Supplementary Information

It is easiest to view this document in *Mathematica* or MathPlayer (available as a free download at http://www.wolfram.com/producs/-player/).

1. Automation for the IM model: Three genes in two demes

1.1 Set up

Notation

Lineages are labelled by the set of genes to which they are ancestral. Thus, lineages at the tips are ancestral to a single gene, and are labelled $\{a\}$, $\{b\}$, …. A deme containing lineages $\{b\}$ and $\{c\}$ is denoted $\{\{b\},\{c\}\}$, and two demes - one containing lineage $\{a\}$ and the other containing $\{b\}$ and $\{c\}$ - is denoted $\{\{\{a\}\},\{\{b\},\{c\}\}\}$. If populations can split, we also need to define the ancestry of the demes in a similar way. $\{\{x\},\{y\}\}$ denotes two demes, ancestral to the present-day demes x and y. The single ancestral deme that existed before the split is denoted $\{\{x, y\}\}$. Note that a single lineage must be ancestral to every gene, and a single deme must be ancestral to every present-day deme. Thus, the content of the lists that define the genealogy and the population phylogeny stays the same - only the nesting changes.

The generating function has the form GF[ω, $\{\{\{a\}\},\{\{b\},\{c\}\}\}$, M, $\{\{x\},\{y\}\}$, Λ]. $\omega[\{a\}]$ corresponds to branch $\{a\}$, which is ancestral to a; $\Lambda[\{x, y\}]$ is the split rate of population $\{x, y\}$. $M = 4\,N\,m$ is the scaled migration rate

In the text, this is denoted more compactly as $\psi[a, b \setminus c]$. tidyNotation[ψ] gives something like this notation, to make the output more readable.

Solving the recursions

This procedure is simple, but not very efficient given that it does not exploit all the symmetries, which can drastically reduce the number of equations needed. However, this part is extremely fast relative to later steps.

makeAllEqns automates the recursions for the IM model. Here we assume a sampling configuartion {a/b,c}.

```
eqs = makeAllEqns[GF[ω, {{{a}}, {{b}, {c}}}, M, {{x}, {y}}, Λ]]; vars = GetVars[eqs]
```

```
{GF[ω, {{{a}, {b, c}}}, M, {{x, y}}, Λ], GF[ω, {{{b}, {a, c}}}, M, {{x, y}}, Λ],
 GF[ω, {{{c}, {a, b}}}, M, {{x, y}}, Λ], GF[ω, {{{a}, {b}, {c}}}, M, {{x, y}}, Λ],
 GF[ω, {{}, {{a}, {b, c}}}, M, {{x}, {y}}, Λ], GF[ω, {{}, {{b}, {a, c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{}, {{c}, {a, b}}}, M, {{x}, {y}}, Λ], GF[ω, {{}, {{a}, {b}, {c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{a}}, {{b, c}}}, M, {{x}, {y}}, Λ], GF[ω, {{{a}}, {{b}, {c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{b}}, {{a, c}}}, M, {{x}, {y}}, Λ], GF[ω, {{{b}}, {{a}, {c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{c}}, {{a, b}}}, M, {{x}, {y}}, Λ], GF[ω, {{{c}}, {{a}, {b}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{a, b}}, {{c}}}, M, {{x}, {y}}, Λ], GF[ω, {{{a, c}}, {{b}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{b, c}}, {{a}}}, M, {{x}, {y}}, Λ], GF[ω, {{{a}, {b}}, {{c}}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{a}, {c}}, {{b}}}, M, {{x}, {y}}, Λ], GF[ω, {{{a}, {b, c}}, {}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{b}, {c}}, {{a}}}, M, {{x}, {y}}, Λ], GF[ω, {{{b}, {a, c}}, {}}, M, {{x}, {y}}, Λ],
 GF[ω, {{{c}, {a, b}}, {}}, M, {{x}, {y}}, Λ], GF[ω, {{{a}, {b}, {c}}, {}}, M, {{x}, {y}}, Λ]}
```

Next, we choose those equations that involve 1 deme, and solve them. First/@eqs1 lists the GF[] that we need to solve for:

```
eqs1 = selectEqns[eqs, {1, All}];
soln1 = Solve[eqs1, First /@ eqs1][[1]]
```

$$\left\{GF[\omega, \{\{\{a\}, \{b, c\}\}\}, M, \{\{x, y\}\}, \Lambda] \to -\frac{1}{-1 - \omega[\{a\}] - \omega[\{b, c\}]},\right.$$
$$GF[\omega, \{\{\{b\}, \{a, c\}\}\}, M, \{\{x, y\}\}, \Lambda] \to -\frac{1}{-1 - \omega[\{b\}] - \omega[\{a, c\}]},$$
$$GF[\omega, \{\{\{c\}, \{a, b\}\}\}, M, \{\{x, y\}\}, \Lambda] \to -\frac{1}{-1 - \omega[\{c\}] - \omega[\{a, b\}]},$$
$$GF[\omega, \{\{\{a\}, \{b\}, \{c\}\}\}, M, \{\{x, y\}\}, \Lambda] \to$$
$$1/((3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])(1+\omega[\{c\}]+\omega[\{a, b\}])) -$$
$$1/((3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])(-1-\omega[\{b\}]-\omega[\{a, c\}])) -$$
$$\left.1/((3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])(-1-\omega[\{a\}]-\omega[\{b, c\}]))\right\}$$

We then choose those that involve 2 genes in 2 demes:

```
eqs2 = selectEqns[eqs, {2, 2}];
soln2 = Solve[eqs2, First /@ eqs2][[1]];
```

This needs to be simplified, by using the solutions for all the 1-deme cases (stored in soln1). This is the solution for all configurations with two genes in two demes. Note that this is inefficient: there are 12 configurations in general, but only three kinds for the symmetric model (where both demes have equal popultaion size and migration is symmetric) - the genes can be in the same deme or different demes.

These are the solutions for two genes with two demes, given in the "tidy notation". $\psi_{\{\},\{\{a\},\{b,c\}\}}$ denotes an empty deme, and a deme containing two lineages - one ancestral to $\{a\}$, the other to $\{b, c\}$.

```
soln2Simp = soln2 /. soln1 // Simplify;
soln2Simp /. tidyNotation[ψ] /. {ω_{x_,y_}} :> ωL - ωComplement[{a,b,c},{x,y}], Λ_{x},{y} → Λ} // Simplify
```

$$\left\{\psi_{\{\},\{\{a\},\{b,c\}\}} \to \left(M+\Lambda+2M\Lambda+\Lambda^2+(1+M+2\Lambda)\omega_L+\omega_L^2\right)/\right.$$
$$\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right), \psi_{\{\},\{\{b\},\{a,c\}\}} \to$$
$$\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+M+2\Lambda)\omega_L+\omega_L^2\right)/\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right),$$
$$\psi_{\{\},\{\{c\},\{a,b\}\}} \to \left(M+\Lambda+2M\Lambda+\Lambda^2+(1+M+2\Lambda)\omega_L+\omega_L^2\right)/$$
$$\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right), \psi_{\{\{a\}\},\{\{b,c\}\}} \to$$
$$\left(M+\Lambda+2M\Lambda+\Lambda^2+(M+\Lambda)\omega_L\right)/\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right),$$
$$\psi_{\{\{b\}\},\{\{a,c\}\}} \to \left(M+\Lambda+2M\Lambda+\Lambda^2+(M+\Lambda)\omega_L\right)/$$
$$\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right), \psi_{\{\{c\}\},\{\{a,b\}\}} \to$$
$$\left(M+\Lambda+2M\Lambda+\Lambda^2+(M+\Lambda)\omega_L\right)/\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right),$$
$$\psi_{\{\{a,b\}\},\{\{c\}\}} \to \left(M+\Lambda+2M\Lambda+\Lambda^2+(M+\Lambda)\omega_L\right)/$$
$$\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right), \psi_{\{\{a,c\}\},\{\{b\}\}} \to$$
$$\left(M+\Lambda+2M\Lambda+\Lambda^2+(M+\Lambda)\omega_L\right)/\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right),$$
$$\psi_{\{\{b,c\}\},\{\{a\}\}} \to \left(M+\Lambda+2M\Lambda+\Lambda^2+(M+\Lambda)\omega_L\right)/$$
$$\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right), \psi_{\{\{a\},\{b,c\}\},\{\}} \to$$
$$\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+M+2\Lambda)\omega_L+\omega_L^2\right)/\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right),$$
$$\psi_{\{\{b\},\{a,c\}\},\{\}} \to \left(M+\Lambda+2M\Lambda+\Lambda^2+(1+M+2\Lambda)\omega_L+\omega_L^2\right)/$$
$$\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right), \psi_{\{\{c\},\{a,b\}\},\{\}} \to$$
$$\left.\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+M+2\Lambda)\omega_L+\omega_L^2\right)/\left((1+\omega_L)\left(M+\Lambda+2M\Lambda+\Lambda^2+(1+2M+2\Lambda)\omega_L+\omega_L^2\right)\right)\right\}$$

We have rewritten this in terms of ω_L, which refers to the sum of the ω's for the two lineages involved.

Now we solve for 3 genes in two demes:

```
eqs3 = selectEqns[eqs, {2, 3}];
soln3 = Solve[eqs3, First /@ eqs3][[1]];
soln3Simp = soln3 /. soln1 /. soln2Simp;
```

As a check, if we set $\omega \to 0$, the GF is always 1, independent of Λ:

```
soln3Simp /. {ω[_] → 0} // Simplify
```

```
{GF[ω, {{}, {{a}, {b}, {c}}}, M, {{x}, {y}}, Λ] → 1,
 GF[ω, {{{a}}, {{b}, {c}}}, M, {{x}, {y}}, Λ] → 1, GF[ω, {{{b}}, {{a}, {c}}}, M, {{x}, {y}}, Λ] → 1,
 GF[ω, {{{c}}, {{a}, {b}}}, M, {{x}, {y}}, Λ] → 1, GF[ω, {{{a}, {b}}, {{c}}}, M, {{x}, {y}}, Λ] → 1,
 GF[ω, {{{a}, {c}}, {{b}}}, M, {{x}, {y}}, Λ] → 1, GF[ω, {{{b}, {c}}, {{a}}}, M, {{x}, {y}}, Λ] → 1,
 GF[ω, {{{a}, {b}, {c}}, {}}, M, {{x}, {y}}, Λ] → 1}
```

■ 1.2 Sumaries for exponentially distributed split times

▫ *The total length of the genealogy*

A relatively simple expression can be obtained for the distribution of total length of the genealogy, $T = t_{\{a\}} + t_{\{b\}} + \ldots$, for a given Λ by setting all the ω to be the same, so that $\psi = E[\exp(-\omega T)]$

```
ss = soln3Simp /. ω[_] → ω /. tidyNotation[ψ] /. Λ{x},{y} → Λ // Simplify
```

$$\Big\{\psi_{\{\},\{\{a\},\{b\},\{c\}\}} \to \big(\Lambda^4 + 5\,\Lambda^3\,(1+2\,\omega) + \Lambda^2\,\big(7+36\,\omega+37\,\omega^2\big) + 6\,\omega\,\big(1+6\,\omega+11\,\omega^2+6\,\omega^3\big) + \Lambda\,\big(3+32\,\omega+85\,\omega^2+60\,\omega^3\big) + M^2\,\big(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)\big) + M\,\big(4\,\Lambda^3+\Lambda^2\,(14+27\,\omega)+\Lambda\,\big(13+56\,\omega+53\,\omega^2\big)+3\,\big(1+7\,\omega+14\,\omega^2+8\,\omega^3\big)\big)\big)\Big/\big((1+\omega)\,(1+2\,\omega)\,\big(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2\big)\,\big(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega)\big)\big),$$

$$\psi_{\{\{a\}\},\{\{b\},\{c\}\}} \to \big(M^2\,\big(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)\big) + M\,\big(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,\big(13+46\,\omega+37\,\omega^2\big)\big) + \Lambda\,\big(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,\big(7+26\,\omega+21\,\omega^2\big)+3\,\big(1+6\,\omega+11\,\omega^2+6\,\omega^3\big)\big)\big)\Big/\big((1+\omega)\,(1+2\,\omega)\,\big(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2\big)\,\big(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega)\big)\big),$$

$$\psi_{\{\{b\}\},\{\{a\},\{c\}\}} \to \big(M^2\,\big(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)\big) + M\,\big(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,\big(13+46\,\omega+37\,\omega^2\big)\big) + \Lambda\,\big(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,\big(7+26\,\omega+21\,\omega^2\big)+3\,\big(1+6\,\omega+11\,\omega^2+6\,\omega^3\big)\big)\big)\Big/\big((1+\omega)\,(1+2\,\omega)\,\big(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2\big)\,\big(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega)\big)\big),$$

$$\psi_{\{\{c\}\},\{\{a\},\{b\}\}} \to \big(M^2\,\big(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)\big) + M\,\big(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,\big(13+46\,\omega+37\,\omega^2\big)\big) + \Lambda\,\big(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,\big(7+26\,\omega+21\,\omega^2\big)+3\,\big(1+6\,\omega+11\,\omega^2+6\,\omega^3\big)\big)\big)\Big/\big((1+\omega)\,(1+2\,\omega)\,\big(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2\big)\,\big(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega)\big)\big),$$

$$\psi_{\{\{a\},\{b\}\},\{\{c\}\}} \to \big(M^2\,\big(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)\big) + M\,\big(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,\big(13+46\,\omega+37\,\omega^2\big)\big) + \Lambda\,\big(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,\big(7+26\,\omega+21\,\omega^2\big)+3\,\big(1+6\,\omega+11\,\omega^2+6\,\omega^3\big)\big)\big)\Big/\big((1+\omega)\,(1+2\,\omega)\,\big(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2\big)\,\big(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega)\big)\big),$$

$$\psi_{\{\{a\},\{c\}\},\{\{b\}\}} \to \big(M^2\,\big(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)\big) + M\,\big(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,\big(13+46\,\omega+37\,\omega^2\big)\big) + \Lambda\,\big(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,\big(7+26\,\omega+21\,\omega^2\big)+3\,\big(1+6\,\omega+11\,\omega^2+6\,\omega^3\big)\big)\big)\Big/\big((1+\omega)\,(1+2\,\omega)\,\big(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2\big)\,\big(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega)\big)\big),$$

