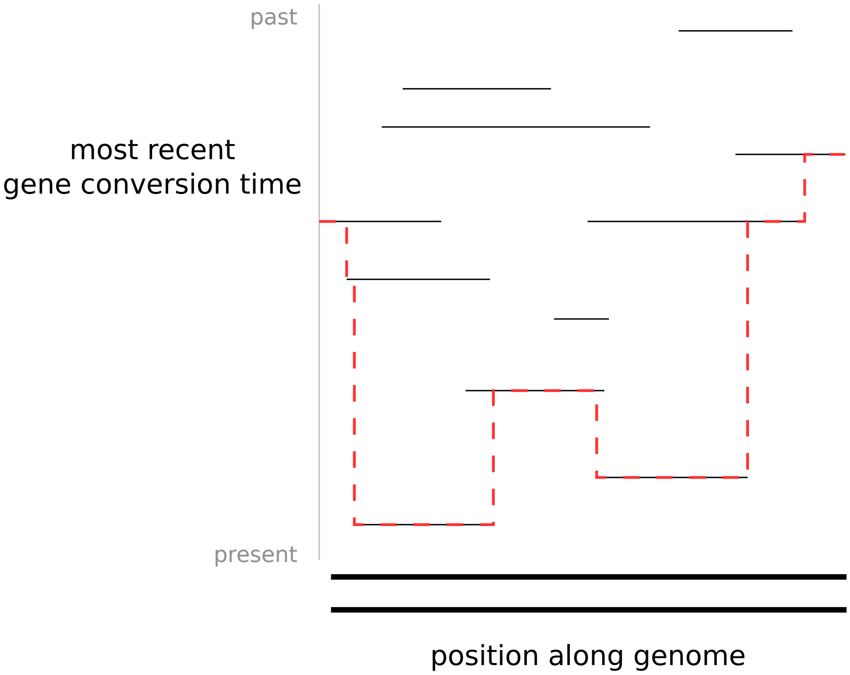

Figure 1: Schematic of the gene conversion ancestral process. Gene conversion events occur as a Poisson process in time and along the genome, and each gene conversion tract has an exponential length distribution. At each point along the two aligned chromosomes, the time of the most recent gene conversion event is marked with a dashed red line. We are interested in modeling the distribution of this dashed red line across a genome, since it determines patterns of genetic variation.

1 Gene conversion

We are interested in modeling the effects of gene conversion on genome-wide patterns of genetic variation within clonal asexual lineages. We will begin with a diploid asexual lineage and then proceed to the more complex case of triploids. We assume that the asexual lineage is derived from diploid sexual ancestors.

We work in continuous time, with time scaled by $2N_0$, where N_0 is the (diploid) population size of the sexual ancestor of the asexual lineage. Assume that this diploid asexual lineage is derived from its sexual ancestors at time $t = T_d$ in the past, with the present being $t = 0$. For the moment we will also model distances along the genome as continuous variables.

We assume that during the asexual phase of the lineage's ancestry, gene conversion initiation events occur as a Poisson process in both time and space (i.e., along the genome). For the moment, we also assume that each gene conversion tract is exponentially distributed in length, although we will return to this assumption later.

In sexually reproducing species, gene conversion can be thought of as a double-recombination event in the coalescent process. In clonally reproducing asexual lineages, gene conversion instead acts analogously to coalescence, since chromosomes are "trapped" in the same lineage and each gene conversion event causes two chromosomes to share ancestry at that point. Since all genetic differences separating two chromosomes are due to mutations that occurred since their shared ancestor, we are intersted in the most recent gene conversion time at each locus: the distribution across the genome of these most-recent gene conversion times is our object of interest. See Fig. 1 for an illustration of the process under consideration.

With our assumption of a two-dimensional Poisson process, the probability that a gene conversion event is initiated in the intervals $(x, x + dx)$ along the genome and $(t, t + dt)$ back in time is $2\lambda_c dxdt$ for some rate of gene conversion λ_c. The factor of two is due to the fact that a gene conversion event can occur on either chromosome. The length of a particular gene conversion tract has the density $\lambda_L e^{-\lambda_L x} dx$ so that the mean gene conversion tract length is $1/\lambda_L$.

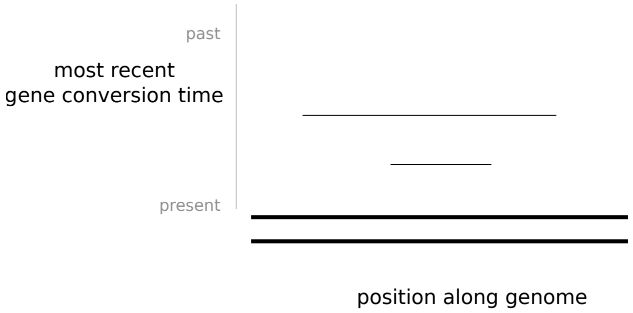

Figure 2: Example of full overlap in gene conversion tracts. Because of events like this, the gene-conversion process cannot be described by a first-order Markov model. In our model we assume that events such as this cannot occur, enabling a description by a first-order Markov process.

