Strategies for calculating blockwise likelihoods under the coalescent

Konrad Lohse[1], Martin Chmelik[2], Simon H. Martin[3], Nicholas H. Barton[2]

[1]Institute of Evolutionary Biology

University of Edinburgh

Kings Buildings

Edinburgh EH9 3JT, UK

[2]Institute of Science and Technology

Am Campus 1

A-3400 Klosterneuburg

Austria

[3]Zoology Department

University of Cambridge

UK

Running head:

Keywords: Maximum likelihood, population divergence, gene flow, structured coalescent, generating function

Proofs to be sent to:

Konrad Lohse

Institute of Evolutionary Biology

University of Edinburgh

Kings Buildings

Edinburgh EH9 3JT, UK

Abstract

The inference of demographic history from genome data is hindered by a lack of efficient computational approaches. In particular, it has proven difficult to exploit the information contained in the distribution of genealogies across the genome. We have previously shown that the generating function (GF) of genealogies can be used to analytically compute likelihoods of demographic models from configurations of mutations in short sequence blocks. Although the GF has a simple, recursive form (Lohse *et al.*, 2011), the size of such likelihood computations explodes quickly with the number of individuals and applications of this framework have so far been limited to small samples (pairs and triplets). Here we investigate several strategies for exploiting the inherent symmetries of the coalescent and approximating the models that include reversible events. In particular, we show that the GF of genealogies can be decomposed into a set of equivalence classes which allows likelihood calculations from non-trivial samples. As an example, we implement block-wise likelihood calculations for a model of isolation with migration (IM) and two diploid samples without phase and outgroup information and compare the power of this approach to that of minimal pairwise samples. We demonstrate the new inference scheme with an analysis of two individual genomes from the sister species *Heliconius melpomene rosina* and *Heliconius cyndo*.

Genomes contain a wealth of information about the demographic and selective history of populations. However, efficiently extracting this information by fitting explicit models of population history remains a considerable computational challenge. Because it is currently not feasible to base such inferences on a complete description of the ancestral process of coalescence and recombination, inference methods generally rely on simplifying assumptions about recombination. In the most extreme case, inferences are based on the site frequency spectrum (SFS) and so ignore information contained in the physical linkage of sites altogether by treating variable sites as unlinked (Gutenkunst *et al.*, 2009; Excoffier *et al.*, 2013). Because the SFS is a function only of the expected length of genealogical branches (Griffiths & Tavaré, 1998; Chen, 2012), this greatly simplifies likelihood computations. However, much of the information about past demography is sacrificed in the process. Other methods approximate recombination along the genome as a Markov process (Li & Durbin, 2011; Harris & Nielsen, 2013). However, these methods are computationally intensive, limited to simple models (Schiffels & Durbin, 2014) and/or pairwise samples (Li & Durbin, 2011; Mailund *et al.*, 2012) and require well assembled genomes which are still only available for a handful of species.

A different class of methods assumes that recombination can be ignored within sufficiently short blocks of sequence (Hey & Nielsen, 2004; Yang, 2002). The benefit of this "multi-locus assumption" is that it gives a tractable framework for analysing linked sites, and so captures the information contained in the distribution of genealogical branches. Multi-locus methods are also attractive in practice because they naturally apply to RAD data or partially assembled genomes that can now be generated for any species (e.g. Davey & Blaxter, 2011; Hearn *et al.*, 2014).

For small samples, the probability of seeing a particular configuration of mutations at a locus can be obtained analytically. For example, Wilkinson-Herbots (2008) and Wang & Hey (2010) have derived the distribution of pairwise differences under a model of isolation with migration (IM) and Wilkinson-Herbots (2012) has extended this to a history where migration is limited to an initial period. Yang (2002) derives the

probability of mutational configurations under a divergence model for three populations and a single sample from each and Zhu & Yang (2012) have included migration between the most recently diverged pair of populations in this model. However, all of these particular cases can be calculated using a general procedure based on the generating function for the genealogy (Lohse *et al.*, 2011). Here, we explain how this can be efficiently computed for larger samples than has hitherto been possible.

The generating function of genealogies

Assuming an infinite sites mutation model and an outgroup to polarize mutations, the information in a non-recombining block of sequence can be summarized as a vector $\underline{k}$ of counts of mutations on all possible genealogical branches which are labelled by the individuals they are connected to. We have previously shown that the probability of seeing k_s mutations on branch s can be calculated directly from the Laplace Transform or generating function (GF) of genealogical branches (Lohse *et al.*, 2011). This gives a framework for computing likelihoods for any model (including the non-equilibrium histories mentioned above) and sampling scheme. Full details are given in (Lohse *et al.*, 2011). Briefly, denoting the vector of all possible branches $\underline{t}$, the GF is defined as $\psi[\underline{\omega} = E[e^{-\underline{\omega}.\underline{t}}]$, where $\underline{\omega}$ is a vector of dummy variables corresponding to $\underline{t}$. Setting the $\underline{\omega}$ to zero necessarily gives one, the total probability; differentiating with respect to ω_i and setting the $\underline{\omega}$ to zero gives (minus) the expected coalescence time. If we assume an infinite sites mutation model, the probability of seeing k_s mutations on branch s is (Lohse *et al.*, 2011, eq. 1):

$$P\left[k_S\right] = E\left[e^{-\mu t_S}\frac{\left(\mu t_S\right)^{k_S}}{k_S!}\right] = \frac{(-\mu)^{k_S}}{k_S!}\left(\frac{\partial^{k_S}\psi}{\partial {\omega_S}^{k_S}}\right)_{\omega_S=\mu} \quad (1)$$

Using the GF rather than the distribution of branches itself to compute likelihoods is convenient because we avoid the Felsenstein (1988) integral and because the GF has a very simple form: going backwards in time, the GF is a recursion over successive events in the history of the sample (Lohse *et al.*, 2011, eq. 4):

$$\psi[\Omega] = \frac{\sum_i \lambda_i \psi\left[\Omega_i\right]}{\left(\sum_i \lambda_i + \sum_{|S|=1} \omega_S\right)} \tag{2}$$

where Ω denotes the sampling configuration (i.e. the location and state of lineages) before some event i and Ω_i the sampling configuration afterwards. Events during this interval occur with a total rate $\sum_i \lambda_i$. The numerator is a sum over all the possible events i each weighted by its rate λ_i. Equation 2 applies to any history that consist of exponentially distributed events. As outlined by Lohse *et al.* (2011) the GF for models involving discrete events (population splits, bottlenecks, selective sweeps) can be found by inverting the GF of the analogous continuous model. If we know the GF, assuming an exponential rate of discrete events at rate Λ, then taking the inverse LT wrt Λ gives the GF for any fixed time of the event.