$$\psi_{\{\{b\},\{c\}\},\{\{a\}\}} \to \big(M^2\,\big(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)\big) + M\,\big(4\,\Lambda^3+3\,(1+\omega)\,(1+2\,\omega)^2+\Lambda^2\,(14+23\,\omega)+\Lambda\,\big(13+46\,\omega+37\,\omega^2\big)\big) + \Lambda\,\big(\Lambda^3+\Lambda^2\,(5+8\,\omega)+\Lambda\,\big(7+26\,\omega+21\,\omega^2\big)+3\,\big(1+6\,\omega+11\,\omega^2+6\,\omega^3\big)\big)\big)\Big/\big((1+\omega)\,(1+2\,\omega)\,\big(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2\big)\,\big(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega)\big)\big),$$

$$\psi_{\{\{a\},\{b\},\{c\}\},\{\}} \to \big(\Lambda^4 + 5\,\Lambda^3\,(1+2\,\omega) + \Lambda^2\,\big(7+36\,\omega+37\,\omega^2\big) + 6\,\omega\,\big(1+6\,\omega+11\,\omega^2+6\,\omega^3\big) + \Lambda\,\big(3+32\,\omega+85\,\omega^2+60\,\omega^3\big) + M^2\,\big(3+4\,\Lambda^2+9\,\omega+6\,\omega^2+2\,\Lambda\,(4+7\,\omega)\big) + M\,\big(4\,\Lambda^3+\Lambda^2\,(14+27\,\omega)+\Lambda\,\big(13+56\,\omega+53\,\omega^2\big)+3\,\big(1+7\,\omega+14\,\omega^2+8\,\omega^3\big)\big)\big)\Big/\big((1+\omega)\,(1+2\,\omega)\,\big(M+\Lambda+2\,M\,\Lambda+\Lambda^2+2\,\omega+4\,M\,\omega+4\,\Lambda\,\omega+4\,\omega^2\big)\,\big(3+\Lambda^2+12\,\omega+9\,\omega^2+\Lambda\,(4+6\,\omega)+M\,(3+2\,\Lambda+6\,\omega)\big)\big)\Big\}$$

These are the variables in the more compact notation:

```
vars = GetVars[soln3Simp] /. tidyNotation[ψ]
```

$$\{\psi_{\{\},\{\{a\},\{b\},\{c\}\}},\ \psi_{\{\{a\}\},\{\{b\},\{c\}\}},\ \psi_{\{\{b\}\},\{\{a\},\{c\}\}},\ \psi_{\{\{c\}\},\{\{a\},\{b\}\}},\ \psi_{\{\{a\},\{b\}\},\{\{c\}\}},\ \psi_{\{\{a\},\{c\}\},\{\{b\}\}},\ \psi_{\{\{b\},\{c\}\},\{\{a\}\}},\ \psi_{\{\{a\},\{b\},\{c\}\},\{\}}\}$$

We only have to worry about two kinds of configuration. For three genes in the same deme, Λ makes no difference:

```
Take[vars, 2] /. ss /. M → 0 // Simplify
```

$$\left\{\frac{1}{1+3\,\omega+2\,\omega^2},\ \frac{\Lambda}{(1+\omega)\,(1+2\,\omega)\,(\Lambda+2\,\omega)}\right\}$$

The distribution does depend on M when Λ=0:

```
Take[vars, 2] /. ss /. Λ → 0 // Simplify
```

$$\left\{\left(M+M^2+4\,M\,\omega+2\,\omega\,(1+3\,\omega)\right)\Big/\left(\left(1+M+4\,\omega+2\,M\,\omega+3\,\omega^2\right)\,(M+4\,M\,\omega+2\,\omega\,(1+2\,\omega))\right),\right.$$
$$\left.(M\,(1+M+2\,\omega))\Big/\left(\left(1+M+4\,\omega+2\,M\,\omega+3\,\omega^2\right)\,(M+4\,M\,\omega+2\,\omega\,(1+2\,\omega))\right)\right\}$$

However, the mean length of the genealogy for three genes in the same deme is independent of M for $\Lambda = 0$ - an extension of the result for two genes. This is the full expression for mean length as a function of Λ and M:

```
mnL = (-D[# /. ss, ω] & /@ Take[vars, 2]) /. ω → 0 // Simplify
```

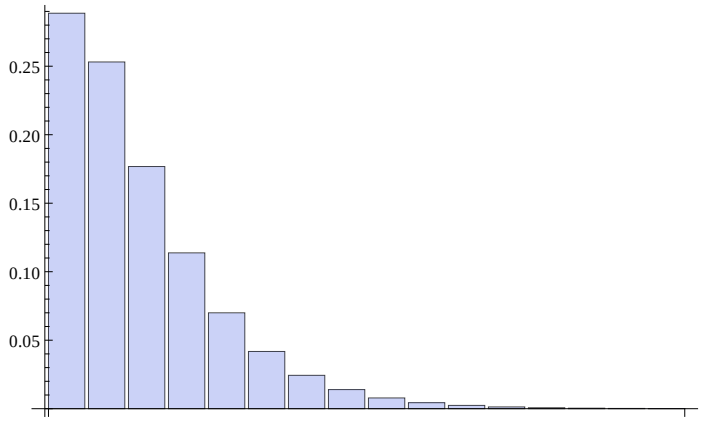
$$\left\{\frac{3\,(1+\Lambda)\,(2\,M+\Lambda)}{M+\Lambda+2\,M\,\Lambda+\Lambda^2},\ \frac{(1+\Lambda)\,(2+6\,M+3\,\Lambda)}{M+\Lambda+2\,M\,\Lambda+\Lambda^2}\right\}$$

of segregating sites

The probability that there are X segregating sites in total is $E\left[e^{-\theta\,t/2}\frac{(\theta\,t/2)^X}{X!}\right]$.

This gives the distribution of # of segregating sites, for M=0.6 , θ=1, Λ=0.7 (three genes in the same deme). Recall that Λ is the rate of splits in scaled time: we are assuming that T is exponentially distributed with mean 1/Λ.

```
cc = CoefficientList[Series[vars[[1]] /. ss /. {Λ → 0.7,
  M → 0.6, ω -> 1/2 - x}, {x, 0, 15}], x] (1/2)^(Range[0, 15]);
BarChart[cc]
{cc, Total[cc]}
```




```
{{0.28869, 0.253114, 0.176776, 0.11377, 0.0700139, 0.0417864, 0.0243667, 0.0139495, 0.00786732,
  0.00438275, 0.00241661, 0.00132101, 0.000716804, 0.000386489, 0.000207243, 0.000110592}, 0.999876}
```

1.3 Sumaries for specific T

The total length of the genealogy

We can get expressions directly in terms of the split time by taking the ILT wrt Λ:

```
ilT = InverseLaplaceTransform[Λ^-1 Take[vars, 2] /. ss, Λ, T] // Simplify;
```

This is the mean length of the genealogy in the IM model with three genes, with sampling configurations {a/b,c} and {a,b,c/∅}:

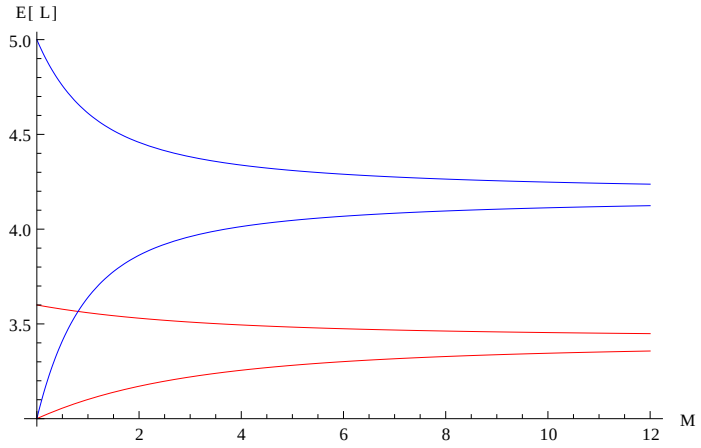
```
mn = FullSimplify[-D[ilT, ω] /. ω → 0, T > 0]
```

$$\left\{-\left(3\, e^{-\frac{1}{2}\left(1+2M+\sqrt{1+4M^2}\right)T}\left(-1-2M+\sqrt{1+4M^2}-4\, e^{\frac{1}{2}\left(1+2M+\sqrt{1+4M^2}\right)T}\sqrt{1+4M^2}+e^{\sqrt{1+4M^2}\,T}\left(1+2M+\sqrt{1+4M^2}\right)\right)\right)\Big/\left(2\sqrt{1+4M^2}\right),\ \frac{1}{2M\sqrt{1+4M^2}}\right.$$

$$\left.e^{-\frac{1}{2}\left(1+2M+\sqrt{1+4M^2}\right)T}\left(2-2\sqrt{1+4M^2}+4\, e^{\frac{1}{2}\left(1+2M+\sqrt{1+4M^2}\right)T}(1+3M)\sqrt{1+4M^2}+3M\left(1+2M-\sqrt{1+4M^2}\right)-e^{\sqrt{1+4M^2}\,T}\left(2\left(1+\sqrt{1+4M^2}\right)+3M\left(1+2M+\sqrt{1+4M^2}\right)\right)\right)\right\}$$

This shows how the expected length depends on M, for two different diveregence times T=0.3, 1 (red, blue)

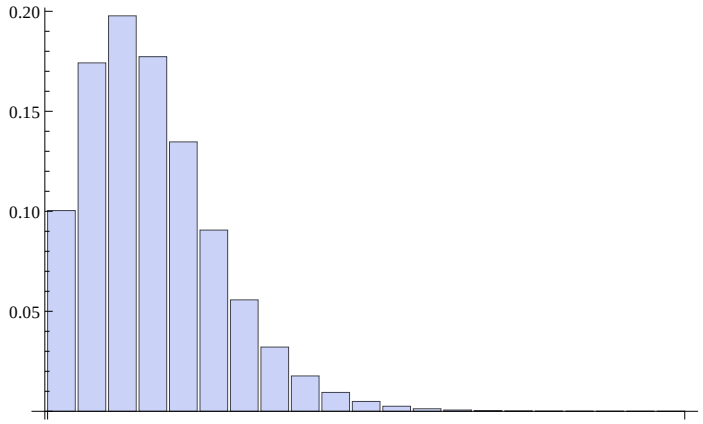
```
Plot[{mn /. {T -> 0.3}, mn /. {T → 1}}, {M, 0, 12},
 PlotStyle → {Red, Blue}, AxesLabel → {"M", "E[L]"}]
```



of segregating sites

This shows the probability distribution for the total number of segrating sites X for $T = 2$, M=0.6 and θ=1.

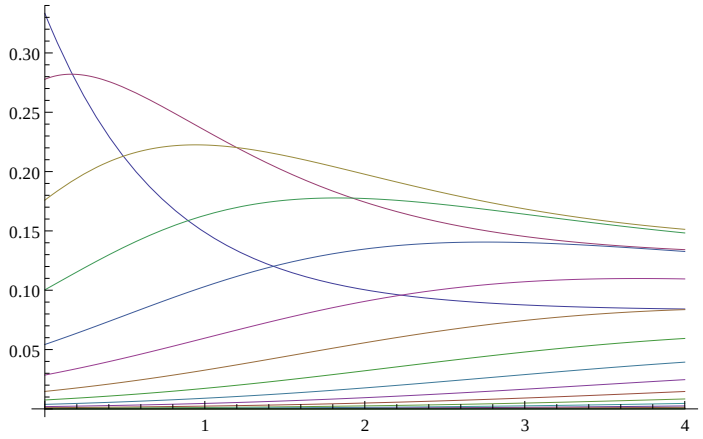
```
ccT = CoefficientList[Series[ilT[[2]] /. {T → 2, M → 0.6, ω → 1/2 - x}, {x, 0, 20}], x] (1/2)^Range[0, 20];
BarChart[ccT]
{ccT, Total[ccT]}
```




```
{{0.100361, 0.17422, 0.197747, 0.177302, 0.13472, 0.0906167, 0.0557276, 0.0321221, 0.0176861,
  0.0094335, 0.00492414, 0.00253334, 0.00129083, 0.000653541, 0.000329491, 0.000165655,
  0.0000831307, 0.0000416664, 0.0000208673, 0.0000104435, 5.23015 × 10^-6}, 0.999995}
```

This shows the probability of 0, 1,…,20 mutations as a function of T; $M = 0.6$, $\theta = 1$.

```
ccT2 = CoefficientList[Series[ilT[[2]] /. {M → 0.6, ω → 1/2 - x}, {x, 0, 20}], x] (1/2)^Range[0, 20];
Plot[ccT2, {T, 0, 4}]
```



Topological probabilities

The probability of a particular topology can be found from the LP by taking the limit of the dummy variables corresponding to internal branches incompatible with that topology. For example, to find the probability of a topology {a/b,c} we take the limit of ω_{ab} and ω_{ac} at infinity and set all other ω to zero.

```
{probtopab =
  (Limit[soln3Simp[[2, 2]] /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
   Simplify, probtopac =
  (Limit[soln3Simp[[2, 2]] /. {ω[{a, b}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
   Simplify,
 probtopbc = (Limit[soln3Simp[[2, 2]] /. {ω[{a, b}] → z α, ω[{a, c}] → z, Λ[_] → Λ}, z → ∞] /.
     ω[_] → 0) // Simplify}
```

$$\left\{\frac{2\,M+\Lambda}{3+6\,M+3\,\Lambda},\ \frac{2\,M+\Lambda}{3+6\,M+3\,\Lambda},\ \frac{3+2\,M+\Lambda}{3+6\,M+3\,\Lambda}\right\}$$

