mathcal{T}^1(\mathcal{L}', \mathcal{M}')$
r^1
$\mathcal{T}^0(\mathcal{L}')$
r^0
r^2
$\mathcal{T}^2(\mathcal{L}', \mathcal{M}')$
● $\mathcal{L}'$
○ $\mathcal{M}'$

FIG. 4. *Illustration of concepts given in Definition 5.3(2). The subtree $\mathcal{T}^0(\mathcal{L}')$ is spanned by leaves in $\mathcal{L}'$. Considering $\mathcal{T} \setminus \mathcal{T}^0(\mathcal{L}')$, the tree falls in three parts. Two of them, which lead to a leaf in $\mathcal{M}'$, are denoted $\mathcal{T}^k(\mathcal{L}', \mathcal{M}')$, $k = 1, 2$. Roots of the trees are r^0, r^1, r^2.*

$\Gamma_{\mathcal{T}} = (\mathcal{G}_t)_{t\in\mathcal{T}}$ be the tree-indexed Markov chain from Definition 2.2 with $\mathcal{G}_r \sim \mathcal{POI}(\frac{\theta}{\rho})$. Then,

$$\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \Big\backslash \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t \sim \mathcal{POI}\biggl(\frac{\theta}{\rho} e^{-\rho/2\ell(\mathcal{T}^0(\mathcal{L}'))} \prod_{i=1}^{k} (1 - p_{\mathcal{T}^i(\mathcal{L}',\mathcal{M}')}(r_i)) \cdot \lambda_I\biggr).$$

In addition, if $\mathcal{L}'', \mathcal{M}'' \subseteq \mathcal{L}$ with $\mathcal{L}'' \cap \mathcal{M}'' = \varnothing$, then

$$\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \Big\backslash \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t \quad \textit{and} \quad \bigcap_{t\in\mathcal{L}''} \mathcal{G}_t \Big\backslash \bigcup_{t\in\mathcal{M}''} \mathcal{G}_t$$

are independent if $\mathcal{L}' \cap \mathcal{M}'' \neq \varnothing$ or $\mathcal{L}'' \cap \mathcal{M}' \neq \varnothing$.

REMARK 5.6. The pairwise independence in the proposition can be extended to independence of any number of random measures $\bigcap_{t\in\mathcal{L}_i} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}_i} \mathcal{G}_t$, $i = 1, \dots, n$, provided $\mathcal{L}_i \cap \mathcal{M}_j \neq \varnothing$ or $\mathcal{L}_j \cap \mathcal{M}_i \neq \varnothing$ holds for any pair $i \neq j$.

PROOF OF PROPOSITION 5.5. Given $\mathcal{T}$, rooted at $r \in \mathcal{T}$, we have assumed that $\mathcal{G}_r = \mathcal{POI}(\frac{\theta}{\rho} \cdot \lambda_I)$, that is, the tree-indexed Markov chain is in equilibrium. Consequently, $\mathcal{G}_{r_0} \sim \mathcal{G}_r$. Every gene in $\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t$ must have been present in $\mathcal{G}_{r_0}$. In addition, every gene in $\mathcal{G}_{r_0}$ has the same chance p to be present in $\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t$. This already shows that $\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t$ is a thinning of a Poisson measure and hence is Poisson with intensity $\frac{\theta}{\rho} p$. It is important to note that a gene present in $\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t$ must not be lost on the whole subtree $\mathcal{T}^0(\mathcal{L}')$, which occurs with probability $e^{-\rho/2\ell(\mathcal{T}^0(\mathcal{L}'))}$ and must be lost on any subtree leading to a leaf in $\mathcal{M}'$. However, the chance that a gene is lost along one such subtree is given through the survival function. In the subtree i, we have a root r_i connecting the subtree to the tree spanned by $\mathcal{L}'$ and so $1 - p_{\mathcal{T}^i(\mathcal{L}',\mathcal{M}')}(r_i)$ is the probability that the gene is lost in all leaves in $\mathcal{M}'$.

For the independence property assume that $\mathcal{L}' \cap \mathcal{M}'' \neq \varnothing$ or $\mathcal{L}'' \cap \mathcal{M}' \neq \varnothing$. Observe that $(\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t) \cap (\bigcap_{t\in\mathcal{L}''} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}''} \mathcal{G}_t) = 0$ in this case, that is, $\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t$ and $\bigcap_{t\in\mathcal{L}''} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}''} \mathcal{G}_t$ arise by different Poisson events along $\mathcal{T}$. The independence follows. □

COROLLARY 5.7. *For the same situation as in Proposition 5.5, if $\mathcal{L}' \cap \mathcal{M}' = \mathcal{L}'' \cap \mathcal{M}'' = \varnothing$,*

$$\mathbb{COV}\biggl[\biggl|\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t\biggr|, \biggl|\bigcap_{t\in\mathcal{L}''} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}''} \mathcal{G}_t\biggr| \biggl| \mathcal{T}\biggr]$$
$$= \mathbb{V}\biggl[\biggl|\bigcap_{t\in\mathcal{L}'\cup\mathcal{L}''} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'\cup\mathcal{M}''} \mathcal{G}_t\biggr| \biggl| \mathcal{T}\biggr].$$

PROOF. We write

$$\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'} \mathcal{G}_t = \Bigl(\bigcap_{t\in\mathcal{L}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'\cup\mathcal{M}''\cup\mathcal{L}''} \mathcal{G}_t\Bigr) \uplus \Bigl(\bigcap_{t\in\mathcal{L}'\cup\mathcal{L}''} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'\cup\mathcal{M}''} \mathcal{G}_t\Bigr)$$
$$\uplus \Bigl(\bigcap_{t\in\mathcal{L}'\cup\mathcal{L}''\cup\mathcal{M}''} \mathcal{G}_t \setminus \bigcap_{t\in\mathcal{M}'} \mathcal{G}_t\Bigr),$$
$$\bigcap_{t\in\mathcal{L}''} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}''} \mathcal{G}_t = \Bigl(\bigcap_{t\in\mathcal{L}''} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'\cup\mathcal{M}''\cup\mathcal{L}'} \mathcal{G}_t\Bigr) \uplus \Bigl(\bigcap_{t\in\mathcal{L}'\cup\mathcal{L}''} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}'\cup\mathcal{M}''} \mathcal{G}_t\Bigr)$$
$$\uplus \Bigl(\bigcap_{t\in\mathcal{L}'\cup\mathcal{L}''\cup\mathcal{M}'} \mathcal{G}_t \setminus \bigcup_{t\in\mathcal{M}''} \mathcal{G}_t\Bigr).$$

By the independence statement in Proposition 5.5, only the covariances of the two second terms in both equalities do not vanish. The result follows. □

5.2. *Gene content for two individuals.* The simplest case in Proposition 5.5 arises if $\mathcal{T}$ has only two leaves. This case was already studied in the proof of Proposition 4.4. We extend our analysis by the next result.

PROPOSITION 5.8 (Gene content for two individuals). *For $1 \le i \neq j \le n$,*

$$\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j|] = \frac{\theta}{1+\rho},$$
$$\mathbb{V}[|\mathcal{G}_i \setminus \mathcal{G}_j|] = \frac{\theta^2}{(1+\rho)^2(1+2\rho)} + \frac{\theta}{1+\rho},$$
$$\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_j \setminus \mathcal{G}_i|] = \frac{\theta^2}{(1+\rho)^2(1+2\rho)}.$$

PROOF. We use Proposition 5.5. It suffices to assume that $\mathcal{T}$ is a tree connecting individuals i and j, that is, $\mathcal{L} = \{i, j\}$. First we assume that the coalescence time T of the two individuals is given. Under this assumption, Proposition 5.5 tells us that

$$\mathcal{G}_i \setminus \mathcal{G}_j \sim \mathcal{POI}\biggl(\frac{\theta}{\rho}(1 - e^{-\rho T}) \cdot \lambda_I\biggr)$$

and, using the fact that expectation and variance are equal for a Poisson distributed random variable,

$$\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||T] = \mathbb{V}[|\mathcal{G}_i \setminus \mathcal{G}_j||T] = \frac{\theta}{\rho}(1 - e^{-\rho T}),$$

such that we obtain

$$\begin{aligned}
\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j|] &= \mathbb{E}\bigl[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||T]\bigr] = \mathbb{E}\biggl[\frac{\theta}{\rho}(1 - e^{-\rho T})\biggr] = \frac{\theta}{\rho}\frac{\rho}{1+\rho} = \frac{\theta}{1+\rho},\\
\mathbb{V}[|\mathcal{G}_i \setminus \mathcal{G}_j|] &= \mathbb{V}\bigl[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||T]\bigr] + \mathbb{E}\bigl[\mathbb{V}[|\mathcal{G}_i \setminus \mathcal{G}_j||T]\bigr]\\
&= \mathbb{V}\biggl[\frac{\theta}{\rho}(1 - e^{-\rho T})\biggr] + \mathbb{E}\biggl[\frac{\theta}{\rho}(1 - e^{-\rho T})\biggr]\\
&= \frac{\theta^2}{\rho^2}\biggl(\frac{1}{1+2\rho} - \frac{1}{(1+\rho)^2}\biggr) + \frac{\theta}{\rho}\biggl(1 - \frac{1}{1+\rho}\biggr)\\
&= \frac{\theta^2}{(1+\rho)^2(1+2\rho)} + \frac{\theta}{1+\rho}.
\end{aligned}$$