In principle, the GF recursion applies to any sample size and model and can be automated using symbolic software (such as *Mathematica*). In practice however, likelihood calculations based on the GF have so far been limited to pairs and triplets: Lohse *et al.* (2011) computed likelihoods for an IM model with unidirectional migration for three sampled genomes and Lohse *et al.* (2012) and Hearn *et al.* (2014) derived likelihoods for a range of divergence histories for a single genome from each of three populations with instantaneous admixture, including the model used by Green *et al.* (2010) to infer Neandertal admixture into modern humans (Lohse & Frantz, 2014).

There are several serious challenges in applying the GF framework to larger samples of individuals. First, the number of sample configurations (and hence GF equations) grows super-exponentially with sample size. Thus, the task of solving the GF and differentiating it to tabulate probabilities for all possible mutational configurations quickly becomes computationally prohibitive. Second, models involving reversible state transitions, such as two-way migration or recombination between loci, include a potentially infinite number of events. Solving the GF for such cases involves matrix inversions (Hobolth *et al.*, 2011; Lohse *et al.*, 2011). Third, while assuming infinite sites mutations may be convenient mathematically and realistic

for closely related sequences, this assumption becomes problematic for more distantly related outgroups that are used to polarise mutations in practice. Finally, being able to uniquely map mutations onto genealogical branches assumes phased data that are rarely available for diploid organisms given the limitations of current sequencing technologies.

In the first part of this paper, we discuss each of these problems in turn and introduce several strategies to remedy the explosion of terms and computation time. These arguments apply generally, irrespective of the peculiarities of particular demographic models and sampling schemes and suggest a computational "pipeline" for likelihood calculations for non-trivial samples of individuals (up to $n = 6$). As a concrete example, we implement maximum likelihood calculations for a model of isolation with continuous migration (IM) between two populations for unphased and unpolarized data from two diploid individuals and compare the power of this scheme to that of existing results for pairwise samples (Wilkinson-Herbots, 2008). Finally, to illustrate the new method, we estimate divergence and migration between the buttefly species *Heliconius melpomene* and *H. cyndo* using 150bp intergenic blocks (Martin *et al.*, 2013).

Methods

Partitioning the GF into equivalence classes

The GF is a sum over all possible sequences of events in the history of a sample; Edwards (1970) called them "labelled histories". Considering only coalescence events, each labelled history corresponds to a unique ranked topology (i.e. one where the order of nodes is known). Because the GF is defined in terms of genealogical branches and each topology is specified by a unique set of branches, an intuitive strategy for computing likelihoods is to partition the GF into the contributions from different topologies. To condition on a certain topology, we simply set GF terms that are incompatible with it to 0 (Lohse *et al.*, 2011). Importantly

Table 1: Fundamental quantities of genealogies for small samples (n). The total number of: branches, ordered and unordered topologies, topological shapes and equivalence classes for a sample from two populations. * the total number of mutational configurations for the 2 population IM model with up to k_m per mutations per branch.

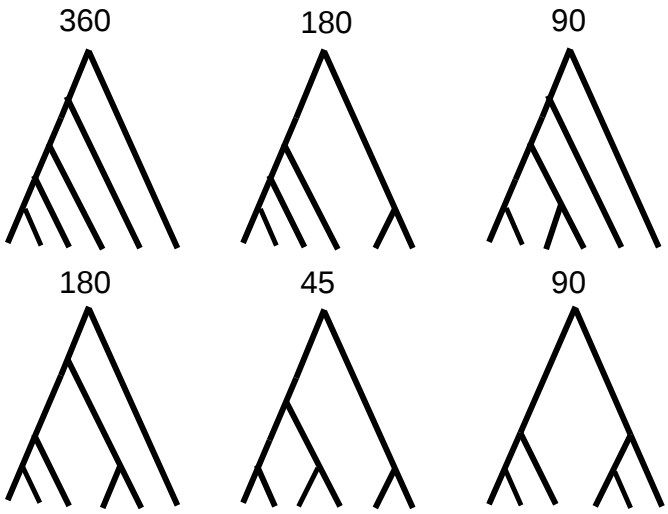
n	branches	ranked topologies	unranked topologies	EC 1 pop.	EC 2 pop.	# of config, $k_m = 3$
	$2^n - 2$	$\frac{(n!(n-1)!)}{2^{(n-1)}}$	$(2n-3)!!$	(Felsenstein, 2003)		$(2 + k_m)^{2(n-1)}$
3	6	3	3	1	2	625
4	14	18	15	2	6	15625
6	62	2,700	945	6	49	9765625
8	254	1,587,600	135,135	23	560	6103515625
10	1022	2,571,912,000	34,469,425	98	7,139	3814697265625

however, such incompatible events still contribute to the total rate $\sum_i \lambda_i$ of events in the denominator of equation 2. Then, setting all ω in the topology-conditioned GF to zero gives the probability of that particular topology. Although conditioning on a particular topology gives a GF with a manageable number of terms, it is clearly not practical to do this for all possible topologies given their sheer number even for moderate n (Table 1).