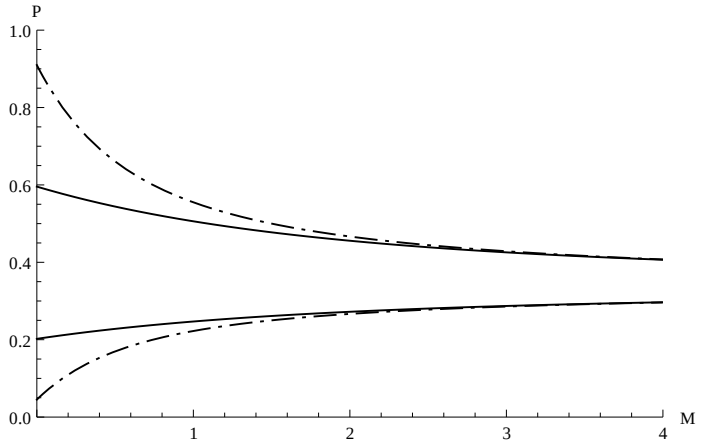
The above sum to one as they should. For a specific time we need to take the ILT of the above and divide by Λ:

```
{probab = InverseLaplaceTransform[probtopab / Λ, Λ, T],
 probac = InverseLaplaceTransform[probtopac / Λ, Λ, T],
 probbc = InverseLaplaceTransform[probtopbc / Λ, Λ, T]}
```

$$\left\{\frac{e^{-(1+2\,M)\,T}}{3\,(1+2\,M)}+\frac{2\,M}{3\,(1+2\,M)},\ \frac{e^{-(1+2\,M)\,T}}{3\,(1+2\,M)}+\frac{2\,M}{3\,(1+2\,M)},\ -\frac{2\,e^{-(1+2\,M)\,T}}{3\,(1+2\,M)}+\frac{3+2\,M}{3\,(1+2\,M)}\right\}$$

This plots topological probabilities for a triplet with sampling configuration {a/b,c} in the symmetric IM model against the scaled migration rate M for two splitting time, T=0.5 (solid lines) and T=2 (dashed lines). The chance of observing an incongruent genealogy {c,{a,b}} or {b{a,c}} (below) increases with M as congruent topologies {a,{b,c}} (above) become less likely.

```
Plot[{{probab, probbc} /. T -> 0.5, {probab, probbc} /. T -> 2},
 {M, 0, 4}, PlotRange -> {{0, 4}, {0, 1}}, AxesLabel → {"M", "P"},
 PlotStyle → {{AbsoluteThickness[1], GrayLevel[0]},
   {AbsoluteThickness[1], GrayLevel[0], AbsoluteDashing[{5, 1, 5}]}}]
```



For samples taken from the same deme, the topologies have the same probability as expected.

```
{(Limit[soln3Simp[[1, 2]] /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
   Simplify,
 (Limit[soln3Simp[[1, 2]] /. {ω[{a, b}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
   Simplify,
 (Limit[soln3Simp[[1, 2]] /. {ω[{a, b}] → z α, ω[{a, c}] → z, Λ[_] → Λ}, z → ∞] /. ω[_] → 0) //
   Simplify}
```

$$\left\{\frac{1}{3},\ \frac{1}{3},\ \frac{1}{3}\right\}$$

- ***The # of mutations on the internal branch for a given topology***

To find the GF for a particular internal branch conditional on a topology, we take the limit of the ω inconsistent with this topology at infinity and again set ω corresponding to external branches to zero. For branches {a,b} and {b,c} we have:

```
limSolGen2demab = Limit[soln3Simp[[2, 2]] /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞];
limSolGen2dem2ab = limSolGen2demab //. {ω[{a, b}] → ωAB, ω[_] → ω} // Simplify;
limSolGen2dembc = Limit[soln3Simp[[2, 2]] /. {ω[{a, b}] → z α, ω[{a, c}] → z, Λ[_] → Λ}, z → ∞];
limSolGen2dem2bc = limSolGen2dembc //. {ω[{b, c}] → ωBC, ω[_] → ω} // Simplify;
```

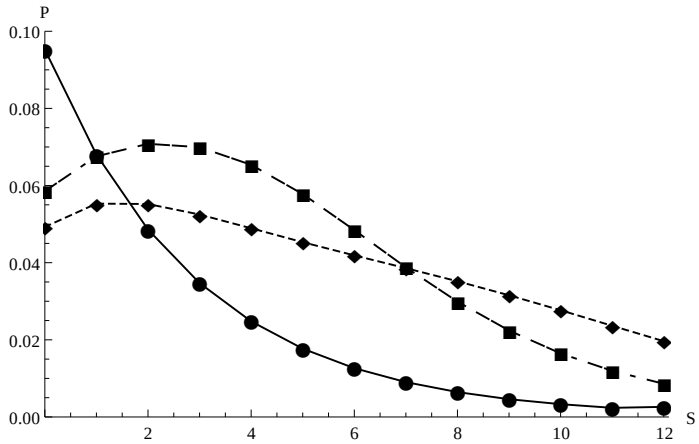
To condition on particular time, we take the ILT at T.

```
iltab = InverseLaplaceTransform[Λ^-1 limSolGen2dem2ab, Λ, T] // Simplify;
iltbc = InverseLaplaceTransform[Λ^-1 limSolGen2dem2bc, Λ, T] // Simplify;

km = 12;
clab =
  Table[List[Table[i, {i, 0, km}], CoefficientList[Series[iltab /. {ω → 0, ωAB :> 5/2 - yAB, M → 0.8},
      {yAB, 0, km}], yAB] Table[(5/2)^i, {i, 0, km}] // Chop] // Thread, {T, 0, 4, 2}];
clbc = Table[List[Table[i, {i, 0, km}], CoefficientList[Series[iltbc /. {ω -> 0, ωBC :> 5/2 - yBC,
      M → 0.8}, {yBC, 0, km}], yBC] Table[(5/2)^i, {i, 0, km}] // Chop] // Thread, {T, 0, 4, 2}];
```

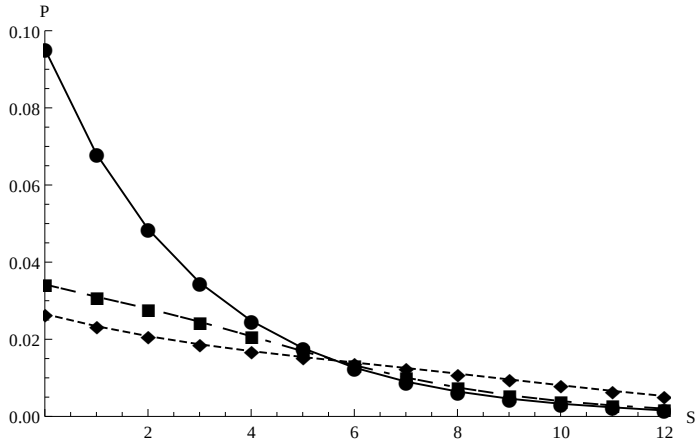
This shows the distribution of the number of mutations on internal the branch {bc} (corresponding to a topology congruent with the sampling configuration) for θ=5, M=0.8 for three different splitting times T = 0 (circles), T = 2 (squares), T = 4 (diamonds):

```
ListPlot[{clbc[[1]], clbc[[2]], clbc[[3]]}, PlotRange → {{0, 12.1}, {0, 0.1}}, PlotJoined → True,
 Mesh → All, PlotMarkers → {Automatic, Medium}, MeshStyle → {GrayLevel[0]}, AxesLabel → {"S", "P"},
 PlotStyle → {{AbsoluteThickness[1], GrayLevel[0]}, {AbsoluteThickness[1], GrayLevel[0],
    AbsoluteDashing[{7, 2, 7}]}, {AbsoluteThickness[1], GrayLevel[0], AbsoluteDashing[{3, 3, 3}]}}]
```



This shows the distribution of the number of mutations on internal the branch {a,b} for θ=5, M=0.8 for three different splitting times T=0 (circles), T= 2 (squares), T=4 (diamonds):

```
ListPlot[{clab[[1]], clab[[2]], clab[[3]}, PlotRange → {{0, 12.1}, {0, 0.1}}, PlotJoined → True,
 Mesh → All, PlotMarkers → {Automatic, Medium}, MeshStyle → {GrayLevel[0]}, AxesLabel → {"S", "P"},
 PlotStyle → {{AbsoluteThickness[1], GrayLevel[0]}, {AbsoluteThickness[1], GrayLevel[0],
    AbsoluteDashing[{7, 2, 7}]}, {AbsoluteThickness[1], GrayLevel[0], AbsoluteDashing[{3, 3, 3}]}}]
```



■ 1.4 Full results

▫ *Probabilities of mutational configurations for a given topology with exponentially distributed split times*

So far, we have derived results for the total number of segregating sites, by replacing all the branch-specific ω_S by a single ω. Now, we turn to the harder problem of finding the joint probabilities of specific configurations of mutations. This can be done by realising that the GF must be a sum of three terms, each corresponding to a different topology. We obtain the GF for a spcific topology explicitly - both for fixed Λ and for a specific split time, T. When we see an informative mutation (i.e. one shared by two of the leaves), we can just use these expressions to calculate likelihoods. If we only see singletons, we must sum over all three topologies.

Suppose that we observe at least one $\{a, b\}$ mutation. Then, we can delete any terms that depend on $\omega_{\{b,c\}}$ or $\omega_{\{a,c\}}$. The simplest way to do this is to set any terms with these in the denominator to zero. We just do this for three genes with sampling configuration {a/b,c} by taking the second row of soln3Simp:

```
soln3Simp[[2, 2]] /. {Λ[_] → 0.7, M → 0.6} /.
 {aa__/(bb__ - ω[{b, c}]) :> 0, aa__/(bb__ - ω[{a, c}]) :> 0, aa__/(bb__ + ω[{b, c}]) :> 0, aa__/(bb__ + ω[{a, c}]) :> 0};
```

This method *fails:* it mistakenly deletes terms that have $\omega[\{a, c\}]$ or $\omega[\{b, c\}]$ in the numerator as well as the denominator. *Mathematica*'s built in Limit[...] function gives the right answer - and without the need to specify Λ or M:

```
limSolGen = Limit[soln3Simp[[2, 2]] /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞];
limSolGen2 = limSolGen //. {ω[{a}] → ωS - ω[{b}] - ω[{c}], ω[{a, b}] → ωAB - ω[{c}]} // Simplify;
```

Necessarily, the remaining terms depend only on $\omega_S = \omega_{\{a\}} + \omega_{\{b\}} + \omega_{\{c\}}$ and on $\omega_{AB} = \omega_{\{a,b\}} + \omega_{\{c\}}$, which correspond to the number of mutations in the intervals before and after the coalescence of the a and b lineages. The table shows their joint probability distribution obtained by inverting w.r.t. ω_S (top to bottom) and ω_{AB} (left to right). In this example, Λ=0.7, $M = 0.6$ and θ=1.

```
km = 10;
cl =
 CoefficientList[Series[limSolGen2 /. {Λ → 0.7, M → 0.6} /. {ωS :→ 1/2 - yS, ωAB :→ 1/2 - yAB}, {yS, 0, km},
   {yAB, 0, km}], {yS, yAB}] Table[(1/2)^(i+j), {i, 0, km}, {j, 0, km}] // Chop; cl // MatrixForm
```

0.0905292	0.0359305	0.0139067	0.00527662	0.00197083	0.000726866	0.00026533
0.0314364	0.0125158	0.0048477	0.00183919	0.000686699	0.000253161	0.0000923767
0.00831566	0.00331877	0.00128632	0.00048805	0.000182197	0.0000671558	0.0000244997
0.00198247	0.000792698	0.000307428	0.00011666	0.0000435501	0.0000160508	5.8551×10^{-6}
0.000448601	0.000179639	0.0000697056	0.0000264559	9.87661×10^{-6}	3.64009×10^{-6}	$1.32782 \times 10^{-}$
0.0000985425	0.0000395058	0.0000153364	5.8218×10^{-6}	2.17358×10^{-6}	8.01111×10^{-7}	2.9223×10^{-7}
0.0000212579	8.52985×10^{-6}	3.31257×10^{-6}	1.25769×10^{-6}	4.69602×10^{-7}	1.73089×10^{-7}	$6.31414 \times 10^{-}$
4.53323×10^{-6}	1.82022×10^{-6}	7.07099×10^{-7}	2.68507×10^{-7}	1.00265×10^{-7}	3.69582×10^{-8}	$1.34826 \times 10^{-}$
9.59436×10^{-7}	3.85443×10^{-7}	1.49769×10^{-7}	5.68794×10^{-8}	2.12413×10^{-8}	7.83007×10^{-9}	$2.85656 \times 10^{-}$
2.02042×10^{-7}	8.12006×10^{-8}	3.15578×10^{-8}	1.19863×10^{-8}	4.47651×10^{-9}	1.65023×10^{-9}	$6.02058 \times 10^{-}$
4.24028×10^{-8}	1.70469×10^{-8}	6.62609×10^{-9}	2.51695×10^{-9}	9.40055×10^{-10}	3.46558×10^{-10}	$1.2644 \times 10^{-}$

Note that the first column represents the probability that there is no $\{a, b\}$ mutation - contrary to the assumption. It should be deleted. If it is included, then the total is equal to the probability of an ab topology, as expected.

```
{Total[First /@ cl], probtopac /. {Λ → 0.7, M -> 0.6}, Total[Total[cl]]}
```

```
{0.132838, 0.218391, 0.218387}
```

Probabilities of mutational configurations for a given topology with a specific T

Now, we try doing the same for a specific *time* rather than a specific Λ. That requires that we keep the expressions as functions of Λ and then take the inverse Laplace transform. The expression is ugly, but not too large. Note that the full GF is obtained just by summing the two other terms for the two other possible topologies:

```
ilt = InverseLaplaceTransform[Λ^-1 limSolGen2, Λ, T] // Simplify;
ilt /. {ωS → ωS, ωAB → ωAB} //.
 {Sqrt[1 + 4 M + 16 M^2] → α, 1 + 8 M + Sqrt[1 + 64 M^2] + 2 ωAB → 2 β, Sqrt[1 + 64 M^2] → γ};
```

```
km = 10;
cl = CoefficientList[Series[ilt /. {ωS :> 1/2 - yS, ωAB :> 1/2 - yAB, M → 0.6, T → 2}, {yS, 0, km},
      {yAB, 0, km}], {yS, yAB}] Table[(1/2)^(i+j), {i, 0, km}, {j, 0, km}] // Chop; cl // TableForm
```