In addition, given T, $\mathcal{G}_i \setminus \mathcal{G}_j$ and $\mathcal{G}_j \setminus \mathcal{G}_i$ are independent by Corollary 5.7. Hence,

$$\begin{aligned}
\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_j \setminus \mathcal{G}_i|] &= \mathbb{COV}\bigl[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||T], \mathbb{E}[|\mathcal{G}_j \setminus \mathcal{G}_i||T]\bigr]\\
&= \mathbb{V}\biggl[\frac{\theta}{\rho}(1 - e^{-\rho T})\biggr] = \frac{\theta^2}{(1+\rho)^2(1+2\rho)}. \qquad \square
\end{aligned}$$

5.3. *Gene content for three individuals.* Similar to Proposition 5.8, we use the general setting of Proposition 5.5 in order to prove results about the joint distribution of gene content in three individuals.

PROPOSITION 5.9 (Gene content for three individuals). *For $i, j, k \in \{1, \dots, n\}$ pairwise different*,

$$\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_i \setminus \mathcal{G}_k|] = \frac{\theta^2}{(1+\rho)^2(1+2\rho)(3+2\rho)} + \frac{\theta}{2+\rho}, \tag{5.2}$$

$$\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_i|] = \frac{\theta^2}{(1+\rho)^2(1+2\rho)(3+2\rho)}, \tag{5.3}$$

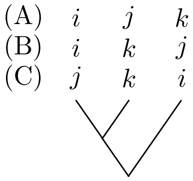

FIG. 5. *The* 3 *cases for a genealogical tree connecting three individuals* i, j, k.

$$\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_j \setminus \mathcal{G}_k|] = \frac{\theta^2}{(1+\rho)^2(1+2\rho)(3+2\rho)}, \tag{5.4}$$

$$\begin{aligned}\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_j|] &= \frac{\theta^2}{(1+\rho)^2(1+2\rho)(3+2\rho)} \\ &\quad + \frac{\theta}{(1+\rho)(2+\rho)}.\end{aligned} \tag{5.5}$$

PROOF. We use Proposition 5.5 again. Let $\mathcal{T}$ be the tree connecting three individuals i, j and k, that is, $\mathcal{L} = \{i, j, k\}$. Assume the random times T_2, T_3 during which the coalescent has 2, 3 lines, respectively, and one of the three possible tree topologies, illustrated in Figure 5, are given. We use

$$\begin{aligned}&\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_i \setminus \mathcal{G}_k|] \\ &\quad = \mathbb{COV}[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}], \mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_k||\mathcal{T}]] \\ &\qquad + \mathbb{E}[\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_i \setminus \mathcal{G}_k||\mathcal{T}]]\end{aligned} \tag{5.6}$$

and similar equalities for the other cases. We compute both parts of the right-hand side separately. For the first part we need to calculate $\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}]$ depending on $\mathcal{T}$:

1. $\mathcal{T} \in \{(A)\}$:

$$\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}] = \int_0^{2T_3} \frac{\theta}{2} e^{-\rho/2t}\,dt = \frac{\theta}{\rho}(1 - e^{-\rho T_3});$$

2. $\mathcal{T} \in \{(B), (C)\}$:

$$\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}] = \int_0^{2T_2+2T_3} \frac{\theta}{2} e^{-\rho/2t}\,dt = \frac{\theta}{\rho}(1 - e^{-\rho(T_2+T_3)}).$$

Replacing the pair ij in the last to expressions by ik, jk, ki or kj leads to the same possibilities arising in the genealogies (A), (B), (C). We collect all possibilities in Table 1.

TABLE 1
The three different tree topologies from Figure 5 give rise to two different terms for the conditional expectation of a pair of leaves, depending on the labeling of the pair

	(*A*)	**(*B*)**	**(*C*)**
ij	1.	2.	2.
ji	1.	2.	2.
ik	2.	1.	2.
jk	2.	2.	1.
ki	2.	1.	2.
kj	2.	2.	1.

In Proposition 5.8 we have seen that $\mathbb{E}[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}] = \mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j|] = \frac{\theta}{1+\rho}$ and therefore,

$$\begin{aligned}
&\mathbb{COV}[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}], \mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_k||\mathcal{T}]] \\
&\quad = \frac{2}{3}\frac{\theta^2}{\rho^2}\mathbb{E}\big[\big(1-e^{-\rho(T_2+T_3)}\big)\big(1-e^{-\rho T_3}\big)\big] \\
&\qquad + \frac{1}{3}\frac{\theta^2}{\rho^2}\mathbb{E}\big[\big(1-e^{-\rho(T_2+T_3)}\big)^2\big] - \frac{\theta^2}{(1+\rho)^2} \\
&\quad = \frac{\theta^2}{3\rho^2}\bigg(\frac{2\rho}{3+\rho} - \frac{6}{(1+\rho)(3+\rho)} + \frac{6}{(1+\rho)(3+2\rho)} \\
&\qquad\qquad + 1 - \frac{6}{(1+\rho)(3+\rho)} + \frac{1}{(1+2\rho)(3+2\rho)}\bigg) - \frac{\theta^2}{(1+\rho)^2} \\
&\quad = \frac{\theta^2}{(1+\rho)^2(1+2\rho)(3+2\rho)}.
\end{aligned}$$

Note that this equation also holds for the other three cases in Proposition 5.9, that is, we have computed the first term in (5.6) for all combinations of i, j, k arising in the proposition.

Let us now consider the second part of (5.6). From Corollary 5.7 we see that

$$\mathbb{E}[\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_i||\mathcal{T}]] = \mathbb{E}[\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_j \setminus \mathcal{G}_k||\mathcal{T}]] = 0,$$

which already gives assertions (5.3) and (5.4). Moreover, Corollary 5.7 gives

$$\begin{aligned}
\mathbb{E}[\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_i \setminus \mathcal{G}_k||\mathcal{T}]] &= \mathbb{E}[\mathbb{V}[|\mathcal{G}_i \setminus (\mathcal{G}_j \cup \mathcal{G}_k)||\mathcal{T}]], \\
\mathbb{E}[\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_j||\mathcal{T}]] &= \mathbb{E}[\mathbb{V}[|\mathcal{G}_i \cap \mathcal{G}_k \setminus \mathcal{G}_j||\mathcal{T}]].
\end{aligned}$$

From Proposition 5.5 we know that for given $\mathcal{T}$, $|\mathcal{G}_i \setminus (\mathcal{G}_j \cup \mathcal{G}_k)|$ and $|\mathcal{G}_i \cap \mathcal{G}_k \setminus \mathcal{G}_j|$ are Poisson distributed. Note that $|\mathcal{G}_i \setminus (\mathcal{G}_j \cup \mathcal{G}_k)|$ is the number of genes present in i, but not in j and k. Recalling that $G_k^{(n)}$ denotes the number of genes present in k out of n individuals, it is clear that $\mathbb{E}[|\mathcal{G}_i \setminus (\mathcal{G}_j \cup \mathcal{G}_k)|] = \frac{1}{3}\mathbb{E}[G_1^{(3)}]$, and so using Theorem 5

$$\begin{aligned}\mathbb{E}\big[\mathbb{V}[|\mathcal{G}_i \setminus (\mathcal{G}_j \cup \mathcal{G}_k)||\mathcal{T}]\big] &= \mathbb{E}\big[\mathbb{E}[|\mathcal{G}_i \setminus (\mathcal{G}_j \cup \mathcal{G}_k)||\mathcal{T}]\big] \\ &= \mathbb{E}[|\mathcal{G}_i \setminus (\mathcal{G}_j \cup \mathcal{G}_k)|] = \frac{\theta}{2+\rho}.\end{aligned}$$

