

Calculating likelihoods for coalescence of linear genomes

Nick Barton & Konrad Lohse



A wealth of data, and a simple (neutral) model ...

A genealogy is described by its branch lengths: $\underline{t} = \{t_S\} \quad S \subset \Omega$

A genealogy is described by its branch lengths: $\underline{t} = \{t_S\} \quad S \subset \Omega$

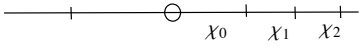
For discrete loci, tracing back to the previous coalescence or recombination:

To apply to a large # of non-recombining blocks, *tabulate* probabilities of possible mutational configurations

Hearn et al., 2014, *Molecular Ecology*: gallwasps (*Biorhiza pallida*)

- >2000 blocks of > 2kb from two replicate triplets

Lohse and Frantz, 2014, arXiv:1307.8263v1: Admixture from *Neandertal*

Linear genomes: 

- recombination in $\chi_0 + \chi_1 + \chi_2$ can occur on $\underline{t}_0, \underline{t}_1, \underline{t}_2$, length L_2^*
- recombination in non-ancestral material matters
- choose the # of genomes and the # of blocks, and follow $\text{Prob}[\underline{t}, \underline{\chi}]$
- blocks may have identical genealogies
- a random point will tend to fall on a long block

Two blocks: $\text{Prob}[t, \mathcal{X}] = P[t] \underline{L}_0 e^{-L_0 \chi_0} \underline{L}_1 e^{-L_1 \chi_1}$
 $\psi = \mathbb{E}[L_0 e^{-L_0 \chi_0} e^{-L_1 \chi_1} e^{-\omega_0 t_0 - \omega_1 t_1}]$

Two blocks: $\text{Prob}[\underline{t}, \underline{\chi}] = P[\underline{t}] \underline{L}_0 e^{-L_0 \chi_0} \underline{L}_1 e^{-L_1 \chi_1}$

$\psi = \mathbb{E}[L_0 e^{-L_0 \chi_0} e^{-L_1 \chi_1} e^{-\omega_0 t_0 - \omega_1 t_1}]$

$$\psi_1[\underline{\Omega}] = \frac{1}{(\lambda_k + k\chi_0 + k\chi_1 + \omega_L)} \left(\sum_{\substack{1 \leq i < j \leq k \\ k \neq 2}} \psi_1[\underline{\Omega}_{i,j}] + \sum_{\alpha=1}^k \psi_0[\underline{\Omega}_\alpha] \right) \quad (1)$$

$$\psi_0[\underline{\Omega}] = \frac{1}{(\lambda_k + k\chi_0 + k\chi_1 + \omega_L)} \left(\sum_{1 \leq i < j \leq k} \psi_0[\underline{\Omega}_{i,j}] \right)$$

Two blocks, two individuals:

$$\omega_0 = \omega_{a_0} + \omega_{b_0}, \omega_1 = \omega_{a_1} + \omega_{b_1}, \omega_L = \omega_0 + \omega_1$$

$$\psi_1[\{\mathbf{a}_0, \mathbf{a}_1\}, \{\mathbf{b}_0, \mathbf{b}_1\}] = \frac{(\psi_0[\{\mathbf{a}_0, \mathbf{a}_1\}, \{\mathbf{b}_0\}, \{\mathbf{b}_1\}] + \psi_0[\{\mathbf{a}_0\}, \{\mathbf{a}_1\}, \{\mathbf{b}_0, \mathbf{b}_1\}])}{(1 + \omega_L + 2\chi_0 + 2\chi_1)}$$

$$\psi_0[\{\mathbf{a}_0\}, \{\mathbf{a}_1\}, \{\mathbf{b}_0, \mathbf{b}_1\}] = \psi_0[\{\mathbf{a}_0, \mathbf{a}_1\}, \{\mathbf{b}_0\}, \{\mathbf{b}_1\}] = \frac{(\psi_0[\{\mathbf{a}_0, \mathbf{a}_1\}, \{\mathbf{b}_0, \mathbf{b}_1\}] + \psi_0[\{\mathbf{a}_0\}, \{\mathbf{b}_0\}] + \psi_0[\{\mathbf{a}_1\}, \{\mathbf{b}_1\}])}{(3 + \omega_L + 2\chi_0 + 2\chi_1)}$$

$$\psi_0[\{\mathbf{a}_0\}, \{\mathbf{b}_0\}] = \frac{1}{1 + 2\chi_0 + \omega_0} \quad \psi_0[\{\mathbf{a}_1\}, \{\mathbf{b}_1\}] = \frac{1}{1 + 2\chi_1 + \omega_1}$$

$$\psi_0[\{\mathbf{a}_0, \mathbf{a}_1\}, \{\mathbf{b}_0, \mathbf{b}_1\}] = \frac{1}{1 + 2\chi_0 + 2\chi_1 + \omega_L}$$