0.0577467	0.0305081	0.0134862	0.00524047	0.00188228	0.000648139	0.00021878
0.0284365	0.0134435	0.00544754	0.00200202	0.00069707	0.000236416	0.0000792902
0.0102248	0.00434233	0.00163887	0.000578277	0.000197283	0.0000663165	0.0000221659
0.00302825	0.00118388	0.000425197	0.000146215	0.0000493046	0.0000164978	5.50554×10^{-6}
0.000764908	0.000282339	0.0000982776	0.0000333009	0.000011162	3.72691×10^{-6}	1.24287×10^{-6}
0.000168863	0.000060004	0.0000204992	6.89109×10^{-6}	2.30303×10^{-6}	7.68233×10^{-7}	2.56125×10^{-7}
0.0000333213	0.0000115524	3.90397×10^{-6}	1.30693×10^{-6}	4.36171×10^{-7}	1.45435×10^{-7}	4.84831×10^{-8}
6.0019×10^{-6}	2.0484×10^{-6}	6.87933×10^{-7}	2.29804×10^{-7}	7.66445×10^{-8}	2.55489×10^{-8}	8.52027×10^{-9}
1.00605×10^{-6}	3.39987×10^{-7}	1.13782×10^{-7}	3.79666×10^{-8}	1.266×10^{-8}	4.21899×10^{-9}	1.43155×10^{-9}
1.59682×10^{-7}	5.36385×10^{-8}	1.79163×10^{-8}	5.97454×10^{-9}	1.99449×10^{-9}	6.56397×10^{-10}	2.41574×10^{-10}
2.43645×10^{-8}	8.15498×10^{-9}	2.72118×10^{-9}	9.05399×10^{-10}	3.0696×10^{-10}	0	1.59591×10^{-10}

Again, the first column represents the probability that there is no $\{a, b\}$ mutation - contrary to the assumption. If it is included, then the total is the probability of an ab topology as expected.

```
{Total[Total[cl]], probab /. {M → 0.6, T → 2}}
```

```
{0.183676, 0.183678}
```

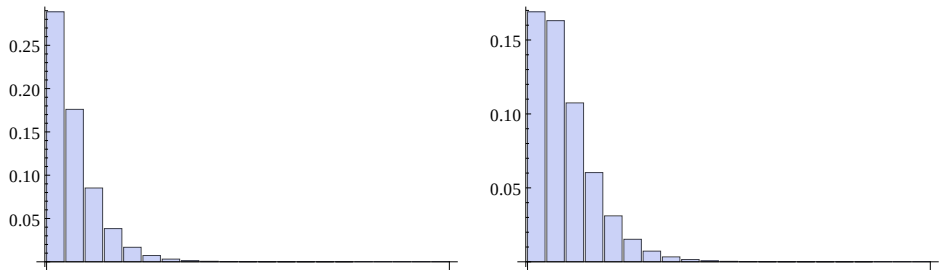
▫ *The # of singletons when there are no informative mutations*

This shows the distribution of the # of singletons for triplets with sampling configurations {a,b,c/∅} (i.e. all samples from the same deme)(left plot) and {a/b,c} (right plot). We assume that there are no mutations on internal branches (i.e., ancestral to two genes) by setting $\omega_{\{\_,\_\}}$ to the scaled mutation rate $\theta/2$. We have chosen specific values θ=1, M=0.6, Λ=0.7.

```
(singletons = soln3Simp /. tidyNotation[ψ] /. {ω{_,_} → 1/2, Λ__ → 0.7, M → 0.6});
{sing1 = CoefficientList[Series[singletons[[1, 2]] /. ω{i_} :> 1/2 - y, {y, 0, 20}], y] (1/2)^Range[0, 20],
 sing2 = CoefficientList[Series[singletons[[2, 2]] /. ω{i_} :> 1/2 - y, {y, 0, 20}], y]
   (1/2)^Range[0, 20]} // TableForm
```

0.28869	0.176037	0.0852381	0.0383407	0.016761	0.00725279	0.00313281	0.00135589
0.169064	0.163077	0.107474	0.0603337	0.0310932	0.0152276	0.00721991	0.00335067

```
Show[GraphicsGrid[{{BarChart[sing1], BarChart[sing2]}}]]
```



The probability that there will be no informative mutations, for genes in the same vs in different demes is the sum of the tables above, but is obtained more directly by setting ω to zero:

```
{singletons⟦1, 2⟧, singletons⟦2, 2⟧} /. ω_{_} → 0
```

```
{0.617854, 0.55963}
```

2. Test on real and simulated data

■ Set up

The above solutions can be used to compute the joint Log likelihood (LogL) of IM model parameters from very large numbers of loci. If we assume for the moment that loci have the same mutation rate, this requires tabulating the LogL for all observed mutational configurations, multiplying by the number of loci with each configuration. For a triplet sample with sampling configuration {a/b,c} there are three topology classes; loci may be topologically congruent (those with a bc mutation), incongruent (those with an ab or ac mutation) or uninformative. Note that we are assuming outgroup rooting such that each locus can be assigned to the three classes unambiguously (ways of dealing with finite sites mutations are discussed in the last section).

For any rooted topology, there are 3 types of mutations. For example, assuming topology {a/b,c}, we need to distinguish mutations on the internal branch (k_{bc}), those on the shorter external branches k_{ex} (since branches connected to b and c have the same length these can be lumped) and mutations on the longer external branch k_a. However, as shown before, we have the constraint $t_a = t_{b\,c} + t_b = t_{b\,c} + t_c$ and thus the GF is a function only of $\omega_{b\,c} - \omega_a$ and of $\omega_{\text{ex}} - \omega_a$, which correspond to the number of mutations in the two coalescence intervals. The joint probability of the three types of observable mutations $P[k_{\text{bc}}, k_{\text{ex}}, k_a]$ can be found by summing over all possible ways these can be partitioned amongst the two coalescent intervals:

$$P[k_{\text{bc}}, k_{\text{ex}}, k_a] = \sum_{j=0}^{k_a} \binom{k_{\text{ex}} + k_a - j}{k_a - j} \left(\tfrac{1}{3}\right)^{k_a - j} \left(\tfrac{2}{3}\right)^{k_{\text{ex}}} \binom{k_{\text{bc}} + j}{j} \left(\tfrac{1}{2}\right)^{k_{\text{bc}} + j} P[k_{b\,c} + j, k_{\text{ex}} + k_a - j]$$

We need to evaluate the GF for the number of mutations in each coalescence interval for all 3 topology classes. For the IM model with symmetric migration we have (note that for the topologically uninformative loci, we are only using the distribution of the total number of singleton rather than their full, joint distribution here):

```
limSolGenCON = Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, b}] → z α, ω[{a, c}] → z, Λ[_] → Λ}, z → ∞] //.
    {ω[{b}] → θS - ω[{a}] - ω[{c}], ω[{b, c}] → θBC - ω[{a}]} // Simplify;
```

```
limSolGenINCON = Limit[soln3Simp⟦2, 2⟧ /. {ω[{a, c}] → z α, ω[{b, c}] → z, Λ[_] → Λ}, z → ∞] //.
    {ω[{a}] → θS - ω[{b}] - ω[{c}], ω[{a, b}] → θAB - ω[{c}]} // Simplify;
```

```
limSolGenNOTOP = soln3Simp⟦2, 2⟧ /. {ω[{a, b}] → θ/2, ω[{b, c}] → θ/2,
    ω[{a, c}] → θ/2, Λ[_] → Λ, ω[{a}] → ωS, ω[{c}] → ωS, ω[{b}] → ωS} // Simplify;
```

```
ilt2typesCON = InverseLaplaceTransform[Λ^-1 limSolGenCON, Λ, T] // Simplify;
```

```
ilt2typesINCON = InverseLaplaceTransform[Λ^-1 limSolGenINCON, Λ, T] // Simplify;
```

```
ilt2typesNOTOP = InverseLaplaceTransform[Λ^-1 limSolGenNOTOP, Λ, T] // Simplify;
```

■ IM with asymmetric migration

Allowing for migration in one direction only greatly simplifies the problem. For a sample of three genes ((a) sampled one and (b) and (c) from the other), we can write down the GF by hand. Assuming that only lineage (a) can have been affected by migration (for migration in the reverse direction see section below) we have 6 equations:

```
asym = {ψ[{}, {{a}, {b, c}}] == 1/(1 + ω[{a}] + ω[{b, c}]),
   ψ[{}, {{b}, {a, c}}] == 1/(1 + ω[{b}] + ω[{a, c}]),
   ψ[{}, {{c}, {a, b}}] == 1/(1 + ω[{c}] + ω[{a, b}]),
   ψ[{}, {{a}, {b}, {c}}] == 1/(3 + ω[{a}] + ω[{b}] + ω[{c}])
     (ψ[{}, {{a}, {b, c}}] + ψ[{}, {{b}, {a, c}}] + ψ[{}, {{c}, {a, b}}]),
   ψ[{{a}}, {{b, c}}] == 1/(Λ + (M / 2) + ω[{a}] + ω[{b, c}]) (Λ + (M / 2)) ψ[{}, {{a}, {b, c}}],
   ψ[{{a}}, {{b}, {c}}] == 1/(Λ + 1 + (M / 2) + ω[{a}] + ω[{b}] + ω[{c}])
     ((Λ + (M / 2)) ψ[{}, {{a}, {b}, {c}}] + ψ[{{a}}, {{b, c}}])};
```

▫ *GF conditional on topology*

Solving the above gives the GF for a sample (a, (b,c)):

```
asymGF = (Solve[asym, First /@ asym])[[1, -1, 2]] // Simplify
```

```
((M + 2 Λ)
   ((2 + ω[{b}] + ω[{c}] + ω[{a, b}] + ω[{a, c}]) / ((1 + ω[{c}] + ω[{a, b}]) (1 + ω[{b}] + ω[{a, c}])) +
     (6 + M + 2 Λ + 4 ω[{a}] + 2 ω[{b}] + 2 ω[{c}] + 2 ω[{b, c}]) /
      ((1 + ω[{a}] + ω[{b, c}]) (M + 2 Λ + 2 ω[{a}] + 2 ω[{b, c}])))) /
 ((3 + ω[{a}] + ω[{b}] + ω[{c}]) (2 + M + 2 Λ + 2 ω[{a}] + 2 ω[{b}] + 2 ω[{c}]))
```

In this case we can invert wrt Λ to find the GF for a discrete splitting time T. The expression is complex but not vast...

```
asymGF2 = InverseLaplaceTransform[Λ^-1 asymGF, Λ, T] // Simplify
```