Equivalently, with $\mathbb{E}[|\mathcal{G}_i \cap \mathcal{G}_k \setminus \mathcal{G}_j|] = \frac{1}{3}\mathbb{E}[G_2^{(3)}]$,

$$\mathbb{E}\big[\mathbb{V}[|\mathcal{G}_i \cap \mathcal{G}_k \setminus \mathcal{G}_j||\mathcal{T}]\big] = \mathbb{E}[|\mathcal{G}_i \cap \mathcal{G}_k \setminus \mathcal{G}_j|] = \frac{\theta}{(2+\rho)(1+\rho)}. \qquad \square$$

5.4. *Gene content for pairs of two individuals.*

PROPOSITION 5.10 (Gene content for pairs of two individuals). *For $i, j, k, l \in \{1, \dots, n\}$ pairwise different*

$$\begin{aligned}&\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_l|] \\ &\quad = \frac{\theta}{(3+\rho)(2+\rho)} + \frac{2\theta^2}{(1+\rho)^2(3+\rho)(1+2\rho)(3+2\rho)}.\end{aligned}$$

PROOF. The proof is similar to the proof of Proposition 5.9. Analogously to (5.6) we use

$$\begin{aligned}\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_l|] &= \mathbb{COV}\big[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}], \mathbb{E}[|\mathcal{G}_k \setminus \mathcal{G}_l||\mathcal{T}]\big] \\ &\quad + \mathbb{E}\big[\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_l||\mathcal{T}]\big].\end{aligned} \tag{5.7}$$

As $\mathbb{E}[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}]] = \mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j|] = \frac{\theta}{1+\rho}$ we get that

$$\begin{aligned}&\mathbb{COV}\big[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}], \mathbb{E}[|\mathcal{G}_k \setminus \mathcal{G}_l||\mathcal{T}]\big] \\ &\quad = \mathbb{E}\big[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}] \cdot \mathbb{E}[|\mathcal{G}_k \setminus \mathcal{G}_l||\mathcal{T}]\big] - \frac{\theta^2}{(1+\rho)^2}.\end{aligned}$$

Therefore, four different cases occur depending on the topology of the tree seen in Figure 6:

1. $\mathcal{T} \in \{(C), (D), (E), (F), (G), (H), (I), (J)\}$:

$$\frac{\rho^2}{\theta^2}\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}] \cdot \mathbb{E}[|\mathcal{G}_k \setminus \mathcal{G}_l||\mathcal{T}] = \big(1 - e^{-\rho(T_3+T_4)}\big)\big(1 - e^{-\rho(T_2+T_3+T_4)}\big);$$

2. $\mathcal{T} \in \{(A), (B), (K), (L)\}$:

$$\frac{\rho^2}{\theta^2}\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}] \cdot \mathbb{E}[|\mathcal{G}_k \setminus \mathcal{G}_l||\mathcal{T}] = (1 - e^{-\rho T_4})\big(1 - e^{\rho(T_2+T_3+T_4)}\big);$$

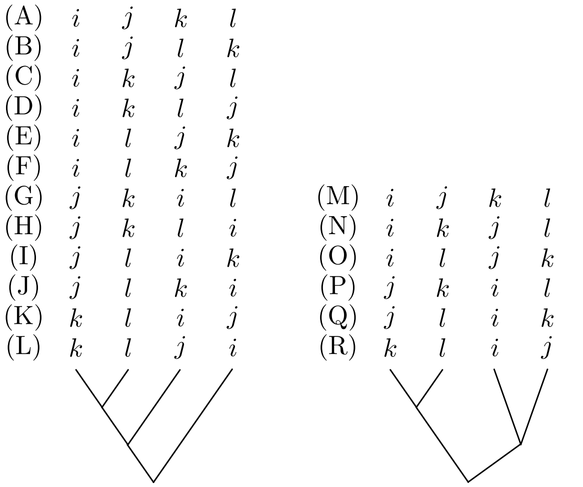

FIG. 6. *The* 18 *equally probable topologies for a genealogical tree connecting four individuals* i, j, k, l.

3. $\mathcal{T} \in \{(M), (R)\}$:

$$\frac{\rho^2}{\theta^2}\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}] \cdot \mathbb{E}[|\mathcal{G}_k \setminus \mathcal{G}_l||\mathcal{T}] = (1 - e^{-\rho T_4})\big(1 - e^{-\rho(T_3+T_4)}\big);$$

4. $\mathcal{T} \in \{(N), (O), (P), (Q)\}$:

$$\frac{\rho^2}{\theta^2}\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}] \cdot \mathbb{E}[|\mathcal{G}_k \setminus \mathcal{G}_l||\mathcal{T}] = \big(1 - e^{-\rho(T_2+T_3+T_4)}\big)^2.$$

Hence, with a little help from MATHEMATICA,

$$\begin{aligned}
&\mathbb{COV}[\mathbb{E}[|\mathcal{G}_i \setminus \mathcal{G}_j||\mathcal{T}], \mathbb{E}[|\mathcal{G}_k \setminus \mathcal{G}_l||\mathcal{T}]] \\
&\quad = \frac{\theta^2}{\rho^2}\bigg(\frac{8}{18}\bigg(1 - \frac{18}{(3+\rho)(6+\rho)} - \frac{18}{(1+\rho)(3+\rho)(6+\rho)} \\
&\qquad\qquad + \frac{18}{(1+\rho)(3+2\rho)(6+2\rho)}\bigg) \\
&\qquad + \frac{4}{18}\bigg(1 - \frac{6}{6+\rho} - \frac{18}{(1+\rho)(3+\rho)(6+\rho)} \\
&\qquad\qquad + \frac{18}{(1+\rho)(3+\rho)(6+2\rho)}\bigg) \\
&\qquad + \frac{2}{18}\bigg(1 - \frac{6}{6+\rho} - \frac{18}{(3+\rho)(6+\rho)} + \frac{18}{(3+\rho)(6+2\rho)}\bigg) \qquad (5.8)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{4}{18}\biggl(1-\frac{36}{(1+\rho)(3+\rho)(6+\rho)} \\
&\qquad\qquad +\frac{18}{(1+2\rho)(3+2\rho)(6+2\rho)}\biggr)\biggr)-\frac{\theta^2}{(1+\rho)^2} \\
&=\frac{2\theta^2}{(1+\rho)^2(3+\rho)(1+2\rho)(3+2\rho)}.
\end{aligned}
$$

For the second term, Corollary 5.7 gives

$$\mathbb{E}\bigl[\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_k|, |\mathcal{G}_j \setminus \mathcal{G}_l||\mathcal{T}]\bigr]=\mathbb{E}[\mathbb{V}[D_{ij,kl}|\mathcal{T}]]$$

with $D_{ij,kl}$ as in (3.3). Given $\mathcal{T}$, $D_{ij,kl}$ is Poisson distributed, hence we obtain from (3.4)

$$\mathbb{E}[\mathbb{V}[D_{ij,kl}|\mathcal{T}]]=\mathbb{E}[\mathbb{E}[D_{ij,kl}|\mathcal{T}]]=\frac{\theta}{(3+\rho)(2+\rho)}. \tag{5.9}$$

Combining (5.7) with (5.8) and (5.9) gives the result. □

5.5. *Proof of Theorem* 2. Using Propositions 5.8, 5.9 and 5.10, it is now easy to prove Theorem 2. We obtain

$$
\begin{aligned}
&n^2(n-1)^2\mathbb{V}[D] \\
&\quad=\sum_{i\neq j}(\mathbb{V}[|\mathcal{G}_i \setminus \mathcal{G}_j|]+\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_j \setminus \mathcal{G}_i|]) \\
&\qquad+\sum_{i,j,k \text{ pwd}}(\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_i \setminus \mathcal{G}_k|]+\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_j \setminus \mathcal{G}_k|] \\
&\qquad\qquad+\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_i|]+\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_j|]) \\
&\qquad+\sum_{i,j,k,l \text{ pwd}}\mathbb{COV}[|\mathcal{G}_i \setminus \mathcal{G}_j|, |\mathcal{G}_k \setminus \mathcal{G}_l|] \\
&\quad=n(n-1)(\mathbb{V}[|\mathcal{G}_1 \setminus \mathcal{G}_2|]+\mathbb{COV}[|\mathcal{G}_1 \setminus \mathcal{G}_2|, |\mathcal{G}_2 \setminus \mathcal{G}_1|]) \\
&\qquad+n(n-1)(n-2)(\mathbb{COV}[|\mathcal{G}_1 \setminus \mathcal{G}_2|, |\mathcal{G}_1 \setminus \mathcal{G}_3|] \\
&\qquad\qquad+\mathbb{COV}[|\mathcal{G}_1 \setminus \mathcal{G}_2|, |\mathcal{G}_2 \setminus \mathcal{G}_3|] \\
&\qquad\qquad+\mathbb{COV}[|\mathcal{G}_1 \setminus \mathcal{G}_2|, |\mathcal{G}_3 \setminus \mathcal{G}_1|] \\
&\qquad\qquad+\mathbb{COV}[|\mathcal{G}_1 \setminus \mathcal{G}_2|, |\mathcal{G}_3 \setminus \mathcal{G}_2|]) \\
&\qquad+n(n-1)(n-2)(n-3)\mathbb{COV}[|\mathcal{G}_1 \setminus \mathcal{G}_2|, |\mathcal{G}_3 \setminus \mathcal{G}_4|],
\end{aligned}
$$

and the result follows by some application of MATHEMATICA.