The marginal distribution of block lengths is:

$$\mathbf{P}[\chi_0] = \mathbb{E}[\mathbf{L}_0 e^{-\mathbf{L}_0 \chi_0}] = \psi_1 |_{\underline{\omega}=0, \chi_1=0} = \frac{2}{(1 + 2 \chi_0)^2} \quad \mathbb{E}[\chi_0] = \infty$$

$$\mathbf{P}[\chi_1] = \mathbb{E}[\mathbf{L}_1 e^{-\mathbf{L}_1 \chi_1}] = \int_0^\infty -\frac{\partial \psi_1}{\partial \chi_1} |_{\underline{\omega}=0, \chi_1=0} d\chi_0 \quad \mathbb{E}[\chi_1] = 0.881$$

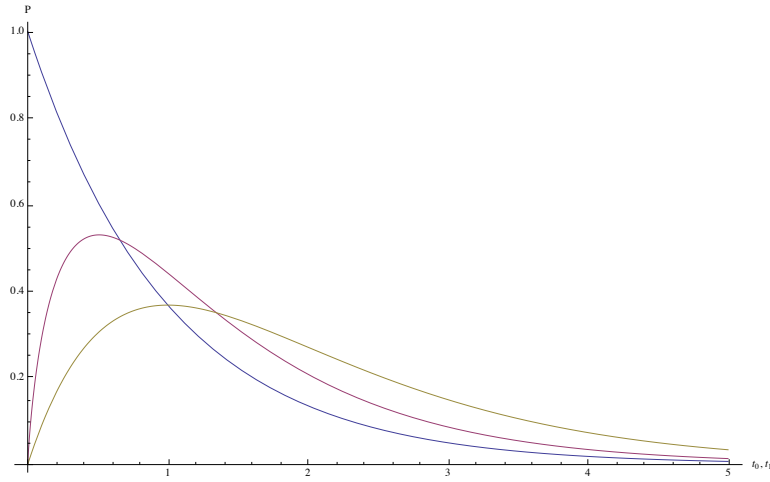
$$= \frac{1}{4 \chi_1} \left(-\frac{2}{(1 + 2 \chi_1)^2 (1 + \chi_1)} + \frac{(1 + 4 \chi_1)}{(1 + 2 \chi_1)^2 (1 + 2 \chi_1)^2 \chi_1^2} \text{Log}\left[\frac{1 + 2 \chi_1}{3 + 2 \chi_1}\right] + \frac{(1 + 2 \chi_1)}{\chi_1 (1 + \chi_1)^2} \text{Log}[3 + 2 \chi_1] \right)$$

Distribution of coalescence times along successive blocks

Distribution at a random point in the genome is e^{-t_0} (left), which has mean 1.

The adjacent genealogy has a higher mean, $1 + \log(27)/8 \sim 1.41$ (middle).

The right curve shows the distribution of sizes of a randomly chosen *block*, $t_{\infty} e^{-t_{\infty}}$, which has mean 2.



The general recursion

$$\psi_{\underline{\theta}}[\underline{\Omega}] = \mathbb{E} \left[\left(\prod_{\alpha=0}^{n-3} \mathbf{L}_{\alpha}^* \right) \mathbf{L}_{n-2} e^{-\underline{\mathbf{L}}^{\circ} \cdot \underline{\lambda}} e^{-\underline{\omega} \cdot \underline{\mathbf{t}}} \right]$$

where the vector $\underline{\theta}$ follows whether or not a recombination is pending.

The general recursion

$$\psi_{\underline{\theta}}[\underline{\Omega}] = \mathbb{E} \left[\left(\prod_{\alpha=0}^{n-3} \mathbf{L}_{\alpha}^* \right) \mathbf{L}_{n-2} e^{-\underline{\mathbf{L}} \cdot \underline{\boldsymbol{\lambda}}} e^{-\underline{\omega} \cdot \underline{\mathbf{t}}} \right]$$

where the vector $\underline{\theta}$ follows whether or not a recombination is pending.

$$\psi_{\underline{\theta}}[\underline{\Omega}] = \frac{1}{(\lambda_k + \sum_{\alpha=0}^{n-1} k_{\alpha}^* \chi_{\alpha} + \omega_L)} \left(\sum_{i < j} \psi_{\underline{\theta}}[\underline{\Omega}_{i,j}] + \sum_{\alpha \in \theta} \psi_{\theta^*}[\underline{\Omega}_{\alpha}] \right)$$