However, we can make use of a fundamental property of the coalescent, which is that samples from the same population are exchangeable. Considering the simplest case of a single population, all ranked topologies are equally likely (Hudson, 1983; Kingman, 1982). In other words, if we could somehow assign each mutation to a particular coalescence (i.e. internode) interval, we could use a much simpler GF, defined in terms of the $(n-1)$ coalescence intervals rather than the $2(n-1)$ branches for inference. This logic underlies demographic methods that use the branch length information contained in well-resolved genealogies (e.g. Nee *et al.*, 1995; Pybus *et al.*, 2002) and coalescent-based derivations of the site frequency spectrum (Griffiths & Tavaré, 1998; Chen, 2012).

However, when analysing short blocks of nuclear sequence, we are generally limited by the number of mutations on any one branch and do not know the order of nodes. Although unranked topologies are not exchangeable even for a sample from a single randomly mating population, their leaf labels are. In other

Figure 1: Unranked, unlabelled topologies define equivalence classes of genealogies. For a sample of $n = 6$ from a single population there are six equivalence classes, the number of equivalent labelled genealogies contained in each class (n_h) is shown above.



words, each unranked, unlabelled topology, or "tree shape" *sensu* Felsenstein (1978, 2003), is an equivalence class that defines a set of identically distributed genealogies (Fig. 1).

Thus, a promising strategy for deriving likelihoods for large samples is to condition the GF on one representative (random labelling) per equivalence class. The full GF can be decomposed into a weighted sum of the GFs for such class representatives:

$$\psi[\underline{\omega}] = \sum_h n_h \psi[\underline{\omega_h}] \tag{3}$$

where, n_h is the size of equivalence class h and $\underline{\omega_h} \subset \underline{\omega}$ is the set of dummy variables that corresponds to the branches of a single class representative in h. There are necessarily many fewer equivalence classes than labelled topologies. Consuider for example a sample of size $n = 6$ from a single population: there are

945 unranked topologies, but only six equivalence classes (Table 1, Fig. 1).

Crucially, the idea of tree shapes as equivalence classes extends to any model and sample. For samples from multiple populations, the equivalence classes are just the permutations of population labels on tree shapes. It is straightforward to generate and enumerate these equivalence classes (Felsenstein, 2003) for a particular sample. For example, for a sample of $n = 6$ from each of two populations (three per population), there are 49 equivalence classes (partially labelled shapes), which can be found by permuting the two population labels on the unlabelled tree shapes in figure 1.

In general, the size of each equivalence class n_h is a function of the number of permutations of individuals on population labels. For n_i individuals from population i, there are $n_i!$ permutations. Since we do not care about the orientation of nodes, each symmetric node (i.e. with identical subclades) in the equivalence class halves the number of unique permutations:

$$n_h = \prod_i n_i!/2^{n_s} \tag{4}$$

,

where n_s is the number of symmetric nodes.

Symmetries in branch lengths

Any tree shape contains at least one further symmetry: there is at least one node which connects to two leaves. Because the branches descending from that node have the same length by definition, we can combine mutations (and hence ω terms) falling on them: E.g. for a triplet genealogy with topology $(a,(b,c))$, we can combine mutations on branch b and c. The joint probability of seeing a configuration with k_a and k_b mutations can be retrieved from $P[k_a + k_b]$ by multiplying with the binomial probability, $P[k_a, k_c] = \frac{1}{2}^{k_a+k_b}\binom{k_a+k_b}{k_a}P[k_a + k_b]$.

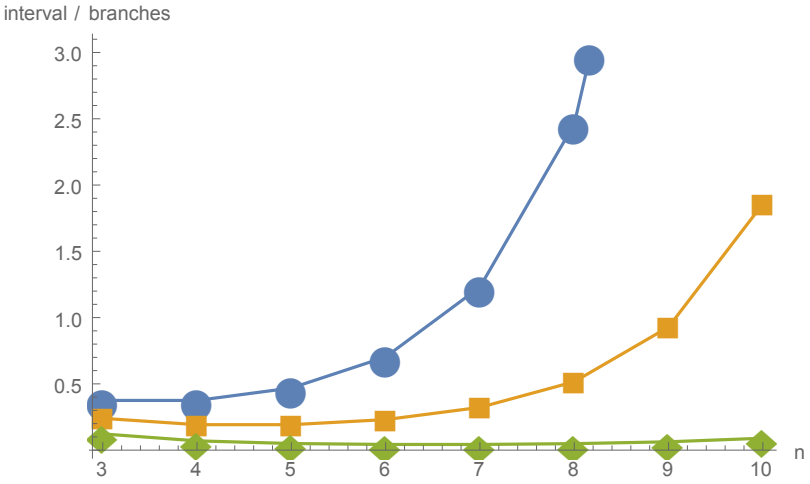
Given an ordered topology, this combinatoric argument extends to all nodes, because the GF only depends on the intervals between successive coalescence events and the number of branches involved in each. We have made use of this in implementing likelihood calculations for triplet samples (Lohse *et al.*, 2011, Supplementary Information). In general, for each genealogy there are $n-1$ coalescence intervals but $2(n-1)$ branches, so for large n, tabulating probabilities of mutational configurations in terms of intervals $P[k_{int}]$ rather than branches ($P[\underline{k}]$) halves the dimensionality of the table of mutational configurations. Obtaining $P[\underline{k}]$ from $P[k_{int}]$ is simple in principle: each assignment of mutations in an interval onto the branches present during that interval is given by a multinomial distribution. For the whole genealogy we have a product of multinomial distributions (one per interval) and $P[\underline{k}]$ is a sum over all ways of assigning mutations from intervals onto branches. For example, for the the topology $((((a),b),c),d)$ we have:

$$P[\underline{k}] = \sum_{j_{c3}=0}^{k_c} \sum_{j_{d2}=0}^{k_d} \sum_{j_{d3}=0}^{k_d - j_{d2}} \left(\frac{1}{4}\right)^{(k_a+k_b+k_c+k_d-j_{c3}-j_{d3}-j_{d2})} \left(\frac{1}{3}\right)^{(k_{ab}+j_{c3}+j_{d3})} \left(\frac{1}{2}\right)^{(k_{abc}+j_{d2})} \binom{k_a+k_b+k_c+k_d-j_{c3}-j_{d3}-j_{d2}}{k_s, k_c-j_{c3}} \binom{k_{ab}+j_{c3}+j_{d3}}{k_{ab}, j_{c3}} \binom{k_{abc}+j_{d2}}{k_{abc}} P[k_{int}]$$

where k_i denotes the number of mutations in interval i. We need two subscripts for the j's corresponding to the branch and the interval in which a mutation happens: j_{c3} is the number of mutations that occur on branch c in the third interval, the number of mutations on branch c that fall in the fourth interval is $k_c - j_{c3}$ and so on. The relationship between the k_i and $\underline{k}$ is given by the topology. For the above topology we have; $k_4 = k_a + k_b + k_c - j_{c3} + k_d - j_{d3} - j_{d2}$, $k_3 = k_{ab} + j_{c3} + j_{d3}$ and $k_2 = k_{abc} + j_{d2}$.

In principle, this logic extends to arbitrary samples. However, there are two reasons why this does not seem promising. Firstly, we are assuming a ranked/ordered topology. In order to find probabilities for an unranked topology we would also need to sum over all orderings of internal nodes. Secondly, converting mutational configurations defined in terms of interval into branches is wasteful compared to directly tabulating

Figure 2: Computing the probability of blockwise mutational configurations directly involves considering mutations on each branch. Alternatively, one can first consider mutational configurations defined in terms of coalescence intervals. However the ratio of the number of configurations defined via intervals over branches (for different global maxima k_m (3,4 and 6 from top to bottom) shows that this only gives a computational saving only for small samples (n).



mutational configurations defined in terms of branches. Suppose that we wanted to find all configurations involving up to k_m on each branch, we would need to allow for up to $k_m n$ mutations during the first interval, $(k_m)n - 1$ during the second, and so on. Including configurations with no mutations, there are a total of $n!(k_m+1)^{(n-1)}$ mutational configurations. If we compare this to the $(k_m+1)^{2(n-1)}$ configurations defined directly in terms of branches, it is clear that computing mutational configurations via internode intervals only gives a substantial saving for rather large k_m (Fig.2).

Approximating models with reversible events

Migration and recombination events are fundamentally different from coalescence and population divergence. Going backwards in time, they do not lead to simpler sample configurations. Thus, the GF for models involving migration and/or recombination is a system of coupled equations the solution of which involves matrix inversion and higher order polynomials and quickly becomes infeasible for large n (Hobolth *et al.*,

2011). As an example, we consider two populations connected by symmetric migration at rate $M = 4Nm$. Given that in practice we are often interested in histories with low or moderate migration, it seems reasonable to consider an approximate model in which the number of migration events is limited. Using a Taylor series expansion, the full GF can be decomposed into histories with $1, 2, \ldots n$ migration events (Lohse *et al.*, 2011) (the same argument applies to recombination between discrete loci). It is crucial to distinguish between M terms in the numerator and denominator. In other words, even if we stop including sampling configurations involving multiple migration events, M still contributes to the total rate $\sum_i \lambda_i$ in the denominator. We can modify the GF for a pair of genes a and b sampled from two populations connected by symmetric migration (Lohse *et al.*, 2011, eq. 9) to include an indicator variable γ that counts the number of migration events:

$$\begin{aligned} \psi^*[a\backslash b] &= \frac{\gamma M}{(M + \omega_a + \omega_b)} \psi^*[a, b\backslash\emptyset] \\ \psi^*[a, b\backslash\emptyset] &= \frac{1}{(1 + M + \omega_a + \omega_b)} (1 + \gamma M \psi^*[a\backslash b]) \end{aligned} \tag{5}$$

Expanding $\psi*$ in γ, the coefficients of $\gamma, \gamma^2 \ldots \gamma^n$ correspond to histories with $1, 2, \ldots n$ migration events. This is analogous to conditioning on a particular topology: the truncated GF does not sum to one (if we set the ω to zero), but rather gives the total probability of seeing no more than n_{max} events. This is convenient in practice because it immediately gives an estimate of the accuracy of this approximation. Expanding the solution of equation 5 around $\gamma = 0$ gives:

$$\psi^*[a\backslash b] = \sum_i \frac{M^i}{((M + \omega_a + \omega_b)(1 + M + \omega_a + \omega_b))^{(i+1)/2}} \tag{6}$$

The GF conditional on there being at most one migration event is

$$\frac{M}{(M+\omega_a+\omega_b)(1+M+\omega_a+\omega_b)} \tag{7}$$

The error of this approximation is:

$$1-\psi[a/b|M_{max}=1]_{\omega_a+\omega_b\to 0}=\frac{M}{M+1} \tag{8}$$

which is just the chance that a migration event occurs before coalescence (see Fig.3). An analogous expansion for the pairwise GF for the IM model (Lohse *et al.*, 2011, eq. 13) gives:

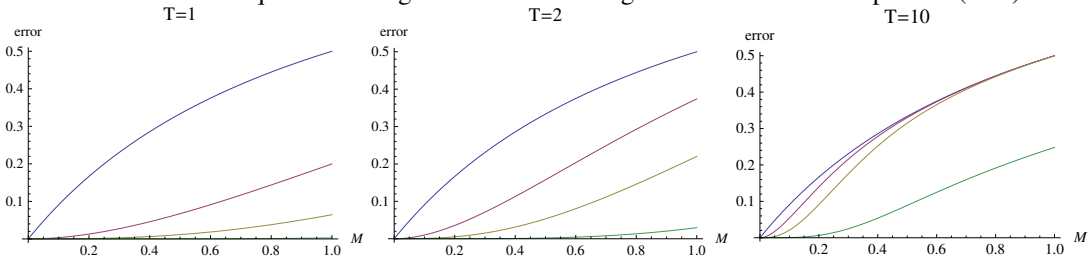
$$\psi[a/b|T,M_{max}=1]=\frac{1}{2}\left(2Me^{-MT}+\frac{2}{1+M}-\frac{2e^{-(M+1)T}M^2}{1+M}\right) \tag{9}$$

Expressions for the GF conditional on a maximum of $2,3,\ldots n$ migration events and for larger samples can be found by automating the GF recursion. While these do not appear to have a simple form, plotting the error against M and T (Fig. 3), shows that for recent divergence ($T<1$) and moderate gene flow ($M<0.5$), histories involving more than two migration events are extremely unlikely ($p<0.01$) and can be ignored to a good approximation. Considering that for large n, coalescence (which occurs at rate $n(n-1)/2$) becomes much more likely than migration (rate Mn), this approximation should be relatively robust to sample size.