$$\Bigg(\Bigg(2\, e^{-\frac{1}{2}T\left(M+2\omega[\{a\}]+2\omega[\{b,c\}]\right)}\,(3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])\,(\omega[\{a\}]+\omega[\{b,c\}])\Bigg)\Big/$$
$$((1+\omega[\{b\}]+\omega[\{c\}]-\omega[\{b,c\}])\,(M+2\,\omega[\{a\}]+2\,\omega[\{b,c\}]))+$$
$$\Bigg(2\, e^{-\frac{1}{2}T\left(2+M+2\omega[\{a\}]+2\omega[\{b\}]+2\omega[\{c\}]\right)}\,(1+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])$$
$$\Big(\omega[\{c\}]+\omega[\{c\}]^2-\omega[\{a,b\}]+\omega[\{c\}]\,\omega[\{a,b\}]-\omega[\{a,c\}]-\omega[\{c\}]\,\omega[\{a,c\}]-2\,\omega[\{a,b\}]$$
$$\omega[\{a,c\}]-\omega[\{b,c\}]+\omega[\{c\}]\,\omega[\{b,c\}]+\omega[\{c\}]^2\,\omega[\{b,c\}]-\omega[\{a,b\}]\,\omega[\{b,c\}]+$$
$$\omega[\{c\}]\,\omega[\{a,b\}]\,\omega[\{b,c\}]-\omega[\{a,c\}]\,\omega[\{b,c\}]-\omega[\{a,b\}]\,\omega[\{a,c\}]\,\omega[\{b,c\}]-$$
$$2\,\omega[\{b,c\}]^2-\omega[\{c\}]\,\omega[\{b,c\}]^2-\omega[\{a,b\}]\,\omega[\{b,c\}]^2-\omega[\{a,c\}]\,\omega[\{b,c\}]^2+$$
$$\omega[\{b\}]^2\,(1+\omega[\{b,c\}])+\omega[\{a\}]\,\Big(1+\omega[\{b\}]^2+\omega[\{c\}]^2-\omega[\{a,b\}]\,\omega[\{a,c\}]+$$
$$\omega[\{c\}]\,(2+\omega[\{a,b\}]-\omega[\{b,c\}])+\omega[\{b\}]\,(2+\omega[\{c\}]+\omega[\{a,c\}]-\omega[\{b,c\}])-$$
$$2\,\omega[\{b,c\}]-\omega[\{a,b\}]\,\omega[\{b,c\}]-\omega[\{a,c\}]\,\omega[\{b,c\}]\Big)+\omega[\{b\}]$$
$$\Big(1-\omega[\{a,b\}]+\omega[\{b,c\}]+\omega[\{c\}]\,\omega[\{b,c\}]-\omega[\{b,c\}]^2+\omega[\{a,c\}]\,(1+\omega[\{b,c\}])\Big)\Big)\Bigg)\Big/$$
$$((2+M+2\,\omega[\{a\}]+2\,\omega[\{b\}]+2\,\omega[\{c\}])\,(1+\omega[\{c\}]+\omega[\{a,b\}])$$
$$(1+\omega[\{b\}]+\omega[\{a,c\}])\,(1+\omega[\{b\}]+\omega[\{c\}]-\omega[\{b,c\}]))+$$
$$\Big(M\,\Big(6+3\,M+8\,\omega[\{c\}]+2\,M\,\omega[\{c\}]+2\,\omega[\{c\}]^2+6\,\omega[\{a,b\}]+2\,M\,\omega[\{a,b\}]+$$
$$2\,\omega[\{c\}]\,\omega[\{a,b\}]+2\,\omega[\{b\}]^2\,(1+\omega[\{c\}]+\omega[\{a,b\}])+6\,\omega[\{a,c\}]+$$
$$2\,M\,\omega[\{a,c\}]+8\,\omega[\{c\}]\,\omega[\{a,c\}]+M\,\omega[\{c\}]\,\omega[\{a,c\}]+2\,\omega[\{c\}]^2\,\omega[\{a,c\}]+$$
$$6\,\omega[\{a,b\}]\,\omega[\{a,c\}]+M\,\omega[\{a,b\}]\,\omega[\{a,c\}]+2\,\omega[\{c\}]\,\omega[\{a,b\}]\,\omega[\{a,c\}]+$$
$$2\,\omega[\{a\}]^2\,(2+\omega[\{b\}]+\omega[\{c\}]+\omega[\{a,b\}]+\omega[\{a,c\}])+6\,\omega[\{b,c\}]+2\,M\,\omega[\{b,c\}]+$$
$$4\,\omega[\{c\}]\,\omega[\{b,c\}]+M\,\omega[\{c\}]\,\omega[\{b,c\}]+4\,\omega[\{a,b\}]\,\omega[\{b,c\}]+M\,\omega[\{a,b\}]\,\omega[\{b,c\}]+$$
$$4\,\omega[\{a,c\}]\,\omega[\{b,c\}]+M\,\omega[\{a,c\}]\,\omega[\{b,c\}]+2\,\omega[\{c\}]\,\omega[\{a,c\}]\,\omega[\{b,c\}]+$$
$$2\,\omega[\{a,b\}]\,\omega[\{a,c\}]\,\omega[\{b,c\}]+4\,\omega[\{b,c\}]^2+2\,\omega[\{c\}]\,\omega[\{b,c\}]^2+$$
$$2\,\omega[\{a,b\}]\,\omega[\{b,c\}]^2+2\,\omega[\{a,c\}]\,\omega[\{b,c\}]^2+\omega[\{b\}]\,\Big(8+2\,M+2\,\omega[\{c\}]^2+2\,\omega[\{a,c\}]+$$
$$4\,\omega[\{b,c\}]+M\,\omega[\{b,c\}]+2\,\omega[\{b,c\}]^2+\omega[\{a,b\}]\,(8+M+2\,\omega[\{a,c\}]+2\,\omega[\{b,c\}])+$$
$$\omega[\{c\}]\,(10+M+2\,\omega[\{a,b\}]+2\,\omega[\{a,c\}]+2\,\omega[\{b,c\}])\Big)+\omega[\{a\}]$$
$$(8+2\,M+6\,\omega[\{a,b\}]+M\,\omega[\{a,b\}]+6\,\omega[\{a,c\}]+M\,\omega[\{a,c\}]+4\,\omega[\{a,b\}]\,\omega[\{a,c\}]+$$
$$8\,\omega[\{b,c\}]+4\,\omega[\{a,b\}]\,\omega[\{b,c\}]+4\,\omega[\{a,c\}]\,\omega[\{b,c\}]+\omega[\{b\}]\,(6+M+4\,\omega[\{c\}]+$$
$$4\,\omega[\{a,b\}]+4\,\omega[\{b,c\}])+\omega[\{c\}]\,(6+M+4\,\omega[\{a,c\}]+4\,\omega[\{b,c\}]))\Big)\Big)\Big/$$
$$((2+M+2\,\omega[\{a\}]+2\,\omega[\{b\}]+2\,\omega[\{c\}])\,(1+\omega[\{c\}]+\omega[\{a,b\}])\,(1+\omega[\{b\}]+\omega[\{a,c\}])$$
$$(M+2\,\omega[\{a\}]+2\,\omega[\{b,c\}]))\Bigg)\Big/$$
$$((3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])\,(1+\omega[\{a\}]+\omega[\{b,c\}]))$$

The GF conditional on a particular topology (congruent or incongruent) can be found by taking the limits as before. The GF only depends on the θBC and θS corresponding to the two coalescence intervals:

```
limSolGenCON2 = Limit[asymGF /. {ω[{a, b}] → z α, ω[{a, c}] → z}, z → ∞] //.
   {ω[{b}] → θS - ω[{a}] - ω[{c}], ω[{b, c}] → θBC - ω[{a}]} // Simplify
```

$$((M+2\,\Lambda)\,(M+2\,(3+\theta BC+\theta S+\Lambda)))/((1+\theta BC)\,(3+\theta S)\,(M+2\,(\theta BC+\Lambda))\,(M+2\,(1+\theta S+\Lambda)))$$

```
limSolGenINCON2 = Limit[asymGF /. {ω[{a, c}] → z α, ω[{b, c}] → z}, z → ∞] //.
   {ω[{a}] → θS - ω[{b}] - ω[{c}], ω[{a, b}] → θAB - ω[{c}]} // Simplify
```

$$\frac{M+2\,\Lambda}{(1+\theta AB)\,(3+\theta S)\,(M+2\,(1+\theta S+\Lambda))}$$

Inverting the above wrt Λ gives:

```
ilt2typesCON2 = InverseLaplaceTransform[Λ^-1 limSolGenCON2, Λ, T] // Simplify
```

$$\left(-\frac{2\, e^{-\frac{1}{2}T(M+2\theta BC)}\,\theta BC\,(3+\theta S)}{(M+2\,\theta BC)\,(-1+\theta BC-\theta S)}-\frac{2\, e^{-\frac{1}{2}T(2+M+2\theta S)}\,(2+\theta BC)\,(1+\theta S)}{(1-\theta BC+\theta S)\,(2+M+2\,\theta S)}+\frac{M\,(M+2\,(3+\theta BC+\theta S))}{(M+2\,\theta BC)\,(2+M+2\,\theta S)}\right)\Big/$$
$$((1+\theta BC)\,(3+\theta S))$$

```
ilt2typesINCON2 = InverseLaplaceTransform[Λ^-1 limSolGenINCON2, Λ, T] // Simplify
```

$$\frac{M+2\, e^{-\frac{1}{2}T(2+M+2\theta S)}\,(1+\theta S)}{(1+\theta AB)\,(3+\theta S)\,(2+M+2\,\theta S)}$$

- ***Check***

Setting all ω to zero the GF must sum to one:

```
{asymGF /. {ω[_] -> 0}, asymGF2 /. {ω[_] → 0}} // Simplify
```

```
{1, 1}
```

Topological probabilities sum to one as they should:

```
{topcon = ilt2typesCON2 /. {θS → 0, θBC → 0},
  topincon = ilt2typesINCON2 /. {θS → 0, θAB → 0}} // Simplify
```

$$\left\{\frac{6-4\, e^{-\frac{1}{2}(2+M)T}+M}{3\,(2+M)},\ \frac{2\, e^{-\frac{1}{2}(2+M)T}+M}{3\,(2+M)}\right\}$$

```
topcon + 2 topincon // FullSimplify
```

```
1
```

- ***GF for topologically uninformative blocks***

To obtain the GF for topologically uninformative blocks we need to sum over all three possible topologies

```
limSolGenNOTOPbc = Limit[asymGF /. {ω[{a, b}] → z α, ω[{a, c}] → z, ω[{b, c}] → θ/2}, z → ∞] /.
   {ω[{c}] → ωsh - ω[{b}], ω[{a}] → ωa} // Simplify;
limSolGenNOTOPab = Limit[asymGF /. {ω[{a, c}] → z α, ω[{b, c}] → z, ω[{a, b}] → θ/2}, z → ∞] /.
   {ω[{a}] → ωsh - ω[{b}], ω[{c}] → ωc} // Simplify;
limSolGenNOTOPac = Limit[asymGF /. {ω[{a, b}] → z α, ω[{b, c}] → z, ω[{a, c}] → θ/2}, z → ∞] /.
   {ω[{a}] → ωsh - ω[{c}], ω[{b}] → ωb} // Simplify;
```

To GF for discrete splitting times are:

```
ilt2typesNOTOPbc = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPbc, Λ, T] // Simplify;
ilt2typesNOTOPab = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPab, Λ, T] // Simplify;
ilt2typesNOTOPac = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPac, Λ, T] // Simplify;
```

- ***GF for Total S***

The GF for the total number of mutations S is found by setting all ω[_] to be the same:

```
GfS = InverseLaplaceTransform[Λ^-1 (asymGF /. {ω[_] → ω} // Simplify), Λ, T] // Simplify
```

$$\frac{M + 4\, e^{-\frac{1}{2} T (M+4\omega)}\, \omega}{(1+\omega)(1+2\omega)(M+4\omega)}$$

This tabulates the pdfF of S:

```
test = probSasym[0.127, 4.2, 0.5, 12]
```

```
{0.0601006, 0.0853815, 0.0888486, 0.08368, 0.075961, 0.0679905,
 0.0605291, 0.0537757, 0.0477385, 0.0423665, 0.0375948, 0.0333591, 0.0296002}
```

```
0.0635 * 2
```

```
0.127
```

```
test = probSasym[0.0635, 4.2, 0.5, 12]
```

```
{0.104485, 0.197259, 0.228484, 0.194915, 0.132674, 0.0759036, 0.0379126,
 0.0170273, 0.00704105, 0.00273326, 0.00101202, 0.000362034, 0.000126404}
```

Which again must sum to one:

```
test // Total
```

```
0.999991
```

Pairwise GF

The GF for the pairwise coalescence times for the asymmetric case is:

$$\psi[\text{diff}] = \frac{1}{\Lambda + M/2 + \omega}\left(M/2 + \Lambda\right)\frac{1}{(1+\omega)};$$

```
InverseLaplaceTransform[ψ[diff] / Λ, Λ, T]
```

$$\frac{\frac{M}{M+2\omega} + \frac{2\, e^{-\frac{1}{2} T (M+2\omega)}\, \omega}{M+2\omega}}{1+\omega}$$

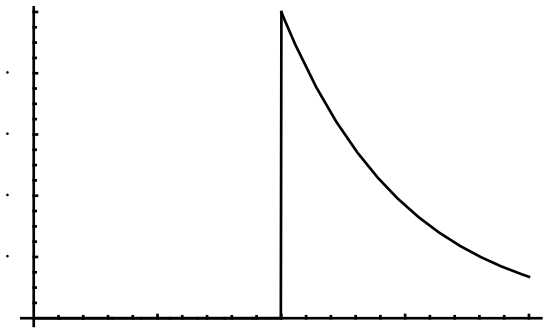
```
InverseLaplaceTransform[ψ[diff] / Λ, Λ, T] /. {M → 0}
```

$$\frac{e^{-T\omega}}{1+\omega}$$

```
PDFcoal = InverseLaplaceTransform[InverseLaplaceTransform[ψ[diff] / Λ, Λ, T], ω, t]
```

$$\frac{1}{-2+M}\, e^{-\frac{1}{2}(2+M)t}\left(\left(-e^{t} + e^{\frac{M t}{2}}\right) M + \left(-2\, e^{\frac{1}{2} M (t-T)+T} + e^{t} M\right) \text{HeavisideTheta}[t-T]\right)$$

```
Plot[PDFcoal /. {M → 0, T → 2}, {t, 0, 4}] // N
```



```
- Graphics -
```

▫ ***Migration in the reverse direction...***

With migration in the reverse direction we have 10 equations:

$$\text{asymREV} = \Big\{\psi[\{\{a\},\{b,c\}\},\{\}] == \frac{1}{1+\omega[\{a\}]+\omega[\{b,c\}]},$$

$$\psi[\{\{b\},\{a,c\}\},\{\}] == \frac{1}{1+\omega[\{b\}]+\omega[\{a,c\}]},$$

$$\psi[\{\{c\},\{a,b\}\},\{\}] == \frac{1}{1+\omega[\{c\}]+\omega[\{a,b\}]},$$

$$\psi[\{\{a,b\}\},\{\{c\}\}] == \frac{1}{\Lambda+M/2+\omega[\{a,b\}]+\omega[\{c\}]}\,(\Lambda+M/2)\,\psi[\{\{c\},\{a,b\}\},\{\}],$$

$$\psi[\{\{a,c\}\},\{\{b\}\}] == \frac{1}{\Lambda+M/2+\omega[\{a,c\}]+\omega[\{b\}]}\,(\Lambda+M/2)\,\psi[\{\{b\},\{a,c\}\},\{\}],$$

$$\psi[\{\{a\},\{b\}\},\{\{c\}\}] == \frac{1}{\Lambda+1+M/2+\omega[\{c\}]+\omega[\{b\}]+\omega[\{a\}]}\,((\Lambda+M/2)\,\psi[\{\{a\},\{b\},\{c\}\},\{\}]+\psi[\{\{a,b\}\},\{\{c\}\}]),$$

$$\psi[\{\{a\},\{c\}\},\{\{b\}\}] == \frac{1}{\Lambda+1+M/2+\omega[\{c\}]+\omega[\{b\}]+\omega[\{a\}]}\,((\Lambda+M/2)\,\psi[\{\{a\},\{b\},\{c\}\},\{\}]+\psi[\{\{a,c\}\},\{\{b\}\}]),$$

$$\psi[\{\{a\},\{b\},\{c\}\},\{\}] == \frac{1}{3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}\,(\psi[\{\{a\},\{b,c\}\},\{\}]+\psi[\{\{b\},\{a,c\}\},\{\}]+\psi[\{\{c\},\{a,b\}\},\{\}]),$$

$$\psi[\{\{a\}\},\{\{b,c\}\}] == \frac{1}{\Lambda+(M/2)+\omega[\{a\}]+\omega[\{b,c\}]}\,(\Lambda+(M/2))\,\psi[\{\{a\},\{b,c\}\},\{\}],$$

$$\psi[\{\{a\}\},\{\{b\},\{c\}\}] == \frac{1}{\Lambda+1+M+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}\,(\Lambda\,\psi[\{\{a\},\{b\},\{c\}\},\{\}]+\psi[\{\{a\}\},\{\{b,c\}\}]+M/2\,\psi[\{\{a\},\{b\}\},\{\{c\}\}]+M/2\,\psi[\{\{a\},\{c\}\},\{\{b\}\}])\Big\};$$