Automating the calculations: three genomes, two blocks

```

Ω2 = {{{{a}}, {{x}}}, {{{b}}, {{y}}}, {{{c}}, {{z}}}};
yy2 = FixedPoint[
  # /. P[s__] := (mb = MakeBlockEqns[P[s]];  $\frac{\text{Total}[mb[[1, 1]] + \text{Total}[mb[[1, 2]]]}{mb[[2]]}$ ) &,
  P[Ω2, ω, {0, 0}, {True}]]

```

Automating the calculations: three genomes, two blocks

```

Ω2 = {{{{a}}, {{{x}}}, {{{b}}, {{{y}}}, {{{c}}, {{{z}}}};
yy2 = FixedPoint [
  # /. P[s__] := (mb = MakeBlockEqns[P[s]];  $\frac{\mathbf{Total}[\mathbf{mb}[[1, 1]]] + \mathbf{Total}[\mathbf{mb}[[1, 2]]]}{\mathbf{mb}[[2]]}$ ) &,
  P[Ω2, ω, {0, 0}, {True}] ]

```

This just follows two dummy variables, corresponding to L_0 , L_1 :

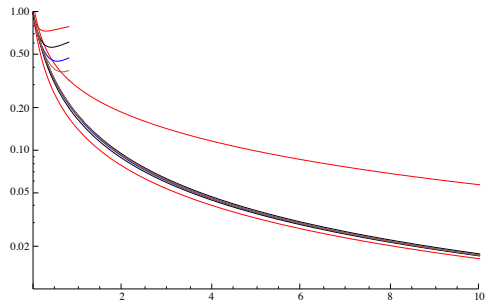
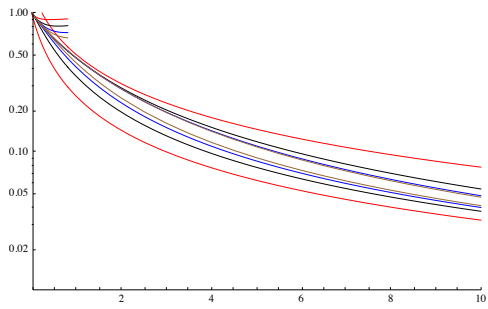
```
y1 = yy2 /. lengthRule[genealogySets[Ω2], ω] // Simplify; y1 /. ω[i_] := ωi
```

$$\frac{6 \left(\frac{1}{1+2\omega_0} + \frac{1}{1+2\omega_1} + \frac{1}{1+2\omega_0+2\omega_1} \right)}{(1+2\omega_0+2\omega_1)(3+2\omega_0+2\omega_1)} + \frac{\frac{1}{(1+\omega_0+\omega_1)(1+2\omega_0+2\omega_1)} + \frac{\frac{1}{1+2\omega_0} + \frac{1}{1+2\omega_1} + \frac{1}{1+2\omega_0+2\omega_1}}{3+2\omega_0+2\omega_1} + \frac{2 \left(\frac{1}{1+2\omega_0} + \frac{2}{1+2\omega_1+2\omega_1} \right)}{3+3\omega_0+2\omega_1} + \frac{2 \left(\frac{1}{1+2\omega_1} + \frac{2}{1+2\omega_0+2\omega_1} \right)}{3+2\omega_0+3\omega_1}}{2+\omega_0+\omega_1}}{3(1+\omega_0+\omega_1)}$$

Luckily, the total probability is 1:

$$\int_0^\infty (\mathbf{y1} /. \omega[1] \rightarrow 0) d\omega \mid [0]$$

Probability that there are no SNP in an interval x . $\gamma = \frac{\mu}{r} = 0.5, 2$ (top, bottom).



Topologies: two recombinations, three genomes

The total probability of the five distinct combinations of topology, and mean lengths of successive blocks.

	overall	$\{(a, b), \{p, q\}, \{x, y\}\}$	$\{(a, b), \{p, q\}, \{x, z\}\}$	$\{(a, b), \{q, r\}, \{x, y\}\}$	$\{(a, b), \{q, r\}, \{x, z\}\}$	$\{(a, b), \{q, r\}, \{y, z\}\}$
#	27	3	6	6	6	6
Prob.	1	0.707	0.108	0.0256	0.0251	0.135
$E[\chi_0]$	0.693	0.656	0.576	0.965	0.738	0.738
$E[\chi_1]$	0.486	0.447	0.595	0.530	0.675	0.680
$E[\chi_2]$	0.425	0.408	0.515	0.413	0.514	0.509

What to do with the machinery...

- effects of low rates of recombination on inference
- understanding the sequential Markov coalescent