6. Proof of Theorem 3. We denote by $\mathcal{T}$ the genealogy connecting the individuals i, j, k, l. As above, we note that $\mathcal{T}$ is uniquely given by the random times T_2, T_3, T_4 during which the coalescent has $2, 3, 4$ lines, respectively, and the tree topology, distinguished by 18 equally probably cases, illustrated in Figure 6. We use

$$\begin{aligned}\mathbb{E}[D_{ij,kl} \cdot D_{ik,jl}] = {} & \mathbb{E}[\mathbb{COV}[D_{ij,kl}, D_{ik,jl}|\mathcal{T}]] \\ & + \mathbb{E}\big[\mathbb{E}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}]\big]\end{aligned} \tag{6.1}$$

and note that

$$\mathbb{COV}[D_{ij,kl}, D_{ik,jl}|\mathcal{T}] = 0$$

by Corollary 5.7. So, we are left with computing the second term in (6.1). The terms $\mathbb{E}[D_{ij,kl}|\mathcal{T}]$ can take six different values, depending on $\mathcal{T}$. We use Proposition 5.5:

1. $\mathcal{T} \in \{(A), (B)\}$:

$$\begin{aligned}&\frac{\rho}{\theta}\mathbb{E}[D_{ij,kl}|\mathcal{T}] \\ &\quad = e^{(-\rho/2)2T_4}\big(1 - e^{-\rho/2T_3} + e^{-\rho/2T_3}\big(1 - e^{-\rho/2(T_3+T_4)}\big) \\ &\qquad\qquad \times \big(1 - e^{-\rho/2(2T_2+T_3+T_4)}\big)\big) \\ &\quad = e^{(-\rho/2)2T_4} - e^{-\rho/2(2T_3+3T_4)} - e^{-\rho/2(2T_2+2T_3+3T_4)} \\ &\qquad + e^{-\rho/2(2T_2+3T_3+4T_4)};\end{aligned}$$

2. $\mathcal{T} \in \{(C), (E), (G), (I)\}$:

$$\begin{aligned}\frac{\rho}{\theta}\mathbb{E}[D_{ij,kl}|\mathcal{T}] &= e^{-\rho/2(2T_3+2T_4)}(1 - e^{-\rho/2T_4})\big(1 - e^{-\rho/2(2T_2+T_3+T_4)}\big) \\ &= e^{-\rho/2(2T_3+2T_4)} - e^{-\rho/2(2T_3+3T_4)} - e^{-\rho/2(2T_2+3T_3+3T_4)} \\ &\quad + e^{-\rho/2(2T_2+3T_3+4T_4)};\end{aligned}$$

3. $\mathcal{T} \in \{(D), (F), (G), (H), (J), (N), (O), (P), (Q)\}$:

$$\begin{aligned}\frac{\rho}{\theta}\mathbb{E}[D_{ij,kl}|\mathcal{T}] &= e^{-\rho/2(2T_2+2T_3+2T_4)}(1 - e^{-\rho/2T_4})\big(1 - e^{-\rho/2(T_3+T_4)}\big) \\ &= e^{-\rho/2(2T_2+2T_3+2T_4)} - e^{-\rho/2(2T_2+2T_3+3T_4)} \\ &\quad - e^{-\rho/2(2T_2+3T_3+3T_4)} + e^{-\rho/2(2T_2+3T_3+4T_4)};\end{aligned}$$

4. $\mathcal{T} \in \{(K), (L)\}$:

$$\begin{aligned}\frac{\rho}{\theta}\mathbb{E}[D_{ij,kl}|\mathcal{T}] &= e^{-\rho/2(2T_2+2T_3+2T_4)}\big(1 - 2e^{-\rho/2(T_3+T_4)} + e^{-\rho/2(T_3+2T_4)}\big) \\ &= e^{-\rho/2(2T_2+2T_3+2T_4)} - 2e^{-\rho/2(2T_2+3T_3+3T_4)} \\ &\quad + e^{-\rho/2(2T_2+3T_3+4T_4)};\end{aligned}$$

TABLE 2

For every tree topology of Figure 6, the two pairs ij, kl as well as ik, jl fall into one of the six cases for the conditional expectation; see below (6.1)

	(*A*)	**(*B*)**	**(*C*)**	**(*D*)**	**(*E*)**	**(*F*)**	**(*G*)**	**(*H*)**	**(*I*)**
ij, kl	1.	1.	2.	3.	2.	3.	2.	3.	2.
ik, jl	2.	3.	1.	1.	3.	2.	2.	3.	4.
	(*J*)	**(*K*)**	**(*L*)**	**(*M*)**	**(*N*)**	**(*O*)**	**(*P*)**	**(*Q*)**	**(*R*)**
ij, kl	3.	4.	4.	5.	3.	3.	3.	3.	6.
ik, jl	4.	2.	3.	3.	5.	3.	3.	6.	3.

5. $\mathcal{T} \in \{(M)\}$:

$$\frac{\rho}{\theta}\mathbb{E}[D_{ij,kl}|\mathcal{T}] = e^{(-\rho/2)2T_4}\bigl(1 - 2e^{-\rho/2(2T_2+2T_3+T_4)} + e^{-\rho/2(2T_2+3T_3+2T_4)}\bigr)$$
$$= e^{(-\rho/2)2T_4} - 2e^{-\rho/2(2T_2+2T_3+3T_4)} + e^{-\rho/2(2T_2+3T_3+4T_4)};$$

6. $\mathcal{T} \in \{(R)\}$:

$$\frac{\rho}{\theta}\mathbb{E}[D_{ij,kl}|\mathcal{T}] = e^{-\rho/2(2T_3+2T_4)}\bigl(1 - 2e^{-\rho/2(2T_2+T_3+T_4)} + e^{-\rho/2(2T_2+T_3+2T_4)}\bigr)$$
$$= e^{-\rho/2(2T_3+2T_4)} - 2e^{-\rho/2(2T_2+3T_3+3T_4)} + e^{-\rho/2(2T_2+3T_3+4T_4)}.$$

Relabeling i, j, k, l by i, k, j, l changes these cases. Table 2 gives the responsible terms for $\mathbb{E}[D_{ij,kl}|\mathcal{T}]$ and $\mathbb{E}[D_{ik,jl}|\mathcal{T}]$ for all 18 possible tree topologies. We obtain nine cases for which to compute the second term in (6.1). We abbreviate $\widetilde{e} := e^{-\rho/2}$.