The total number of mutational configurations

In principle, we can compute the probability of seeing arbitrarily many mutations on a particular branch from equation 1. In practice however, the extra information gained by explicitly including configurations with large numbers of mutations (which are very unlikely for short blocks) is limited, while the computational

Figure 3: The error in limiting the number of migration events to n_{max} =1 (eqs. 1 & 2) (red), 2 (yellow) and 4 (green) for a pairwise sample in the IM model plotted against M for different divergence times T. The results for a model of equilibrium migration without divergence is shown for comparison (blue).



cost increases. An obvious strategy is to tabulate exact probabilities only up to a certain maximum number of mutations k_m per branch and combine residual probabilities for configurations involving more than k_m mutations on one or multiple branches. As described by Lohse *et al.* (2011) and Lohse *et al.* (2012), the residual probability of seeing more than k_m mutations on a particular branch s is given by

$$P[k_s \geq k_m] = \psi[\underline{\omega}]|_{\omega_s \to 0} - \sum_{i=0}^{k_m} P[k_s = i]$$

i.e. we subtract the sum of exact probabilties for configurations involving up to k_m mutations from the marginal probability of seeing branch s.

How many such residual probabilities are there? Assuming that we want to distinguish between all $2(n-1)$ branches in a given equivalence class and use a global k_m for all branches, there are $(k_m + 2)$ possible mutation counts per branch (including those with no mutations or more than k_m mutations on a branch) which gives $(k_m + 2)^{(2n-2)}$ mutational configurations in total. For example, for $n = 6$ and $k_m = 3$ there are 9,765,625 mutational configurations per equivalence class (Table 1). Although this may seem daunting, most of these configurations are extremely unlikely, so a substantial computational saving can be made by choosing branch-specific k_m (see Supporting.nb).

Unknown phase and root

There are at least two further complications for block-wise likelihood computations in practice: First, mapping mutations onto branches assumes that the infinite sites mutation model holds between in and outgroup, which is often unrealistic in practice. Second, given the current limitations of sequencing technology, genomic data are often unphased and one would ideally incorporate phase ambiguity explicitly rather than ignore it (e.g. Lohse & Frantz, 2014) or rely on computational phasing.

When generating the GF, we have labelled branches and corresponding ω variables by the tips (leaf-nodes) they are connected to. Crucially, the full GF expressed as a sum over equivalence class representatives has unique labels for all individuals, i.e. we distinguish samples from the same population. Both unknown phase and root can be incorporated via a simple relabeling of ω variables. To incorporate unknown phase, we simply combine branches with the same set of descendants. Consider for example two samples from each of two populations. This sampling scheme gives six equivalence classes of rooted genealogies (Fig. 4). Combining all branches with the same population labels gives seven ω variables that correspond to unphased site types: ω_a, ω_b, ω_{ab}, ω_{aa}, ω_{bb}, ω_{aab}, ω_{abb}. In the absence of root information, we further combine ω corresponding to the two branches on either side of the root. Thus the set of distinguishable branch variables for the four unrooted branches (which we denote by $*$) is: $\omega_a \rightarrow \omega^*_a$, $\omega_b \rightarrow \omega^*_b$, $\omega_{ab} \rightarrow \omega^*_{ab-ab}$, $\omega_{aa} \rightarrow \omega^*_{aa-bb}$, $\omega_{bb} \rightarrow \omega^*_{aa-bbb}$, $\omega_{aab} \rightarrow \omega^*_b$, $\omega_{abb} \rightarrow \omega^*_a$. The rooted branches contributing to each unrooted branch are indicated in colour in figure 4. Without rooting the six rooted equivalence classes collapses to two unrooted equivalence classes (defined by branches t^*_{aa-bb} and t^*_{ab-ab}) (Fig. 4).

Thus in the absence of phase and root information, the four ω^* variables correspond to four site types: Heterozygous sites that are unique to one individual (either a or b), heterozygous sites that are shared by both and homozyous sites that differ between individuals. Wakeley & Hey (1997) considered these four site types in the context of estimating isolation and migration between two populations.

We can modifying eq. 8 to write the GF of an unrooted genealogy $\psi[\underline{\omega^*}]$ as a sum over unrooted equivalence classes (denoted $h*$), each of which is in turn a sum over rooted equivalence classes:

$$\psi[\underline{\omega^*}] = \sum_{h^*} \sum_{h \in h^*} \psi[\underline{\omega_h} \rightarrow \underline{\omega_h^*}] \tag{10}$$

Results

The various strategies for simplifying likelihood calculations based on the GF outlined above suggest a general "pipeline", each component of which can be automated:

1. Generate all GF terms for a given model (or in the case of discrete events its continuous analog) and sampling scheme.