The GF is:

```
asymGFREV = (Solve[asymREV, First /@ asymREV])[[1, -1, 2]] // Simplify
```

$$\left(\left(\Lambda + \frac{M\left(\frac{M}{2}+\Lambda\right)}{1+\frac{M}{2}+\Lambda+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}\right)\Big/ ((3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])\,(1+\omega[\{c\}]+\omega[\{a,b\}])) + (M\,(M+2\Lambda))/((2+M+2\Lambda+2\,\omega[\{a\}]+2\,\omega[\{b\}]+2\,\omega[\{c\}])\,(1+\omega[\{c\}]+\omega[\{a,b\}])\,(M+2\Lambda+2\,\omega[\{c\}]+2\,\omega[\{a,b\}])) + \left(\frac{\Lambda + \frac{M\left(\frac{M}{2}+\Lambda\right)}{1+\frac{M}{2}+\Lambda+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}}{3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]} + (M\,(M+2\Lambda))/((2+M+2\Lambda+2\,\omega[\{a\}]+2\,\omega[\{b\}]+2\,\omega[\{c\}])\,(M+2\Lambda+2\,\omega[\{b\}]+2\,\omega[\{a,c\}]))\right)\Big/ (1+\omega[\{b\}]+\omega[\{a,c\}]) + \left(\frac{\Lambda + \frac{M\left(\frac{M}{2}+\Lambda\right)}{1+\frac{M}{2}+\Lambda+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]}}{3+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}]} + \frac{M+2\Lambda}{M+2\Lambda+2\,\omega[\{a\}]+2\,\omega[\{b,c\}]}\right)\Big/ (1+\omega[\{a\}]+\omega[\{b,c\}])\Big/ (1+M+\Lambda+\omega[\{a\}]+\omega[\{b\}]+\omega[\{c\}])$$

There are higher order terms in M which are not present with the reverse simpler sampling scheme (a single individual from the receiving population).

```
limSolGenCON2REV = Limit[asymGFREV /. {ω[{a, b}] → z α, ω[{a, c}] → z}, z → ∞] //.
   {ω[{b}] → ΘS - ω[{a}] - ω[{c}], ω[{b, c}] → ΘBC - ω[{a}]} // Simplify
```

$$\left(M^3 + 4\Lambda\,(1+\Theta S+\Lambda)\,(3+\Theta BC+\Theta S+\Lambda) + M^2\,(3+2\,\Theta BC+\Theta S+5\Lambda) + 2M\left(3+\Theta S^2+7\Lambda+3\,\Theta BC\,\Lambda+4\Lambda^2+\Theta S\,(4+3\Lambda)\right)\right)\Big/ ((1+\Theta BC)\,(3+\Theta S)\,(1+M+\Theta S+\Lambda)\,(M+2\,(\Theta BC+\Lambda))\,(M+2\,(1+\Theta S+\Lambda)))$$

```
limSolGenINCON2REV = Limit[asymGFREV /. {ω[{a, c}] → z α, ω[{b, c}] → z}, z → ∞] //.
   {ω[{a}] → ΘS - ω[{b}] - ω[{c}], ω[{a, b}] → ΘAB - ω[{c}]} // Simplify
```

$$\left(M^3 + 4\Lambda\,(\Theta AB+\Lambda)\,(1+\Theta S+\Lambda) + 2M\Lambda\,(4+3\,\Theta AB+2\,\Theta S+4\Lambda) + M^2\,(3+2\,\Theta AB+\Theta S+5\Lambda)\right)\Big/ ((1+\Theta AB)\,(3+\Theta S)\,(1+M+\Theta S+\Lambda)\,(M+2\,(\Theta AB+\Lambda))\,(M+2\,(1+\Theta S+\Lambda)))$$

Inverting the above wrt Λ gives:

```
ilt2typesCON2REV = InverseLaplaceTransform[Λ^-1 limSolGenCON2REV, Λ, T] // Simplify
```

$$\left(\frac{4\,e^{-\frac{1}{2}T\,(2+M+2\,\Theta S)}\,(1+\Theta S)}{2+M+2\,\Theta S} + \frac{4\,e^{-\frac{1}{2}T\,(M+2\,\Theta BC)}\,\Theta BC\,(3+\Theta S)}{(M+2\,\Theta BC)\,(2+M-2\,\Theta BC+2\,\Theta S)} - \left(2\,e^{-T\,(1+M+\Theta S)}\,(M\,(2+\Theta S)+(1+\Theta S)\,(4-\Theta BC+2\,\Theta S))\right)\Big/ ((1+M+\Theta S)\,(2+M-2\,\Theta BC+2\,\Theta S)) + \left(M\left(M^2+M\,(3+2\,\Theta BC+\Theta S)+2\left(3+4\,\Theta S+\Theta S^2\right)\right)\right)\Big/ ((M+2\,\Theta BC)\,(1+M+\Theta S)\,(2+M+2\,\Theta S))\right)\Big/ ((1+\Theta BC)\,(3+\Theta S))$$

```
ilt2typesINCON2REV = InverseLaplaceTransform[Λ^-1 limSolGenINCON2REV, Λ, T] // Simplify
```

$$\left(\frac{2\, e^{-\frac{1}{2}T(2+M+2\Theta S)}(1+\Theta S)(-1-2\Theta AB+\Theta S)}{(1-\Theta AB+\Theta S)(2+M+2\Theta S)}+\frac{M^2(3+M+2\Theta AB+\Theta S)}{(M+2\Theta AB)(1+M+\Theta S)(2+M+2\Theta S)}-\left(2\, e^{-\frac{1}{2}T(M+2\Theta AB)} M\Theta AB(3+\Theta S)\right)\Big/((M+2\Theta AB)(-1+\Theta AB-\Theta S)(2+M-2\Theta AB+2\Theta S))+\frac{2\, e^{-T(1+M+\Theta S)}(M+(2+\Theta AB)(1+\Theta S))}{(1+M+\Theta S)(2+M-2\Theta AB+2\Theta S)}\right)\Big/((1+\Theta AB)(3+\Theta S))$$

To obtain the GF for topologically uninformative blocks we need to sum over all three possible topologies

```
limSolGenNOTOPbcREV = Limit[asymGFREV /. {ω[{a, b}] → z α, ω[{a, c}] → z, ω[{b, c}] → Θ/2}, z → ∞] /.
   {ω[{c}] → ωsh - ω[{b}], ω[{a}] → ωa} // Simplify;
limSolGenNOTOPabREV = Limit[asymGFREV /. {ω[{a, c}] → z α, ω[{b, c}] → z, ω[{a, b}] → Θ/2}, z → ∞] /.
   {ω[{a}] → ωsh - ω[{b}], ω[{c}] → ωc} // Simplify;
limSolGenNOTOPacREV = Limit[asymGFREV /. {ω[{a, b}] → z α, ω[{b, c}] → z, ω[{a, c}] → Θ/2}, z → ∞] /.
   {ω[{a}] → ωsh - ω[{c}], ω[{b}] → ωb} // Simplify;
```

To GF for discrete splitting times are:

```
ilt2typesNOTOPbcREV = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPbcREV, Λ, T] // Simplify;
ilt2typesNOTOPabREV = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPabREV, Λ, T] // Simplify;
ilt2typesNOTOPacREV = InverseLaplaceTransform[Λ^-1 limSolGenNOTOPacREV, Λ, T] // Simplify;
```

Setting all ω to zero the GF has to sum to one:

```
asymGFREV /. {ω[_] -> 0} // Simplify
```

```
1
```

Topological probabilities sum to one as they should:

```
{topconREV = ilt2typesCON2REV /. {Θ S → 0, Θ BC → 0},
  topinconREV = ilt2typesINCON2REV /. {Θ S → 0, Θ AB → 0}} // Simplify
```

$$\left\{\frac{4\, e^{-\frac{1}{2}(2+M)T}-\frac{4\, e^{-(1+M)T}(2+M)}{1+M}+\frac{6+3M+M^2}{1+M}}{3(2+M)},\ \frac{1}{3}\left(\frac{2\, e^{-(1+M)T}}{1+M}-\frac{2\, e^{-\frac{1}{2}(2+M)T}}{2+M}+\frac{M(3+M)}{(1+M)(2+M)}\right)\right\}$$

```
topconREV + 2 topinconREV // FullSimplify
```

```
1
```

■ Wang & Hey reanalysis

■ Importing the data

This imports the Wang & Hey alignments (30,247 loci). The Dsim1/Dsim2/Dmel triplets have been filtered as described in W&H and polarized relative to Dyak. Divergent sites that are invariant in the ingroup are denoted as {1,1,1}, sites with more than two states (either due to backmutation or recombination) are denoted as {1,2,2}, {1,2,3} etc. Sites that are monomorphic in in and outgroup have been stripped, i.e. the order of mutation is retained, the sequence length information is lost. The file is still large (10 Mb):

```
WangHeyRaw = Partition[Import["/home/konrad/Downloads/ALLstripped2", "Table"], 3];
```

```
WangHeyRaw // Length
```

```
30247
```

This turns alignment into lists of site types. The first locus is:

```
WangHeyRaw2 = sitetyp[WangHeyRaw]; WangHeyRaw2[[1]]
```

```
{{1, 1, 1}, {0, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 0, 0},
 {0, 0, 1}, {1, 0, 0}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 2, 2}, {1, 1, 1}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {0, 1, 0}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 0, 0}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {0, 1, 0}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 0, 0}, {1, 1, 1}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1},
 {1, 1, 1}, {1, 1, 1}, {1, 0, 0}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 1, 1}, {1, 0, 0}}
```

E.g.the first locus has 55 variable sites in total, 10 of which are divergent between in and outgroup:

```
{WangHeyRaw2[[1]] // Length, Count[WangHeyRaw2[[1]], {1, 1, 1}]}
```

```
{55, 44}
```

■ Trimming

To keep the number of mutations per block manageable, account for mutational heterogeneity and to minimize the effect of intralocus recombination we will trim each locus to the same outgroup distance. Cutting after 16 divergent (between Dmel and Dyak) sites corresponds to roughly one third of the mean number of divergent sites in the full dataset. There are 2090 loci that fall below this cut-off, i.e. are not informative enough will be ignored:

```
WangHeyTrimRaw = DeleteCases[divcutter[16, #] & /@ WangHeyRaw2, {}];
```

We can the simply count the 6 different mutational types at each locus (in the following order {{1,0,1},{0,1,0},{1,1,0},{0,0,1},{0,1,1},{1,0,0},{1,1,1}}. Sites with multiple segregating states are ignored. Below the counts for the first locus, which only contains one internal mutations on the branch between Dmel and the two Dsim samples:

```
WangHeyTrimCounts = counttyp[#] & /@ WangHeyTrimRaw; WangHeyTrimCounts[[1]]
```

```
{0, 1, 0, 1, 1, 2, 13}
```

As expected by symmetry, the mean number of mutations on internal branches corresponding to the two different incongruent genealogies (1st and 3th value below) is the same:

```
Table[Mean[#[[i]] & /@ WangHeyTrimCounts] // N, {i, 1, 7}]
```

```
{0.126327, 0.662215, 0.122811, 0.628689, 1.81937, 3.03601, 11.8461}
```

17% of loci have no topologically informative mutations:

```
Count[(Plus @@ {#[[1]], #[[3]], #[[5]]}) & /@ WangHeyTrimCounts, 0] / (WangHeyTrimCounts // Length) // N
```

```
0.174663
```

For the triplet analysis conflicting (in terms of the topology) shared derived mutations (i.e. on internal branches) in the same locus are not possible. However, in the W&H data this is the case for 14% of loci:

```
Mean[If[Count[#[[{1, 3, 5}]], 0] < 2, 1, 0] & /@ WangHeyTrimCounts] // N
```

```
0.139823
```

First, we will remove blocks that have more than 2 topologically conflicting mutations (2.2%) . This filters out dubious alignments without biasing against the tails of the coalescence time distribution:

```
WangHeyTrimCounts2 =
 DeleteCases[(If[(Plus @@ Delete[Sort[#[[{1, 3, 5}]]], -1]) < 2, #, r]) & /@ WangHeyTrimCounts, r];
(Length[WangHeyTrimCounts] - Length[WangHeyTrimCounts2]) / Length[WangHeyTrimCounts] // N
```

```
0.0226231
```

Second, we will assume that single incongruent site are backmutations and remove those from each alignment. For 5% of loci, the incongruent, i.e. less frequent topological site cannot be determined (because there are exactly two conflicting shared derived mutations),

these loci will be removed:

```
Count[(Delete[Sort[#[[{1, 3, 5}]], 1]) & /@ WangHeyTrimCounts2, {1, 1}]/
  Length[WangHeyTrimCounts2] // N
```

```
0.050109
```

```
{Count[#[[{1, 3, 5}]] & /@ WangHeyTrimCounts2, {1, 1, 0}],
 Count[#[[{1, 3, 5}]] & /@ WangHeyTrimCounts2, {1, 0, 1}],
 Count[#[[{1, 3, 5}]] & /@ WangHeyTrimCounts2, {0, 1, 1}]}
```

```
{112, 659, 608}
```

```
WangHeyTrimCounts3 = DeleteCases[(incontrim3[#]) & /@ WangHeyTrimCounts2, r];
```

The mean number of mutations is reduced by more then half due to these trimming steps:

```
{Table[Mean[#[[i]] & /@ WangHeyTrimCounts2] // N, {i, 1, 7}],
 Table[Mean[#[[i]] & /@ WangHeyTrimCounts3] // N, {i, 1, 7}]}
```

```
{{0.101272, 0.64335, 0.096439, 0.609012, 1.81192, 3.04895, 11.8968},
 {0.0436479, 0.622279, 0.0403963, 0.588271, 1.85261, 3.0674, 11.9347}}
```

The number of loci removed in the various trimming steps is comparatively small. The only drastic reduction occurs when trimming to a fixed outgroup distance.

```
Length[WangHeyTrimCounts] - Length[WangHeyTrimCounts3]
```

```
2016
```

```
{Length[WangHeyRaw], Length[WangHeyTrimCounts],
 Length[WangHeyTrimCounts2], Length[WangHeyTrimCounts3]}
```

```
{30247, 28157, 27520, 26141}
```

■ Tests on pairwise data

It is quickest to run pairwise analyses (one Dmel , one Dsim individual) to compare the effect of the various trimming steps on parameter estimation and check against the W&H estimates.