$\mathcal{T} \in \{(A), (C)\}$:

$$\frac{\rho^2}{\theta^2}\mathbb{E}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}]$$
$$\begin{aligned} &= \widetilde{e}^{2T_3+4T_4} - \widetilde{e}^{2T_3+5T_4} - \widetilde{e}^{2T_2+3T_3+5T_4} + \widetilde{e}^{2T_2+3T_3+6T_4} \\ &\quad - \widetilde{e}^{4T_3+5T_4} + \widetilde{e}^{4T_3+6T_4} + \widetilde{e}^{2T_2+5T_3+6T_4} - \widetilde{e}^{2T_2+5T_3+7T_4} \\ &\quad - \widetilde{e}^{2T_2+4T_3+5T_4} + \widetilde{e}^{2T_2+4T_3+6T_4} + \widetilde{e}^{4T_2+5T_3+6T_4} - \widetilde{e}^{4T_2+5T_3+7T_4} \\ &\quad + \widetilde{e}^{2T_2+5T_3+6T_4} - \widetilde{e}^{2T_2+5T_3+7T_4} - \widetilde{e}^{4T_2+6T_3+7T_4} + \widetilde{e}^{4T_2+6T_3+8T_4}; \end{aligned}$$

$\mathcal{T} \in \{(B), (D)\}$:

$$\frac{\rho^2}{\theta^2}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}]$$
$$= \widetilde{e}^{2T_2+2T_3+4T_4} - \widetilde{e}^{2T_2+2T_3+5T_4} - \widetilde{e}^{2T_2+3T_3+5T_4} + \widetilde{e}^{2T_2+3T_3+6T_4}$$

$$
\begin{aligned}
&- \widetilde{e}^{2T_2+4T_3+5T_4} + \widetilde{e}^{2T_2+4T_3+6T_4} + \widetilde{e}^{2T_2+5T_3+6T_4} - \widetilde{e}^{2T_2+5T_3+7T_4}\\
&- \widetilde{e}^{4T_2+4T_3+5T_4} + \widetilde{e}^{4T_2+4T_3+6T_4} + \widetilde{e}^{4T_2+5T_3+6T_4} - \widetilde{e}^{4T_2+5T_3+7T_4}\\
&+ \widetilde{e}^{4T_2+5T_3+6T_4} - \widetilde{e}^{4T_2+5T_3+7T_4} - \widetilde{e}^{4T_2+6T_3+7T_4} + \widetilde{e}^{4T_2+6T_3+8T_4};
\end{aligned}
$$

$\mathcal{T} \in \{(E), (F)\}$:

$$
\begin{aligned}
&\frac{\rho^2}{\theta^2}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}]\\
&\quad = \widetilde{e}^{2T_2+4T_3+4T_4} - \widetilde{e}^{2T_2+4T_3+5T_4} - \widetilde{e}^{2T_2+5T_3+5T_4} + \widetilde{e}^{2T_2+5T_3+6T_4}\\
&\qquad - \widetilde{e}^{2T_2+4T_3+5T_4} + \widetilde{e}^{2T_2+4T_3+6T_4} + \widetilde{e}^{2T_2+5T_3+6T_4} - \widetilde{e}^{2T_2+5T_3+7T_4}\\
&\qquad - \widetilde{e}^{4T_2+5T_3+5T_4} + \widetilde{e}^{4T_2+5T_3+6T_4} + \widetilde{e}^{4T_2+6T_3+6T_4} - \widetilde{e}^{4T_2+6T_3+7T_4}\\
&\qquad + \widetilde{e}^{4T_2+5T_3+6T_4} - \widetilde{e}^{4T_2+5T_3+7T_4} - \widetilde{e}^{4T_2+6T_3+7T_4} + \widetilde{e}^{4T_2+6T_3+8T_4};
\end{aligned}
$$

$\mathcal{T} \in \{(G)\}$:

$$
\begin{aligned}
&\frac{\rho^2}{\theta^2}\mathbb{E}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}]\\
&\quad = \widetilde{e}^{4T_3+4T_4} - \widetilde{e}^{4T_3+5T_4} - \widetilde{e}^{2T_2+5T_3+5T_4} + \widetilde{e}^{2T_2+5T_3+6T_4}\\
&\qquad - \widetilde{e}^{4T_3+5T_4} + \widetilde{e}^{4T_3+6T_4} + \widetilde{e}^{2T_2+5T_3+6T_4} - \widetilde{e}^{2T_2+5T_3+7T_4}\\
&\qquad - \widetilde{e}^{2T_2+5T_3+5T_4} + \widetilde{e}^{2T_2+5T_3+6T_4} + \widetilde{e}^{4T_2+6T_3+6T_4} - \widetilde{e}^{4T_2+6T_3+7T_4}\\
&\qquad + \widetilde{e}^{2T_2+5T_3+6T_4} - \widetilde{e}^{2T_2+5T_3+7T_4} - \widetilde{e}^{4T_2+6T_3+7T_4} + \widetilde{e}^{4T_2+6T_3+8T_4};
\end{aligned}
$$

$\mathcal{T} \in \{(H), (O), (P)\}$:

$$
\begin{aligned}
&\frac{\rho^2}{\theta^2}\mathbb{E}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}]\\
&\quad = \widetilde{e}^{4T_2+4T_3+4T_4} - \widetilde{e}^{4T_2+4T_3+5T_4} - \widetilde{e}^{4T_2+5T_3+5T_4} + \widetilde{e}^{4T_2+5T_3+6T_4}\\
&\qquad - \widetilde{e}^{4T_2+4T_3+5T_4} + \widetilde{e}^{4T_2+4T_3+6T_4} + \widetilde{e}^{4T_2+5T_3+6T_4} - \widetilde{e}^{4T_2+5T_3+7T_4}\\
&\qquad - \widetilde{e}^{4T_2+5T_3+5T_4} + \widetilde{e}^{4T_2+5T_3+6T_4} + \widetilde{e}^{4T_2+6T_3+6T_4} - \widetilde{e}^{4T_2+6T_3+7T_4}\\
&\qquad + \widetilde{e}^{4T_2+5T_3+6T_4} - \widetilde{e}^{4T_2+5T_3+7T_4} - \widetilde{e}^{4T_2+6T_3+7T_4} + \widetilde{e}^{4T_2+6T_3+8T_4};
\end{aligned}
$$

$\mathcal{T} \in \{(I), (K)\}$:

$$
\begin{aligned}
&\frac{\rho^2}{\theta^2}\mathbb{E}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}]\\
&\quad = \widetilde{e}^{2T_2+4T_3+4T_4} - 2\widetilde{e}^{2T_2+5T_3+5T_4} + \widetilde{e}^{2T_2+5T_3+6T_4}\\
&\qquad - \widetilde{e}^{2T_2+4T_3+5T_4} + 2\widetilde{e}^{2T_2+5T_3+6T_4} - \widetilde{e}^{2T_2+5T_3+7T_4}\\
&\qquad - \widetilde{e}^{4T_2+5T_3+5T_4} + 2\widetilde{e}^{4T_2+6T_3+6T_4} - \widetilde{e}^{4T_2+6T_3+7T_4}\\
&\qquad + \widetilde{e}^{4T_2+5T_3+6T_4} - 2\widetilde{e}^{4T_2+6T_3+7T_4} + \widetilde{e}^{4T_2+6T_3+8T_4};
\end{aligned}
$$

$\mathcal{T} \in \{(J), (L)\}$:

$$\begin{aligned}
&\frac{\rho^2}{\theta^2}\mathbb{E}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}] \\
&\quad = \tilde{e}^{4T_2+4T_3+4T_4} - 2\tilde{e}^{4T_2+5T_3+5T_4} + \tilde{e}^{4T_2+5T_3+6T_4} \\
&\qquad - \tilde{e}^{4T_2+4T_3+5T_4} + 2\tilde{e}^{4T_2+5T_3+6T_4} - \tilde{e}^{4T_2+5T_3+7T_4} \\
&\qquad - \tilde{e}^{4T_2+5T_3+5T_4} + 2\tilde{e}^{4T_2+6T_3+6T_4} - \tilde{e}^{4T_2+6T_3+7T_4} \\
&\qquad + \tilde{e}^{4T_2+5T_3+6T_4} - 2\tilde{e}^{4T_2+6T_3+7T_4} + \tilde{e}^{4T_2+6T_3+8T_4};
\end{aligned}$$

$\mathcal{T} \in \{(M), (N)\}$:

$$\begin{aligned}
&\frac{\rho^2}{\theta^2}\mathbb{E}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}] \\
&\quad = \tilde{e}^{2T_2+2T_3+4T_4} - 2\tilde{e}^{4T_2+4T_3+5T_4} + \tilde{e}^{4T_2+5T_3+6T_4} \\
&\qquad - \tilde{e}^{2T_2+2T_3+5T_4} + 2\tilde{e}^{4T_2+4T_3+6T_4} - \tilde{e}^{4T_2+5T_3+7T_4} \\
&\qquad - \tilde{e}^{2T_2+3T_3+5T_4} + 2\tilde{e}^{4T_2+5T_3+6T_4} - \tilde{e}^{4T_2+6T_3+7T_4} \\
&\qquad + \tilde{e}^{2T_2+3T_3+6T_4} - 2\tilde{e}^{4T_2+5T_3+7T_4} + \tilde{e}^{4T_2+6T_3+8T_4};
\end{aligned}$$

$\mathcal{T} \in \{(Q), (R)\}$:

$$\begin{aligned}
&\frac{\rho^2}{\theta^2}\mathbb{E}[D_{ij,kl}|\mathcal{T}] \cdot \mathbb{E}[D_{ik,jl}|\mathcal{T}] \\
&\quad = \tilde{e}^{2T_2+4T_3+4T_4} - 2\tilde{e}^{4T_2+5T_3+5T_4} + \tilde{e}^{4T_2+5T_3+6T_4} \\
&\qquad - \tilde{e}^{2T_2+4T_3+5T_4} + 2\tilde{e}^{4T_2+5T_3+6T_4} - \tilde{e}^{4T_2+5T_3+7T_4} \\
&\qquad - \tilde{e}^{2T_2+5T_3+5T_4} + 2\tilde{e}^{4T_2+6T_3+6T_4} - \tilde{e}^{4T_2+6T_3+7T_4} \\
&\qquad + \tilde{e}^{2T_2+5T_3+6T_4} - 2\tilde{e}^{4T_2+6T_3+7T_4} + \tilde{e}^{4T_2+6T_3+8T_4}.
\end{aligned}$$

Combining the last equations and using MATHEMATICA, we obtain the desired result.