2. Generate all equivalence classes h and enumerate their sizes n_h.

3. Condition the GF on one representative within each h and solve the GF successively.

4. Take the Inverse Laplace Transform with respect to the rate parameters that correspond to discrete events.

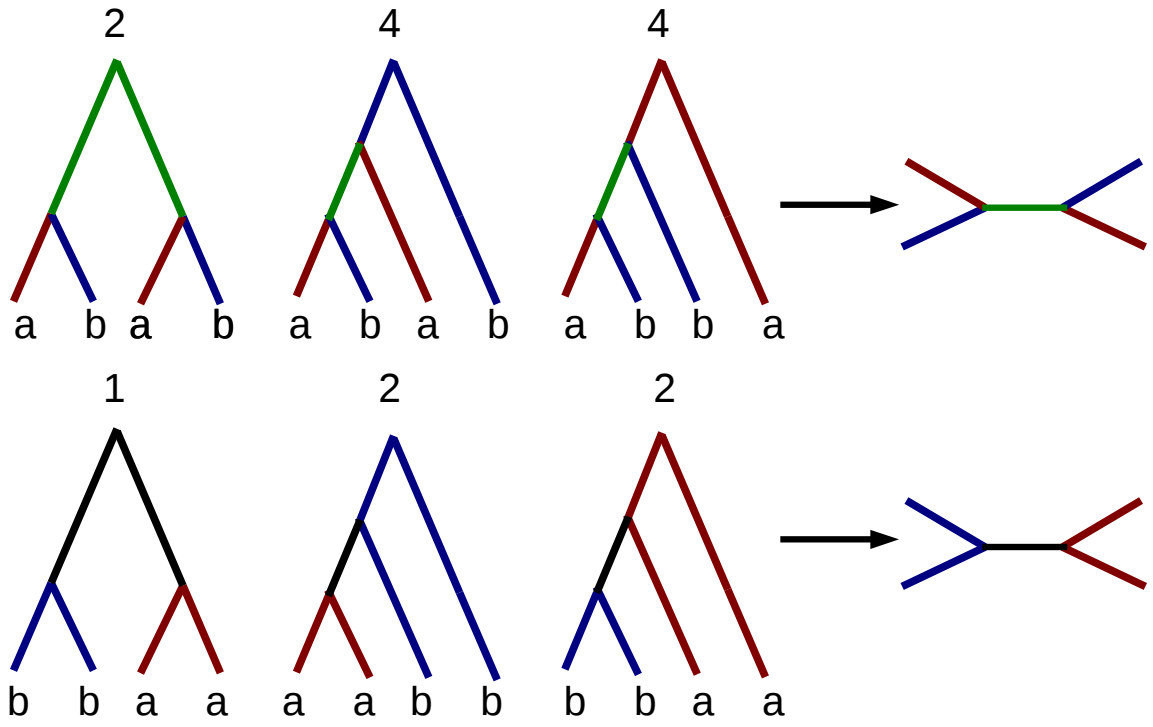
5. Re-label ω variables to combine branches and equivalence classes that are indistinguishable in the absence of root and/or phase information.

6. Find sensible k_m cut-offs for each mutation type from the data.

7. Tabulate probabilities for all mutational configurations in each equivalence class.

To demonstrate how this works, we have implemented block-wise likelihood computations for a model of isolation (at time $T \times 2N_e$ generations) with migration (at rate $M = 4N_e m$ migrants per generation)

Figure 4: For a sample of two sequences from each of two populations (a and b), there are six classes of equivalent, rooted genealogies (left); their sizes n_h are shown above. Without root information, these collapse to two unrooted genealogies (right). Without phase information, there are four mutation types that map to specific branches in the rooted genealogy: heterozygous sites unique to one sample (t_a^* and t_b^*, red and blue), shared heterozygous sites (t_{ab-ab}^*, green) and fixed, homozygous differences (t_{aa-bb}^*, black).



(IM) between two populations (labelled a and b) for unrooted and unphased data. We further assume that migration is unidirectional from a to b and that both populations and their common ancestral population are of the same effective size. Code to generate, simplify and solve the GF under this model is provided in the Supporting *Mathematica* notebook. Our automation can be used to solve the GF for samples of up to $n = 6$ (step 1–3 above). However, the inversion step (4 above) and the computation of $P[\underline{k}]$ (6 above) become prohibitively slow for larger n.

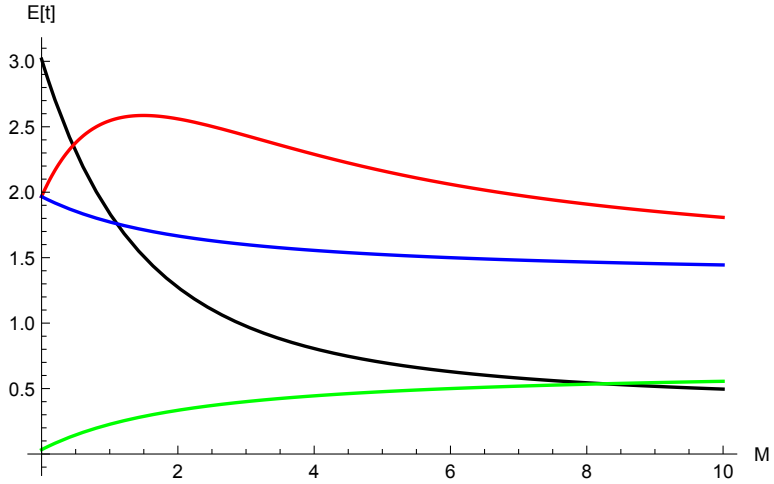
Below, we first consider some properties of unrooted, unphased genealogical branches for $n = 4$, i.e. the special case of a single diploid sample per population. We then investigate the power of likelihood calculations based on mutational configurations defined in terms of these branches and then apply these calculations to an example dataset from two species of *Heliconius* butterflies.