Going back in time at a particular locus, the probability that that particular locus is affected by a gene conversion event in the time interval $(t, t + dt)$ is

$$\begin{aligned} &= 2\lambda_c dt \int_0^\infty dx\, \mathrm{P}(L > x) \\ &= 2\lambda_c dt\, \mathrm{E}[L] \\ &= 2\frac{\lambda_c}{\lambda_L} dt, \end{aligned}$$

where L is the random length of a gene conversion tract. The factor of 2 is due to the fact that a gene conversion event can occur on either ancestral chromosome. In the second equality we use the fact that $\int_0^\infty \mathrm{P}(L > x)dx = \mathrm{E}[L]$ for any non-negative random variable L. (Thus, this holds for all gene conversion tract length distributions.) In other words, the rate of encountering a gene conversion tract that causes shared ancestry between the two chromosomes at a particular location in the genome is the rate of initiation of these events in time multiplied by the mean length of a tract.

Under the assumption of an exponential tract length distribution, the end of the current tract is encountered with probability $\lambda_L dx$. It is also possible that we encounter a more recent gene conversion event that causes a more recent common ancestor event between the two chromosomes. Given the current gene conversion time s, this happens at rate $2\lambda_c s dx$. Thus the total rate of encountering the a change in gene conversion time is $(\lambda_L + 2\lambda_c s)dx$.

Given that a more recent gene conversion time is encountered, the age of this gene conversion time is uniformly distributed in $[0, s)$. Given that the end of the tract is encountered, the next gene conversion event is encountered (going back in time from s) at rate $2\lambda_c/\lambda_L dt$, as before. **Crucially, this assumes that gene conversion events do not fully overlap.** In reality, gene conversion events will fully overlap, and the full process cannot be described by a Markov jump process. We will make the assumption that gene conversion tracts do not overlap completely (example shown in Fig. 2), and thus our model can be described by a first-order Markov chain. It is possible to make the model a second-order Markov chain and allow *pairwise* overlap, similar to the approach in Yin et al. 2009 (*Bioinformatics*, 25:231-239), but this nearly squares the number of states to consider, so the viability of this will have to be considered later.

Considering the above, the local gene conversion time changes along the genome at rate $(\lambda_L + 2s\lambda_c)dx$, and at a transition site, the density of next gene conversion time is

$$r(t|s)dt = \begin{cases} \frac{2\lambda_c}{2s\lambda_c + \lambda_L} dt & 0 < t < s \\ \frac{2\lambda_c}{2s\lambda_c + \lambda_L} e^{-\frac{2\lambda_c (t-s)}{\lambda_L}} dt & t > s. \end{cases} \tag{1}$$

1.1 Gene conversion in triploids

To extend this picture to triploids, we assume that each genome consists of three homologous chromosomes and that gene conversion occurs between each potential pair of source and target chromosomes with equal probability. To describe this gene-conversion ancestral process, two gene conversion times are required. Again using λ_c as the rate of gene conversion initiation sites across the genome, and λ_L as the rate of termination of a given tract, joint distribution of "coalescence" times is

$$\begin{aligned} f_{S_3,S_2}(s_3,s_2)ds_3ds_2 &= \frac{6\lambda_c}{\lambda_L}e^{-\frac{6\lambda_c s_3}{\lambda_L}}\frac{2\lambda_c}{\lambda_L}e^{-\frac{2\lambda_c(s_2-s_3)}{\lambda_L}}ds_3ds_2 \\ &= \frac{12\lambda_c^2}{\lambda_L^2}e^{-\frac{2\lambda_c(s_2+2s_3)}{\lambda_L}}ds_3ds_2. \end{aligned} \tag{2}$$

Given a current configuration (s_3, s_2), the rate of leaving that state is $2\lambda_L + 6\lambda_c s_3 + 2\lambda_c(s_2 - s_3)$. The transition kernel at a point of change is

$$q(t_3,t_2|s_3,s_2) = \begin{cases} \frac{6\lambda_c s_3}{\lambda_T}\frac{1}{s_3}\delta(t_2-s_2), & 0 < t_3 < s_3, t_2 = s_2 \\ \frac{2\lambda_c(s_2-s_3)}{\lambda_T}\frac{1}{s_2-s_3}\delta(t_3-s_3), & t_3 = s_3, s_3 < t_2 < s_2 \\ \frac{\lambda_L}{\lambda_T}\frac{2\lambda_c}{\lambda_L}e^{-\frac{2\lambda_c(t_2-s_2)}{\lambda_L}}\delta(t_3-s_3), & t_3 = s_3, t_2 > s_2 \\ \frac{\lambda_L}{\lambda_T}\frac{6\lambda_c}{\lambda_L}e^{-\frac{6\lambda_c(t_3-s_3)}{\lambda_L}}\delta(t_2-s_2), & t_3 > s_3, t_2 = s_2 \\ \frac{\lambda_L}{\lambda_T}e^{-\frac{6\lambda_c(s_2-s_3)}{\lambda_L}}\frac{2\lambda_c}{\lambda_L}e^{-\frac{2\lambda_c(t_2-s_2)}{\lambda_L}}, & t_3 = s_2, t_2 > s_2. \end{cases} \tag{3}$$