7. Proof of Theorems 4 and 5. Theorems 4 and 5 provide expectations for the size of the dispensable genome and the gene frequency spectrum, respectively. Recalling that $G_k^{(n)}$ is the number of genes in frequency $k = 1, \dots, n$ in a sample of size n, it is clear that

$$G = \sum_{k=1}^{n} G_k^{(n)},$$

where G is given by (3.5). In addition, if $\mathbb{E}[G_1^{(i)}]$, $i = 1, \dots,$ is known from Theorem 5,

$$\mathbb{E}[G] = \sum_{i=1}^{n} \frac{1}{i} \mathbb{E}[G_1^{(i)}] = \sum_{i=1}^{n} \frac{1}{i} \frac{\theta i}{i-1+\rho} = \theta \sum_{i=0}^{n-1} \frac{1}{i+\rho}.$$

Hence, the result for the expected number of genes in Theorem 4 can easily be proved once we have established Theorem 5. However, we take an alternative route and give an independent proof of Theorem 4.

7.1. *An independent proof of Theorem* 4. Recall the survival function $p_{\mathcal{T}}$ from Definition 5.1. Consider the coalescent, denoted by $\mathcal{T}$, started with n lines, rooted at r, the most recent common ancestor of the sample. As shown in Proposition 5.2, $p_{\mathcal{T}} : \mathcal{T} \to [0, 1]$ gives the probability that a mutation that arises at $t \in \mathcal{T}$ is not lost in at least one leaf. Hence, given $\mathcal{T}$, we find that $G \sim \mathcal{POI}(\pi(\mathcal{T}))$ with

$$\pi(\mathcal{T}) := \frac{\theta}{2} \int_{\mathcal{T}} p_{\mathcal{T}}(t)\, dt.$$

Next, we consider a random coalescent $\mathcal{T}$ with additional *loss events* at rate $\frac{\rho}{2}$ along the tree. We say that $t \in \mathcal{T}$ is unlost if there is a leaf in $i \in \mathcal{T}$ such that the path $[t, i]$ is not hit by a loss event. Given $\mathcal{T}$, note that

$$p_{\mathcal{T}}(t) = \mathbb{P}[t \text{ unlost} | \mathcal{T}]$$

by Proposition 5.2.

To prove (3.10), we write immediately, using the above arguments,

$$\begin{aligned} \mathbb{E}[G] &= \mathbb{E}\left[\frac{\theta}{2} \int_{\mathcal{T}} p_{\mathcal{T}}(t)\, dt\right] = \mathbb{E}\left[\frac{\theta}{2} \int_{\mathcal{T}} 1(t \text{ not lost})\, dt\right] \\ &= \frac{\theta}{2} \mathbb{E}[\text{length of unlost lines in } \mathcal{T}]. \end{aligned} \tag{7.1}$$

To compute the expected length of unlost lines we note that all lines are unlost near the leaves. The number of unlost lines decreases either by a coalescence or by a gene loss event. When there are k unlost lines left, any line is lost at rate $\frac{\rho}{2}$ and two lines coalesce at rate $\binom{k}{2}$. Hence, the time until there are $k-1$ unlost lines is $\exp(\frac{\rho}{2}k + \binom{k}{2})$ distributed. Thus,

$$\mathbb{E}[G] = \frac{\theta}{2} \sum_{k=1}^{n} \frac{k}{\binom{k}{2} + \rho/2k} = \theta \sum_{k=1}^{n} \frac{1}{k-1+\rho} = \theta \sum_{k=0}^{n-1} \frac{1}{k+\rho},$$

and we have proven (3.10).

Next we show how to obtain the recursion for $\mathbb{V}[G]$ given in (3.11). Using the fact that, given $\mathcal{T}$, the number of genes G is Poisson distributed with rate $\pi(\mathcal{T})$,

$$\begin{aligned}\mathbb{V}[G] &= \mathbb{E}[\mathbb{E}[G^2|\mathcal{T}]] - \mathbb{E}[G]^2 = \mathbb{E}[\pi(\mathcal{T}) + \pi(\mathcal{T})^2] - \mathbb{E}[G]^2\\ &= \mathbb{E}[G] - \mathbb{E}[G]^2 + \frac{\theta^2}{4}\mathbb{E}\biggl[\int\int p_{\mathcal{T}}(s)p_{\mathcal{T}}(t)\,ds\,dt\biggr].\end{aligned}$$

Since $\mathbb{E}[G]$ is known, it remains to compute the last term in the last display. Consider two independent Poisson processes $\mathcal{P}_1$ and $\mathcal{P}_2$ along the tree $\mathcal{T}$, each at rate $\frac{\rho}{2}$, describing gene loss. As above, we say that a point $s \in \mathcal{T}$ is k-unlost if there is a leaf $i \in \mathcal{T}$ such that the path $[s, i]$ is not hit by an event in $\mathcal{P}_k$. We denote by L_k the length of k-unlost points in $\mathcal{T}$, $k = 1, 2$. Using the same reasoning as in (7.1),

$$\mathbb{E}\biggl[\int\int p_{\mathcal{T}}(s)p_{\mathcal{T}}(t)\,ds\,dt\biggr] = \mathbb{E}[L_1L_2].$$

The latter expectation can be derived via the following construction: in the tree $\mathcal{T}$ with the two independent Poisson loss processes $\mathcal{P}_1$ and $\mathcal{P}_2$, denote by $K_1(\tau)$ the number of lines which are both 1- and 2-unlost some distance τ from the treetop, by $K_2(\tau)$ the number of lines which are 2-lost but 1-unlost and by $K_3(\tau)$ the number of lines which are 1-lost but 2-unlost by time τ. Clearly,

$$L_1 = \int_0^\infty \bigl(K_1(\tau) + K_2(\tau)\bigr)\,d\tau, \qquad L_2 = \int_0^\infty \bigl(K_1(\tau) + K_3(\tau)\bigr)\,d\tau.$$

In addition, $K = (K(\tau))_{\tau\geq 0} = (K_1(\tau), K_2(\tau), K_3(\tau))_{\tau\geq 0}$ is a Markov jump process with the following rates from (k_1, k_2, k_3) to

New state	At rate
$\underline{k}'_1 = (k_1 - 1, k_2, k_3)$	$\lambda_1 = \binom{k_1}{2}$
$\underline{k}'_2 = (k_1, k_2 - 1, k_3)$	$\lambda_2 = \binom{k_2}{2} + k_1k_2 + \frac{\rho}{2}k_2$
$\underline{k}'_3 = (k_1, k_2, k_3 - 1)$	$\lambda_3 = \binom{k_3}{2} + k_1k_3 + \frac{\rho}{2}k_3$
$\underline{k}'_4 = (k_1 + 1, k_2 - 1, k_3 - 1)$	$\lambda_4 = k_2k_3$
$\underline{k}'_5 = (k_1 - 1, k_2 + 1, k_3)$	$\lambda_5 = \frac{\rho}{2}k_1$
$\underline{k}'_6 = (k_1 - 1, k_2, k_3 + 1)$	$\lambda_6 = \frac{\rho}{2}k_1$

Note that $\underline{k}'_1, \dots, \underline{k}'_6, \lambda_1, \dots, \lambda_6$ are as defined in Definition 3.4. In each transition, $2k_1 + k_2 + k_3$ is not increasing and therefore hits 0 after a finite number of transitions.