The distribution of unrooted branches under the IM model

We can find the expected length of any branch or combination of branches s from the GF as: $E[t_s] = -\partial\psi[\underline{\omega}]/\partial\omega_s\big|_{\underline{\omega}\to 0}$ The expressions for the expected lengths of rooted branches are cumbersome (see Supporting.nb). Surprisingly however, the expected length of the four unrooted branches (t^*_{aa-bb}, t^*_{ab-ab}, t^*_a, t^*_b), each of which is a sum over the underlying rooted branches (Fig. 9) have a relatively simple form:

$$
\begin{aligned}
E[t^*_{aa-bb}] &= \frac{e^{-(2+M)T}(-6e^T M^2 - 24e^{\frac{1}{2}(4+M)T}(1+M) + 2(1+M) + e^{(2+M)T} + (24 + 24M + 7M^2 + M^3)}{3M(1+M)(2+M)} \\
E[t^*_{ab-ab}] &= \frac{2(2e^{-(2+M)T} + M)}{3(2+M)} \\
E[t^*_a] &= \frac{4e^{-(2+M)T}(3e^T M - 1 - M - 6e^{\frac{1}{2}(4+M)T}(1+M) + e^{(2+M)T}(9 + 7M + 7M^2))}{3M(1+M)(2+M)} \\
E[t^*_b] &= \frac{4(3 - e^{-(2+M)T} + M)}{3(2+M)}
\end{aligned}
\tag{11}
$$

Figure 5: The expected length of unrooted genealogical branches (eq. 11) for a sample of $n = 4$ under the IM model of two populations (a and b) with asymmetric migration and population divergence time $T = 1.5$ ($\times 2N_e$ generations). Colours correspond to those in figure 2.

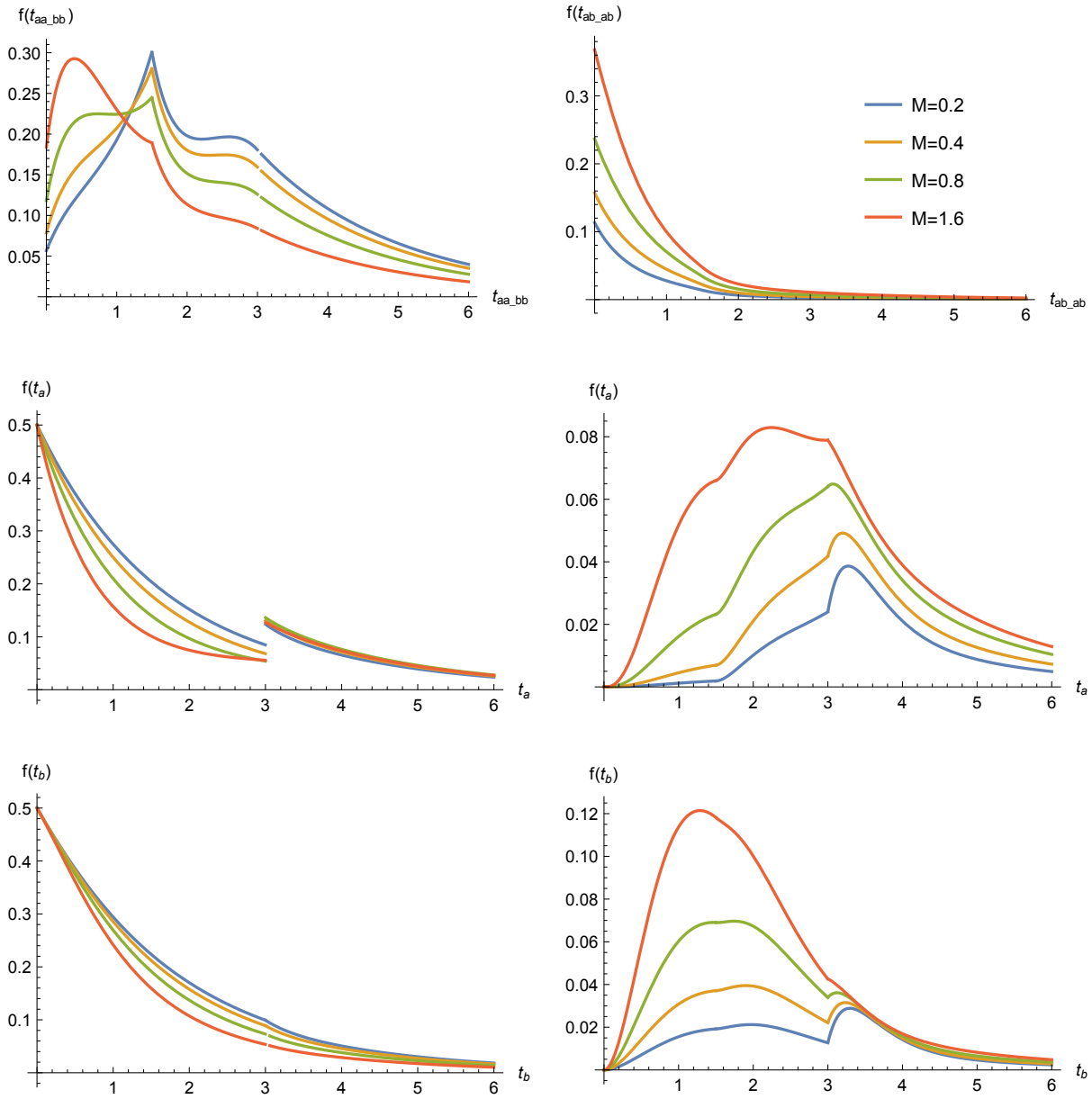


Similarly, the probability of the two possible unrooted topologies reduces to:

$$\begin{aligned} p[t^*_{aa-bb}] &= \frac{4e^{(2+M)T} + 2M}{3(2+M)} \\ p[t^*_{ab-ab}] &= 1 - p[t_{aa_bb}] \end{aligned} \tag{12}$$

We can recover the full distribution of the t^*s from the GF by taking the Inverse Laplace Transform (using *Mathematica*) with respect to the corresponding ω^*. While this does not yield simple expressions (see Supporting.nb), examining figure 6 shows that the branch length distributions are multi-modal and have discontinuities at $2T$ and that the relative size of the modes depends strongly on model parameters. For example, the first mode of t^*_{aa-bb} depends strongly on M. This suggests that much of the information about population history is contained in the shape of the branch length distribution rather than its expectation.