▫ *Full dataset*

This throws out one of the Dsim individuals and condenses the data into counts of pairwise differences within the ingroup (S_in) and between ingroup and outgroup (S_out). Sites with more than two states (backmutations) are counted both in S_in and S_out, so the only difference to the W&H analysis is that we are fitting simpler IM models (with only one migration rate) and are assuming infinite sites mutations.

```
WangHeyPairs = topair[#] & /@ WangHeyRaw2; WangHeyPairs[[1]]
```

```
{10, 51}
```

We need to tabulate LogL of M and T for all observed values of S_in and S_out. There are 79*260= potential combinations.

```
{Table[Max[#[[i]] & /@ WangHeyPairs], {i, 1, 2}],
 Table[Min[#[[i]] & /@ WangHeyPairs], {i, 1, 2}], Table[Mean[#[[i]] & /@ WangHeyPairs] // N, {i, 1, 2}]}
```

```
{{79, 260}, {0, 0}, {18.0691, 46.5621}}
```

Scaling locus specific mutation rates based on the number of observed S_out values and tabulating all LogL exactly would take very long,; a much faster alternative is to bin contigs according to their outgroup divergence, 10 bins should be enough:

```
tabu = Table[Select[ WangHeyPairs, #[[2]] > (260 / 10) * i && #[[2]] < (260 / 10) * (i + 1) &], {i, 0, 9}];
bincounts = Table[Table[Count[#[[1]] & /@ tabu[[i]], k], {k, 0, 79}], {i, 1, 10}];
```

The mutation rate scalars (relative to the mean divergence across all blocks) for the bins are:

```
meanmut = Mean[WangHeyPairs] // N;
meanbin = Table[Mean[#[[2]] & /@ tabu[[i]]] // N, {i, 1, 10}] / meanmut[[2]]
```

```
{0.385613, 0.816086, 1.35804, 1.90756, 2.46236, 3.0201, 3.60045, 4.12813, 4.69419, 5.16514}
```

The joint MLE for τ and θ under a simple split model without migration are:

```
splitMLEFull = FindMaximum[
  Plus @@ Table[Total[Table[Log[Psplit[τ, ϴ, {meanbin[[i]], k}]], {k, 0, 79}] * bincounts[[i]],
    {i, 1, 10}], {τ, 0.5, 0.1, 4}, {ϴ, 8, 4, 16}]
```

```
{-94119.3, {τ → 2.18337, ϴ → 5.82369}}
```

The joint MLE for M, τ and θ for IM model with symmetric and asymmetric migration are:

```
imMLEFull = FindMaximum[Plus @@
   Table[Total[Table[Log[WilkHeSim2s[M, τ, ϴ, {meanbin[[i]], k}]], {k, 0, 79}] * bincounts[[i]],
    {i, 1, 10}], {M, 0.05, 0, 0.5}, {τ, 0.5, 0.1, 3}, {ϴ, 8, 4, 16}]
```

```
{-93467.4, {M → 0.0256439, τ → 2.69317, ϴ → 5.14413}}
```

```
imMLEFullasym = FindMaximum[
  Plus @@ Table[Total[Table[Log[asym2s[M, τ, ϴ, {meanbin[[i]], k}]], {k, 0, 79}] * bincounts[[i]],
    {i, 1, 10}], {M, 0.05, 0, 0.5}, {τ, 0.5, 0.1, 3}, {ϴ, 8, 4, 16}]
```

```
{-93466.3, {M → 0.0510174, τ → 2.69555, ϴ → 5.14185}}
```

Note that the MLE for M under the symmetric model is half that inferred for the asymmetric migration model as expected.

▫ *Trimmed to fixed divergence*

Repeating the above for the data (without trimming out backmutations and topologically conflicting mutations):

```
WangHeyTrimPairs2 = topair[#] & /@ WangHeyTrimRaw;
{Mean[WangHeyTrimPairs2] // N, Max[(#[[1]] & /@ WangHeyTrimPairs2)]}
```

```
{{6.45857, 15.9826}, 31}
```

```
tabPairs = Table[Count[(#[[1]] & /@ WangHeyTrimPairs2), i], {i, 0, 31}];
splitMLE = FindMaximum[
  Total[Table[Log[Psplit[τ, ϴ, {1, i}]], {i, 0, 31}] * tabPairs], {{τ, 0.2, 0, 4}, {ϴ, 2, 1, 4}}]
```

```
{-70303.4, {τ → 3.07509, ϴ → 1.58489}}
```

There is still a clear signal of migration (M is slightly lower than in the analysis on the full data):

```
imMLEasym = FindMaximum[Total[Table[Log[asym2s[M, τ, ϴ, {1, i}]], {i, 0, 31}] * tabPairs],
  {M, 0.02, 0, 0.5}, {τ, 3.5, 0.5, 4}, {ϴ, 2, 0.9, 4}]
```

```
{-70180., {M → 0.041833, τ → 3.84235, ϴ → 1.37638}}
```

▫ *Trimmed to fixed divergence, no backmuts and incongruent sites*

What effect does ignoring detectable backmutations and conflicting shared derived mutations have?

```
WangHeyTrimPairs3 = (Plus @@ Drop[Drop[#, 1], -1]) & /@ WangHeyTrimCounts3;
tabPairs3 = Table[Count[WangHeyTrimPairs3, i], {i, 0, Max[WangHeyTrimPairs3]}];
```

```
imMLE3asym =
 FindMaximum[Total[Table[Log[asym2s[M, τ, ϴ, {1, i}]], {i, 0, Max[WangHeyTrimPairs3]}] * tabPairs3],
  {M, 0.1, 0.001, 0.6}, {τ, 3, 0.5, 6}, {ϴ, 2, 0.9, 3}]
```

```
{-65717.1, {M → 0.0933362, τ → 3.33526, ϴ → 1.50898}}
```

■ Triplet analysis

Given the symmetry in the model there are only 3 types of loci, congruent, incongruent and those without parsimony informative sites. Within each class there are 3 types of mutations, those on the shorter external branches, those on the internal branch and those on longer external branch (the counts are listed in this order). The function sitecount sorts loci according to topology. The mutational information at each locus is summarized by counting the number of mutations on each branch. The first locus with a congruent topology has 2 mutations on the shorter external branches, one on the internal branch and one on the longer external branch.

```
WangHeyTrimCounts3 // Length
```

```
26141
```

```
WHcount = sitecount3s[WangHeyTrimCounts3]; WHcount[[1, 1]]
```

```
{2, 2, 1}
```

To make the GF calculation feasible we need to exclude 6 extreme loci with very large numbers of mutations (>16) on any one branch. This should have very little effect on parameter estimates but avoids catastrophic rounding error. We can then summarize the data as counts of distinct mutational configurations in each topology class:

```
WHCount2 = {Select[WHcount[[1]], (Max[#]) < 17 &], Select[WHcount[[2]], (Max[#]) < 15 &],
  Select[WHcount[[3]], (Max[#]) < 14 &]}; max2 = maxcount3s[WHCount2]
```

```
{{16, 12, 16}, {12, 13, 8}, {8, 13, 11}}
```

Note that the most diverse locus still has 26 mutations.

```
Table[Max[Total[#] & /@ (WHCount2[[i]])], {i, 1, 3}]
```

```
{26, 19, 18}
```

```
resWH2 = Table[Table[Count[WHCount2[[r]], {i, j, k}],
    {i, 0, max2[[r, 1]]}, {j, 0, max2[[r, 2]]}, {k, 0, max2[[r, 3]]}], {r, 1, 3}];
```

```
resWH2 // Flatten // Total
```

```
26135
```

How to best tabulate the probabilities of the observed configurations? The simplest approach is to tabulate the probabilities for all possible configurations (given the maximum number of mutations observed on each branch).

The function tripletL computes LogL under the IM model with asymmetric migration. For a single point in parameter space this takes about 1.5 seconds:

```
Timing[tripletL[0.16, 3.3, 1.5, resWH2, max2]]
```

```
{1.45609, -149746.}
```

FindMaximum uses derivatives and finds the MLE estimate in a few minutes:

```
Timing[triplMax =
  FindMaximum[tripletL[M, τ, θ, resWH2, max2], {M, 0.1, 0.001, 0.6}, {τ, 2, 1, 6}, {θ, 1, 1, 3}]]
```

```
{439.859, {-149556., {M → 0.173665, τ → 3.34091, θ → 1.39874}}}
```

▫ *Comparison between sampling schemes and with Wang and Hey*

How do the above MLEs compare to the estimates of W&H. Given that there are various differences in the scaling of parameters (W&H scale both divergence and migration relative to the mutation rate), we need to convert these into absolute values. W&H assume that Dmel and Dyak diverged 10 MYA with 10 generations per year. The μ per locus and generation for the full data and the fixed divergence are:

```
{{imMLEFull[[2, 3, 2]], imMLEFull[[2, 2, 2]], imMLEFull[[2, 1, 2]]},
  {imMLEFullasym[[2, 3, 2]], imMLEFullasym[[2, 2, 2]], imMLEFullasym[[2, 1, 2]]},
  {imMLEasym[[2, 3, 2]], imMLEasym[[2, 2, 2]], imMLEasym[[2, 1, 2]]},
  {imMLE3asym[[2, 3, 2]], imMLE3asym[[2, 2, 2]], imMLE3asym[[2, 1, 2]]} ,
  {triplMax[[2, 3, 2]], triplMax[[2, 2, 2]], triplMax[[2, 1, 2]]}} // TableForm
```

```
5.14413     2.69317     0.0256439
5.14185     2.69555     0.0510174
1.37638     3.84235     0.041833
1.50898     3.33526     0.0933362
1.39874     3.34091     0.173665
```

```
mufull = 0.1 * (meanmut[[2]] / 2) / 10000000 // N; mu1 = 0.1 * (16 / 2) / 10000000 // N;
```

Converting into absolute values is straightforward for T and N_e. Below the MLE estimates for model parameters for the full data(2nd column), the length trimmed data (3rd column), length trimmed data without backmutations and incongruent sites (4th column) and

the same data:

```
Nefs = imMLEFull[[2, 3, 2]] / (4 * mufull);
Tfs = 0.1 * imMLEFull[[2, 2, 2]] * 2 * Nefs;
mfs = imMLEFull[[2, 1, 2]] * (2.44 * 10 ^ 6) / Nefs;

Nef = imMLEFullasym[[2, 3, 2]] / (4 * mufull);
Tf = 0.1 * imMLEFullasym[[2, 2, 2]] * 2 * Nef;
mf = imMLEFullasym[[2, 1, 2]] * (2.44 * 10 ^ 6) / (2 Nef);

Ne1 = imMLEasym[[2, 3, 2]] / (4 * mu1);
T1 = 0.1 * imMLEasym[[2, 2, 2]] * 2 Ne1;
m1 = imMLEasym[[2, 1, 2]] * (2.44 * (0.00513 / 0.0055) * 10 ^ 6) / (2 Ne1);

Ne3 = imMLE3asym[[2, 3, 2]] / (4 * mu1);
T3 = 0.1 * imMLE3asym[[2, 2, 2]] * 2 Ne3;
m3 = imMLE3asym[[2, 1, 2]] (2.44 * (0.00513 / 0.0055) * 10 ^ 6) / (2 Ne3);

Netr = triplMax[[2, 3, 2]] / (4 * mu1);
Ttr = 0.1 * triplMax[[2, 2, 2]] * 2 Netr;
mtr = triplMax[[2, 1, 2]] (2.44 * (0.00513 / 0.0055) * 10 ^ 6) / (2 Netr);

{{Nefs, Nef, Ne1, Ne3, Netr }, {Tfs, Tf, T1, T3, Ttr}, {mfs, mf, m1, m3, mtr}} // TableForm
```

5.52395×10^6	5.5215×10^6	4.30119×10^6	4.71556×10^6	4.37105×10^6
2.97538×10^6	2.9767×10^6	3.30533×10^6	3.14553×10^6	2.92065×10^6
0.0113272	0.0112725	0.0110674	0.0225232	0.0452107

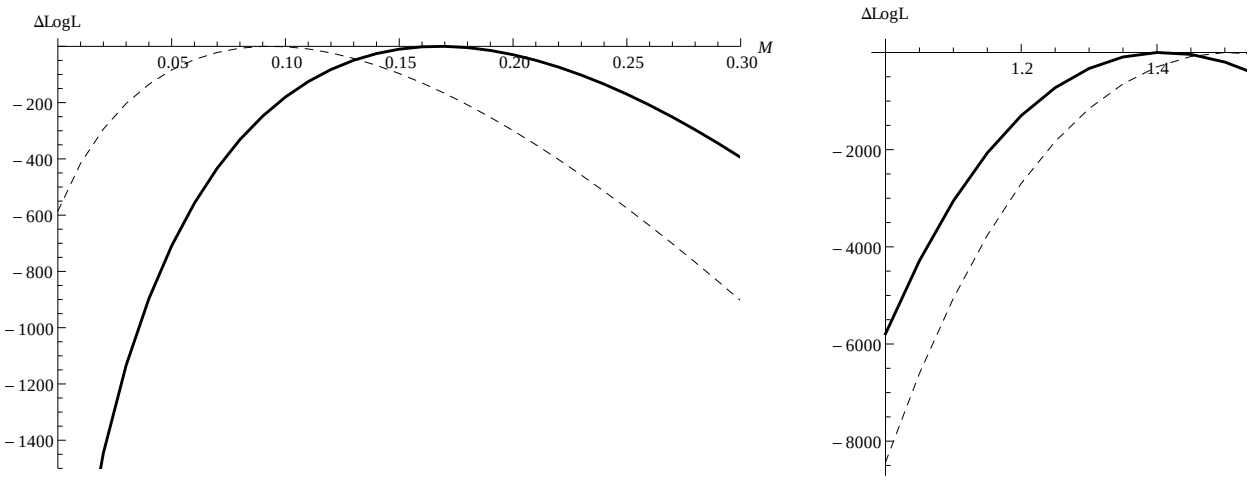
The effective population size and divergence time estimates in the pairwise analysis on the full data agree very well with those of W&H. The effective pop. size is slightly larger than the ancestral N_e estimated by W&H but smaller than their estimate of the Dsim effective population size. Given that our simpler model only contains one N_e parameter, one would expect the MLE for this parameter to be inbetween that of the ancestral population and Dsim.