To obtain (3.11), define the process K with $K(0) = \underline{k} := (k_1, k_2, k_3)$ and define

$$L_1^{\underline{k}} = \int_0^\infty \bigl(K_1(\tau) + K_2(\tau)\bigr)\,d\tau, \qquad L_2^{\underline{k}} = \int_0^\infty \bigl(K_1(\tau) + K_3(\tau)\bigr)\,d\tau$$

such that $L_i \stackrel{d}{=} L_i^{(n,0,0)}$, $i = 1, 2$. We claim that

$$g_{\underline{k}} := \mathbb{E}[L_1^{\underline{k}} L_2^{\underline{k}}]$$

satisfies (3.7) as well as the recursion (3.8).

First, given $k_1 + k_2 = 1$ ($k_1 + k_3 = 1$), there is only one 1-unlost line (2-unlost line) in $\mathcal{T}$. In this case, this line can only be lost by an event in $\mathcal{P}_1$ ($\mathcal{P}_2$), independent of all other coalescence events. Hence, in this case, $L_1^{\underline{k}}$ and $L_2^{\underline{k}}$ are independent, $\mathbb{E}[L_1^{\underline{k}}] = \frac{2}{\rho}$ ($\mathbb{E}[L_2^{\underline{k}}] = \frac{2}{\rho}$) and $\mathbb{E}[L_2^{\underline{k}}] = h_{k_1+k_3}$ ($\mathbb{E}[L_1^{\underline{k}}] = h_{k_1+k_2}$). Combining these results gives (3.7).

Second we show that $\mathbb{E}[L_1^{\underline{k}} L_2^{\underline{k}}]$ satisfies (3.8). Since $\underline{K}$ is a jump process, the first event occurs after an exponential time T with rate $\overline{\lambda}$, which is independent of the new state after the first jump. Conditioning on the first event happening at time T,

$$L_1^{\underline{k}} L_2^{\underline{k}} = \sum_{i=1}^{6} 1_{\{\text{new state is } \underline{k}'_i\}} \bigl((k_1 + k_2)T + L_1^{\underline{k}'_i}\bigr)\bigl((k_1 + k_3)T + L_2^{\underline{k}'_i}\bigr).$$

Taking expectations on both sides shows that $\mathbb{E}[L_1^{\underline{k}} L_2^{\underline{k}}]$ satisfies (3.8). This completes the proof.

7.2. *Proof of Theorem* 5. There are several ways to prove Theorem 5. We present here two approaches, one based on diffusion theory, the other one using an urn model.

PROOF OF THEOREM 5 BASED ON DIFFUSION THEORY. Assume that a gene is present at frequency X_0 at time 0. Then, $(X_t)_{t \geq 0}$ follows the SDE

$$dX = -\frac{\rho}{2} X\,dt + \sqrt{X(1 - X)}\,dW.$$

Frequency spectra for such diffusions have been obtained by Kimura (1964). We follow the arguments given in Durrett (2008), Theorem 7.20. Assume we introduce new genes at frequency $0 < \delta < 1$ into the population at rate

$$\frac{\theta}{2} \frac{1}{\phi(\delta)},$$

where $\mu(x) := -\frac{\rho}{2} x, \sigma^2(x) := x(1 - x)$,

$$\begin{aligned}\psi(y) &:= \exp\Bigl(-2 \int_0^y \frac{\mu(z)}{\sigma^2(z)}\,dz\Bigr) = \exp\Bigl(\rho \int_0^y \frac{1}{1 - z}\,dz\Bigr)\\ &= \exp(-\rho \log(1 - y)) = (1 - y)^{-\rho},\\ \phi(x) &:= \int_0^x \psi(y)\,dy = \frac{1}{1 - \rho}\bigl(1 - (1 - x)^{1-\rho}\bigr).\end{aligned}$$

This rate is consistent in δ: the number of genes at level $\varepsilon > \delta$ is $\frac{\theta}{2}\frac{1}{\phi(\delta)}\frac{\phi(\delta)}{\phi(\varepsilon)}$ since $\frac{\phi(\delta)}{\phi(\varepsilon)}$ is the probability that the gene reaches frequency ε before dying out. Moreover, the Green function for the diffusion—measuring the time until eventual loss of the gene—is given by $2\phi(\delta)m(y)$ for $y > \delta$, where

$$m(y) = \frac{1}{\sigma^2(y)\psi(y)} = \frac{1}{y(1-y)^{1-\rho}}$$

is the density of the speed measure of the diffusion. Hence, we find that the number of genes in frequency x is Poisson with mean

$$g(x)\,dx := \theta\frac{1}{x(1-x)^{1-\rho}}\,dx.$$

Now, the theorem follows since

$$\begin{aligned}\mathbb{E}[G_k] &= \binom{n}{k}\int_0^1 g(x)x^k(1-x)^{n-k}\,dx = \theta\binom{n}{k}\int_0^1 x^{k-1}(1-x)^{n-k-1+\rho}\,dx\\ &= \frac{\theta}{k}\frac{n\cdots(n-k+1)}{(n-1+\rho)\cdots(n-k+\rho)}.\end{aligned}$$ □

PROOF OF THEOREM 5 BASED ON AN URN MODEL. Let $\mathcal{T}$ be the Kingman coalescent and $\Gamma_{\mathcal{T}}$ be the tree-indexed Markov chain from Definition 2.2. First, we focus on loss events for $du \subseteq I$ along the random tree $\mathcal{T}$. (We use the infinitesimal symbol du for notational convenience.) Since du is small, we may safely assume that there is at most one gene in du present in $\bigcup_{i=1}^n \mathcal{G}_i$. Gene loss events in du occur at constant rate $\frac{\rho}{2}$ along each branch. Consider the tree $\mathcal{T}$ from the leaves to the root. Lines coalesce with pair coalescence rate 1, and any line hits a loss event in du at rate $\frac{\rho}{2}$. Upon a loss event we kill the line off the tree. The resulting forest is well known from the family decomposition in the infinite alleles model [e.g., Durrett (2008), page 14]. Using Hoppe's urn, we can also generate the forest forward in time: consider an urn with one colored and one black ball. Choose the colored ball with probability proportional to 1 and the black one with probability proportional to ρ. When choosing a colored ball, put the chosen ball plus one ball of the same color into the urn. When choosing the black ball, put the black ball back together with a ball of a new color. In the next step, again choose any colored ball with probability proportional to 1 and the black balls with probability proportional to ρ. Proceed until there are n colored balls in the urn. Note that, given there are i colored balls in the urn, the chance that the next chosen ball is colored is $\frac{i}{i+\rho} = \frac{\binom{i+1}{2}}{\binom{i+1}{2}+(i+1)\rho/2}$, that is, the chance equals the probability that two among $i+1$ lines coalesce and are not killed off the tree by a gene loss event.

To obtain the correct branch lengths in the tree, when there are i colored balls in the urn, wait an exponential time with rate $\frac{i}{2}(i-1+\rho)$ until adding the next colored ball. This waiting time equals the time the coalescent stays with i lines, when pairs coalesce at rate 1 and single lines are killed at rate $\frac{\rho}{2}$. Hence, by this procedure, balls with the same color belong to the same tree in the forest, and the time the forest spends with i lines is the same as viewing the coalescent backward in time.

So far, Hoppe's urn only described gene loss of the single gene u. Let us add gene gain of a gene in du to the description. During the evolution of Hoppe's urn, which comes with its exponential waiting times, mark all colored balls at rate $\frac{\theta}{2}\,du$. When a marked colored ball is chosen, the added ball is again marked. Here, a mark stands for the presence of the considered gene along the corresponding ancestral line. Since du is small, there is at most one mark along the forest.

For the forest given by the marked Hoppe's urn, we distinguish the times $T_1, \dots, T_n$ when there are $1, \dots, n$ lines present. We say that line l during T_i is of size k iff the ball belonging to this line produces exactly $k-1$ offspring until the urn finishes. Hence,

$$\begin{aligned}\mathbb{E}[G_k] &= \int_I \mathbb{E}[du \in \mathcal{G}_i \text{ for exactly } k \text{ different } i] \\ &= \sum_{i=1}^n \sum_{l=1}^i \mathbb{P}[l\text{th line during } T_i \text{ is of size } k] \\ &\qquad\qquad \times \int \mathbb{P}[\text{mark in } du \text{ on } l\text{th line during } T_i]\end{aligned}$$

and

$$\mathbb{P}[\text{mark in } du \text{ on } l\text{th line during } T_i] = \frac{\theta/2\,du}{i/2(i-1+\rho)} = \frac{\theta}{i(i-1+\rho)}\,du.$$

Let us turn to the probability that the lth line during T_i is of size k. The reasoning below is well known from Pòlya urn models. When starting with $i-1$ unmarked and one marked lines, there are $\binom{n-i}{k-1}$ possibilities at what times $k-1$ marked balls are added when $n-i$ balls are added to the urn in total. For any of these possibilities, the probability is

$$\frac{(k-1)!(i-1+\rho)\cdots(n-k-1+\rho)}{(i+\rho)\cdots(n-1+\rho)}.$$

Putting everything together,

$$\begin{aligned}\mathbb{E}[G_k] &= \sum_{i=1}^n i \binom{n-i}{k-1} \frac{(k-1)!(i-1+\rho)\cdots(n-k-1+\rho)}{(i+\rho)\cdots(n-1+\rho)} \frac{\theta}{i(i-1+\rho)} \\ &= \frac{\theta}{k} \frac{k!}{(n-k+\rho)\cdots(n-1+\rho)} \underbrace{\sum_{i=1}^n \binom{n-i}{k-1}}_{=\binom{n}{k}} \\ &= \frac{\theta}{k} \frac{n\cdots(n-k+1)}{(n-k+\rho)\cdots(n-1+\rho)},\end{aligned}$$

and we are done. □

Acknowledgments. We thank Daniel Huson for pointing out the reference [Huson and Steel (2004)] and we are grateful to Cornelia Borck, Andrej Depperschmidt and Bernhard Haubold for comments on our manuscript.

REFERENCES

BENTLEY, S. (2009). Sequencing the species pan-genome. *Nature Rev. Microbiol.* **7** 258–259.

DUFRESNE, A., OSTROWSKI, M., SCANLAN, D. J., GARCZAREK, L., MAZARD, S., PALENIK, B. P., PAULSEN, I. T., DE MARSAC, N. T., WINCKER, P., DOSSAT, C., FERRIERA, S., JOHNSON, J., POST, A. F., HESS, W. R. and PARTENSKY, F. (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol.* **9** R90.

DURRETT, R. (2008). *Probability Models for DNA Sequence Evolution*, 2nd ed. Springer, New York. MR2439767

DURRETT, R. and POPOVIC, L. (2009). Degenerate diffusions arising from gene duplication models. *Ann. Appl. Probab.* **19** 15–48. MR2498670

DYKHUIZEN, D. E. and GREEN, L. (1991). Recombination in *Escherichia coli* and the definition of biological species. *J. Bacteriol.* **173** 7257–7268.

EHRLICH, G. D., HU, F. Z., SHEN, K., STOODLEY, P. and POST, J. C. (2005). Bacterial plurality as a general mechanism driving persistence in chronic infections. *Clin. Orthop. Relat. Res.* **437** 20–24.

EVANS, S., SHVETS, S. and SLATKIN, M. (2007). Non-equlibrium theory of the allele frequency spectrum. *Theo. Pop. Biol.* **71** 109–119.

EWENS, W. J. (2004). *Mathematical Population Genetics. I. Theoretical Introduction*, 2nd ed. *Interdisciplinary Applied Mathematics* **27**. Springer, New York. MR2026891

FRASER, C., HANAGE, W. P. and SPRATT, B. G. (2007). Recombination and the nature of bacterial speciation. *Science* **315** 476–480.

FU, Y. X. (1995). Statistical properties of segregating sites. *Theo. Pop. Biol.* **48** 172–197.

GRIFFITHS, R. C. (2003). The frequency spectrum of a mutation and its age, in a general diffusion model. *Theo. Pop. Biol.* **64** 241–251.

HILLER, N. L., JANTO, B., HOGG, J. S., BOISSY, R., YU, S., POWELL, E., KEEFE, R., EHRLICH, N. E., SHEN, K., HAYES, J., BARBADORA, K., KLIMKE, W., DERNOVOY, D., TATUSOVA, T., PARKHILL, J., BENTLEY, S. D., POST, J. C., EHRLICH, G. D. and HU, F. Z. (2007). Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: Insights into the pneumococcal supragenome. *J. Bacteriol.* **189** 8186–8195.

HOGG, J. S., HU, F. Z., JANTO, B., BOISSY, R., HAYES, J., KEEFE, R., POST, J. C. and EHRLICH, G. D. (2007). Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* **8** R103.

HUSON, D. H. and STEEL, M. (2004). Phylogenetic trees based on gene content. *Bioinformatics* **20** 2044–2049.

KETTLER, G. C., MARTINY, A. C., HUANG, K., ZUCKER, J., COLEMAN, M. L., RODRIGUE, S., CHEN, F., LAPIDUS, A., FERRIERA, S., JOHNSON, J., STEGLICH, C., CHURCH, G. M., RICHARDSON, P. and CHISHOLM, S. W. (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet.* **3** e231.

KIMURA, M. (1964). Diffusion models in population genetics. *J. Appl. Probab.* **1** 177–232. MR0172727

KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR671034

KUNIN, V. and OUZOUNIS, C. A. (2003). GeneTRACE-reconstruction of gene content of ancestral species. *Bioinformatics* **19** 1412–1416.

LAPIERRE, P. and GOGARTEN, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends in Genetics* **25** 107–110.

LEFÉBURE, T. and STANHOPE, M. J. (2007). Evolution of the core and pan-genome of *Streptococcus*: Positive selection, recombination, and genome composition. *Genome Biol.* **8** R71.

MAIDEN, M. C., BYGRAVES, J. A., FEIL, E., MORELLI, G., RUSSELL, J. E., URWIN, R., ZHANG, Q., ZHOU, J., ZURTH, K., CAUGANT, D. A., FEAVERS, I. M., ACHTMAN, M. and SPRATT, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA* **95** 3140–3145.

MAYNARD-SMITH, J. (1995). Do bacteria have population genetics? In *Population Genetics of Bacteria* 1–12. Cambridge Univ. Press, Cambridge.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. and RAPPUOLI, R. (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15** 589–594.

MÖHLE, M. and SAGITOV, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* **29** 1547–1562. MR1880231

PERNA, N. T., PLUNKETT, G., BURLAND, V., MAU, B., GLASNER, J. D., ROSE, D. J., MAYHEW, G. F., EVANS, P. S., GREGOR, J., KIRKPATRICK, H. A., PÉSFAI, G., HACKETT, J., KLINK, S., BOUTIN, A., SHAO, Y., MILLER, L., GROTBECK, E. J., DAVIS, N. W., LIM, A., DIMALANTA, E. T., POTAMOUSIS, K. D., APODACA, J., ANANTHARAMAN, T. S., LIN, J., YEN, G., SCHWARTZ, D. C., WELCH, R. A. and BLATTNER, F. R. (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409** 529–533.

RILEY, M. A. and LIZOTTE-WANIEWSKI, M. (2009). Population genomics and the bacterial species concept. *Methods Mol. Biol.* **532** 367–377.

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L., DURKIN, A. S., DEBOY, R. T., DAVIDSEN, T. M., MORA, M., SCARSELLI, M., MARGARIT Y ROS, I., PETERSON, J. D., HAUSER, C. R., SUNDARAM, J. P., NELSON, W. C., MADUPU, R., BRINKAC, L. M., DODSON, R. J., ROSOVITZ, M. J., SULLIVAN, S. A., DAUGHERTY, S. C., HAFT, D. H., SELENGUT, J., GWINN, M. L., ZHOU, L., ZAFAR, N., KHOURI, H., RADUNE, D., DIMITROV, G., WATKINS, K., O'CONNOR, K. J., SMITH, S., UTTERBACK, T. R., WHITE, O., RUBENS, C. E., GRANDI, G., MADOFF, L. C., KASPER, D. L., TELFORD, J. L., WESSELS, M. R., RAPPUOLI, R. and FRASER, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci. USA* **102** 13950–13955.

TETTELIN, H., RILEY, D., CATTUTO, C. and MEDINI, D. (2008). Comparative genomics: The bacterial pan-genome. *Curr. Opin. Microbiol.* **11** 472–477.

VULIC, M., DIONISIO, F., TADDEI, F. and RADMAN, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl. Acad. Sci. USA* **94** 9763–9767.

WAKELEY, J. (2008). *Coalescent Theory*: *An Introduction*. Roberts and Company, Colorado.

WRIGHT, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. USA* **24** 253–259.

F. BAUMDICKER
P. PFAFFELHUBER
FAKULTÄT FÜR MATHEMATIK UND PHYSIK
ALBERT-LUDWIGS UNIVERSITY
ECKERSTRASSE 1
D-79104 FREIBURG
GERMANY
E-MAIL: p.p@stochastik.uni-freiburg.de

W. R. HESS
FAKULTÄT FÜR BIOLOGIE
ALBERT-LUDWIGS UNIVERSITY
SCHÄNZLESTR. 1
D-79104 FREIBURG
GERMANY