Note that W&H scale migration as $M=2N_{dmel}\ m$, we are scaling $M=4N_{anc}\ m$. If we taking the larger effective pop. size of the ancestral population compared to Dmel (2.44 Mio) into account, M matches the W&H estimate (0.013) quite well. However, ignoring backmutations and incongruent mutations within blocks results in a marked increase in M and a decrease in N_e, which makes sense given that we are removing polymorphic sites. Strikingly, in the triplet analysis, the estimate of M is further increased compared to the pairwise analysis on the same dataset.

■ Comparing pairwise and triplet results

To visualize the difference between pairwise and triplet estimates, we evaluate a profile through the maximum of the likelihood surface for each parameter (fixing the other two parameters at their MLE):

The ΔLogL (relative to the maximum) for M, T and θ for pairwise (dashed line) and triplet (solid line) reveal not only that the MLE differ between the two sampling schemes (M is higher, θ lower in the triplet analysis), but also that the curvature is the same, i.e. there is no improvement in power by adding a 3rd sample which is unexpected.



What explains the difference between the pairwise and the triplet estimates (in terms of bias and power)? The triplet estimates (in particular M) should be sensitive to any model violation (both the model of sequence evolution and history) that affects the inferred frequencies of incongruent topologies. For relatively old divergence (as here) most incongruent genealogies are expected to be due to migration rather than incomplete lineage sorting. We can use the GF to find the expected frequencies of the three topology classes (congruent, incongruent and uninformative, see Table below) given the MLE for the two sampling schemes and compare these against the observed frequencies. The expected frequency of observable blocks with a congruent topologies is given by the frequency of the congruent genealogies (minus the proportion of those in that have no shared derived mutations).

```
0.821639    0.0211294    0.157232
0.772016    0.0286075    0.199377
0.750067    0.0617563    0.188177
```

There is an excess of incongruent topologies in the data (6.1%), which cannot be explained by the inferred histories. However, the observed frequencies (last row above) match the expectations from triplet MLEs (middle row) much better than those corresponding to the pairwise analysis (1st row).

Given the frequency of sites with more than 2 segregating sites, backmutations in the outgroup branch (which lead to mispolarized sites) are the most likely explanation. To check this we can look at the average number of mutations on each branch in the 3 topology classes. While congruent loci have on average fewer mutations on the two shorter external branches (i.e. those leading to the common ancestor of the two Dsim individuals) (1st row, 1st column) than on the longer external branch (2nd column); this is not the case at all for incongruent loci (2nd row). Thus most loci inferred to have an incongruent topology are due to mispolarized mutations. Given the magnitude of the excess of incongruent loci, it is actually surprising how well the triplet scheme still works!

```
Table[WHCount2[[i]] // Mean // N, {i, 1, 3}] // TableForm
```

```
1.15411     3.19507     2.46722
3.39715     1.15551     1.35812
0.562627    0.563847    2.76759
```

■ Comparison with simulated data

This imports 26141 loci simulated for triplet sampled {{b,c},a} simulated under the IM model with asymmetric migration using Hudson's ms. The values used for simulation were those inferred in the pairwise analysis (T= 3.33, M=0.0933, θ=1.5). The key question is how much statistical power can be gained from analyzing triplet samples compared to pairs ?

```
sim = ReadList["/home/konrad/Downloads/WangHeyTest3rd"][[1]]; Mean[sim] // N
```

```
{0.0178647, 0.719559, 0.020619, 0.727095, 2.33863, 3.03925}
```

Around 84% of the loci are topologically informative:

```
Total[If[#[[1]] > 0 || #[[3]] > 0 || #[[5]] > 0, 1, 0] & /@ sim] / Length[sim] // N
```

```
0.841322
```

```
res = sitecount3s[sim]; max = maxcount3s[res]
```

```
{{14, 15, 16}, {16, 8, 8}, {9, 10, 11}}
```

This summarizes loci in each topological class as counts of distinct mutational configurations:

```
res2 = configcount3s[sim];
```

In contrast to the real data, both the pairwise and triplet results closely match the true values used for simulations:

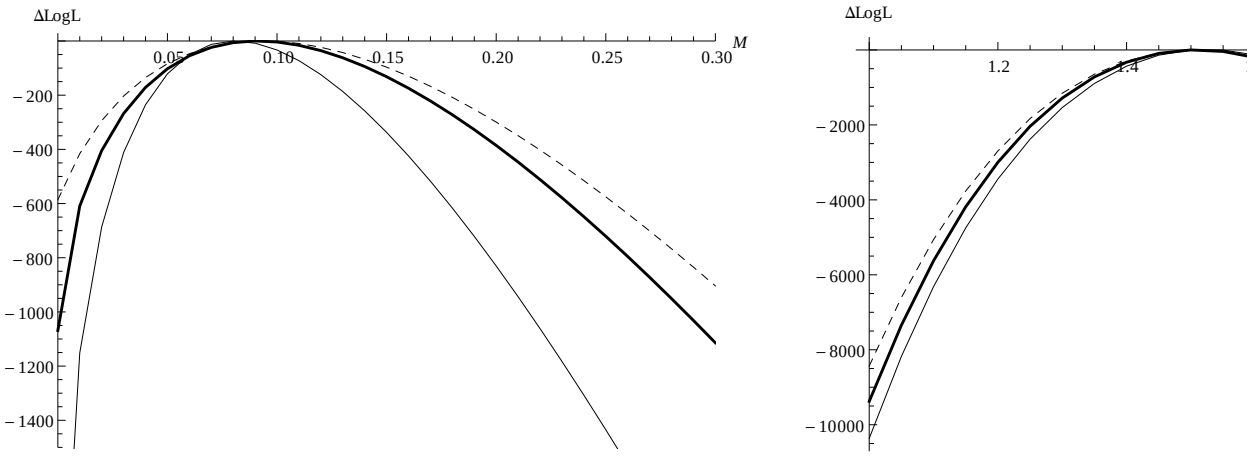
```
simpair = (Flatten[Delete[#, {{1}, {4}}]] // Total) & /@ sim;
simcount = Table[Count[simpair, i], {i, 0, Max[simpair]}];
pairsim = FindMaximum[Total[Table[Log[asym2s[M, τ, θ, {1, i}]], {i, 0, Max[simpair]}] * simcount],
  {M, 0.1, 0.001, 2}, {τ, 4, 1, 12}, {θ, 1, 0.1, 2}]
```

```
{-65619.2, {M → 0.0984103, τ → 3.23947, θ → 1.53159}}
```

```
Timing[triplsim =
  FindMaximum[tripletL[M, τ, θ, res2, max], {M, 0.1, 0.01, 0.5}, {τ, 3.3, 1, 6}, {θ, 1.5, 1, 4}]]
```

```
{376.844, {-151483., {M → 0.0922354, τ → 3.28638, θ → 1.51179}}}
```

This plots the difference in LogL from its maximum (ΔLogL) against T (left) and θ (right) for triplet (solid, thick lines) and pairwise (dashed lines) samples. As expected and in contrast to the inference on the real data, the triplet estimates are narrower. If one uses the reverse triplet sampling scheme i.e. sampling two individuals from the receiving population (see analysis of data simulated for this case with the same parameter values below), the power to infer M increases substantially (thin solid lines):



This imports data simulated under the reverse sampling scheme:

```
simREV = ReadList["/home/konrad/Downloads/WangHeyTestREV"][[1]]; Mean[simREV] // N
```

{{1, 0, 1}, {0, 1, 0}, {1, 1, 0}, {0, 0, 1}, {0, 1, 1}, {1, 0, 0}}

```
{0.0881757, 0.826556, 0.0827053, 0.819555, 2.2545, 3.01067}
```

```
resREV = sitecount3s[simREV]; max = maxcount3s[resREV]
```

```
{{14, 15, 15}, {17, 10, 11}, {10, 10, 11}}
```

```
resREV = configcount3s[simREV];

Timing[tripletLREV[0.0933, 3.3, 1.5, resREV, max]]
```

```
{1.48409, -158940.}
```

Finding the Maximum takes 15 mins....

```
Timing[triplsimREV = FindMaximum[tripletLREV[M, τ, θ, resREV, max],
   {M, 0.1, 0.01, 0.5}, {τ, 3.3, 1, 6}, {θ, 1.5, 1, 4}]]
```

```
{830.436, {-158920., {M → 0.0868254, τ → 3.26594, θ → 1.51019}}}
```

```
Neps = pairsim[[2, 3, 2]] / (4 * mu1);
Tps = 0.1 * pairsim[[2, 2, 2]] * 2 * Neps;
mps = pairsim[[2, 1, 2]] * (2.44 * 10^6) / (2 Neps);

Netrs = triplsim[[2, 3, 2]] / (4 * mu1);
Ttrs = 0.1 * triplsim[[2, 2, 2]] * 2 * Netrs;
mtrs = triplsim[[2, 1, 2]] * (2.44 * 10^6) / (2 Netrs);

{triplsim[[2, 1, 2]], pairsim[[2, 1, 2]]}
```

```
{0.0922354, 0.0984103}
```

```
{{Neps, Netrs }, {Tps, Ttrs}, {mps, mtrs}} // TableForm
```

4.78623×10^6	4.72436×10^6
3.10097×10^6	3.10521×10^6
0.0250846	0.0238185

3. Numbers of configurations

The feasibility of finding a solution for the GF depends on the number of configurations that need to be tracked. In a two-deme migration model, the number of configurations that are possible is determined by the number of ways that j lineages present between successive coalescent events can be distributed across the two populations, and the number of ways that the ancestry of n sampled individuals can be distributed amongst the j lineages present in each successive coalescence event. Specifically, the total number of configurations is:

$$\sum_{j=2}^{n} 2\,(S_{j,2}+1)\,S_{n,j} \tag{1}$$

where $S_{n,j}$ is the Stirling number of the second kind, which gives the number of ways that n lineages can be distributed over j non-empty sets. The sum is over all the intervals during which there were j extant lineages. This number grows dramatically with the number of lineages. For example, there are 92 and 2428 configurations for $n = 4$ and 6 respectively. In the IM model there are an additional $\sum_{j=2}^{n} S_{n,j}$ configurations possible in the ancestral population.

However, if we can find algebraic expressions for the GF with j genes, we do not need to track all these configurations: for example, if we know $\psi[a, b, c \setminus \emptyset]$, we can immediately find $\psi[ab, c, d \setminus \emptyset]$, for example. Therefore, the number of types of configuration that we need to calculate is only:

$$\sum_{j=2}^{n} 2\,(S_{j,2}+1) = 2^{n+1} - 4 \tag{2}$$

For example, this is 28 and 124 for $n = 4$ and 6, respectively.

Although the numbers of configurations with (say) 6 genes would be manageable for numerical calculations, extracting probabilities of mutational configurations requires that we differentiate an algebraic expression, which is given by inverting a large matrix. However, as discussed above the GF can be found directly if it is written as an expansion in M or R, each term corresponding to histories with 0, 1, …

migration or recombination events. The question is now, how many different histories do we need to track, if we allow k mutation or recombination events? Consider migration between two demes. A migration event can occur in j ways during the interval when there are j lineages, and so a single event can occur in $n + (n-1) + \ldots + 2 = (n+2)(n-1)/2$ ways. Multiple events occur independently, and so there are $((n+2)(n-1)/2)^k$ ways that k migration events can occur in the history of n genes. For example, with 4 genes there are 9, 81, 729 ways that 1, 2, 3 migration events can occur, and with 6 genes there are 20, 400, 8000 terms, respectively. With this method, we need to track many more configurations, but each is given by a much more direct calculation.

If we observe a very large number of loci, then we wish to tabulate the probability of every possible configuration of mutations. With n genes, there are $2^n - 2$ branches, and so we have $(2^n - 2)^k$ ways to throw down k mutations onto the branches. For example, even with 3 genes there are 6 possible branches, and $6^{10} \sim 6 * 10^6$ ways to distribute 10 mutations over the branches. However, the number of possibilities that we need to tabulate is much smaller than this, because the probability is determined by a much smaller number of sufficient statistics. With three genes, if we observe no mutations on the internal branches, then the probability depends only on the numbers of singletons, $\{k_a, k_b, k_c\}$, whilst if we see (say) at least one mutation ancestral to $\{a, b\}$, then we know the topology: then, the probability is determined by $\{k_a + k_b, k_{ab}, k_c\}$. In both cases, there are 14 distinct ways to divide 10 mutations over 3 classes of mutation. With more genes, more classes must be tabulated, but their number does not increase catastrophically. For example, with 6 genes and no internal mutations, there are 35 ways to distribute 10 mutations across 6 singleton classes. At the other extreme, if we know the topology, then the probability is determined by 5 independent coalescence times, and so we expect that tabulating the probability of getting $i_1, \ldots i_5$ mutations in each of the five time intervals will allow us to calculate the chance of seeing a particular set of mutations $\underline{k}$. All except the singleton class must contain mutations, and so there are roughly $\sum_{k=1}^{10} \sum_{j=1}^{k} \sum_{i=1}^{j} \sum_{l=0}^{i} 1 = 935$ configurations.

Definitions

- **Automating the recursions**
  - ***makeEqns***
  - ***makeAllEqns***
  - ***TotalRate***
  - ***Mergers***
  - ***reduceEqns***
  - ***tidyNotation***
  - ***numberOfDemes***
  - ***numberOfGenes***
  - ***GetVars***
  - ***configs***
  - ***selectEqns***
- **Data analysis**
  - ***sitecount3s***
  - ***configcount3s***
  - ***maxcount3s***
  - ***pr3s***
  - ***tripletL***
  - ***WilkHeSim2s***
  - ***asym2s***

- ***divcutter***
- ***sitetyp***
- ***topair***
- ***counttyp***
- ***incontrim3***
- ***Psplit***
- ***probSasym***