Figure 6: The length distribution of unrooted genealogical branches for a sample of $n = 4$ under the IM model of two populations (a and b) with asymmetric migration and population divergence at $T = 1.5$ (in $2N_e$ generations).



Power analysis

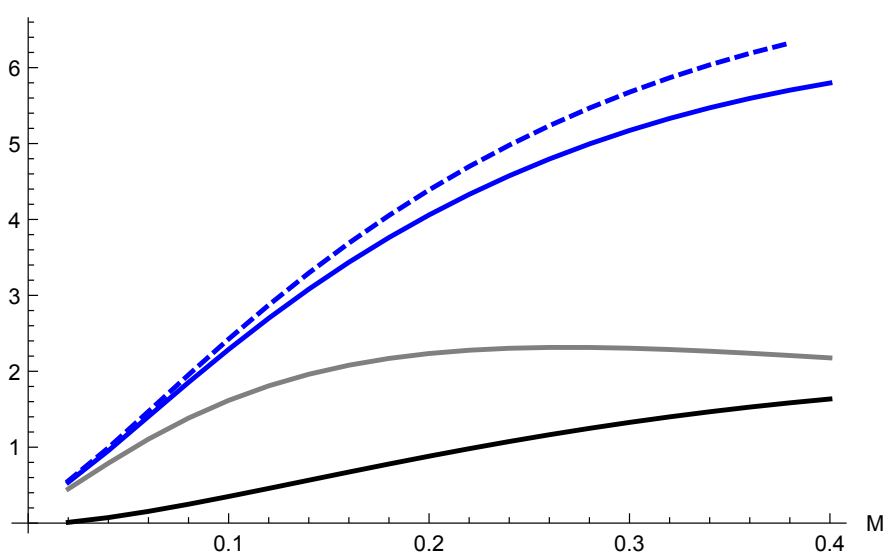
To assess the power of the block-wise maximum likelihood computations, we compared the expected difference in support ($E[\Delta lnL]$) between the IM model and a null model of strict divergence without gene flow between different sampling schemes. Note that assuming that blocks are unlinked, i.e. statistically independent, $E[\Delta lnL]$ increases linearly with the number of blocks. We assumed divergence at $T = 2.5$ and $\theta = 4N_e\mu = 1.5$. Since we are assuming a constant mutation rate per site, the scaled mutation rate θ per block can be thought of as the block lenght. Without gene flow ($M = 0$), $\theta = 1.5$ correspond to an average number of XX mutations per block.

The power to correctly identify an IM history is greater for the full scheme based on four unrooted, unphased samples than a pairwise sample ($n = 2$) (this is analogous to Wilkinson-Herbots (2012) for unidirectional migration). Contrasting the power of the full scheme for $n = 4$ with a likelihood calculation based on the total number of mutations in a sample S_T for the same sample size (black line in figure 6), shows that most of the additional information in the full scheme does not stem from the increase in sample size *per se*, but rather the addition of topology information. Also, the threshold k_m has surprisingly little effect on power. In other words; for realistically short blocks, most information is contained in the joint presence (or absence) of mutation types.

Heliconius analysis

We used the likelihood calculation outlined above to estimate divergence and gene flow between two species of *Heliconius* butterflies. The sister species *H. cydno* and *H. melpomene rosina* occur in sympatry in Central and parts of South America, are known to hybridise in the wild at a low rate (Mallet *et al.*, 2007), and have previously been shown to have experienced postdivergence gene flow (Martin *et al.*, 2013). We sampled 150 bp blocks of intergenic, autosomal sequence for one individual genome of each species from Panama

Figure 7: The power ($E[\Delta lnL]$) to distinguish between an IM model and a null model of strict divergence ($T = 2.5$) from 100 unlinked blocks of length $\theta = 1.5$ for different samples of individuals and summaries of the data: the total number of mutations (S_T) in a sample of $n = 2$ (black) and $n = 4$ (grey), full mutational configurations for unphased, unrooted data for two diploids ($n = 4$) (blue). Solid and dashed lines correspond to different maximum number of mutations per branch, $k_m = 3$ and 4 respectively.



(chi565 and ro2071). These data are part of a larger resequencing study involving high coverage genomes for four individuals of each *H. cydno* and *H. m. rosina* as well as an allopatric population of *H. melpomene* from French Guiana (Martin *et al.*, 2013). We excluded CpG islands and sites with low quality (GQ <30 and MQ<30), excessively low (<10) or high (>200) coverage and only considered sites that passed these filtering criteria in all individuals.

We partitioned the intergenic sequence into blocks of 225bp length and sampled the first 150 bases passing filtering in each block. 6.3% of blocks violated the 4-gamete criterion, i.e. contained both fixed differences and shared heterozygous sites and were removed. This sampling strategy yielded 161,726 blocks with an average per site heterozygosity of 1.7% and 1.5 % in *H. m. rosina* and *H. cydno* respecitively (Fig. 8). Summarizing the data by counting the four mutation types in each block gave 2,337 unique mutational configurations. To visualise the extent of LD between blocks, we plotted the correlation in the total number of segregating sites S between pairs of blocks against their distance (using only scaffolds >200kb) (see Supporting.nb). At a distance of 121 blocks, which corresponds to a physical distance of >27kb, the correlation drops below 0.025 and LD approaches background levels. We initially used all blocks (regardless of linkage) to maximise lnL (using Nelder-Mead simplex optimisation implemented in the *Mathematica* function *NMaximise*) to obtain point estimates of parameters under three models: i) Isolation without gene flow ii) isolation with gene flow from *H. cydno* into *H. m. melpomene* ($IM_{c\rightarrow m}$) and iii) Isolation with gene flow from *H. m. melpomene* into *H. cydno* ($IM_{c\rightarrow m}$). In all cases, we allowed *H. cydno* to have a larger N_e than the ancestor and *H. m. melpomene*. To compare models, we corrected for LD by rescaling ΔlnL with a factor of $1/121$.

We find strong support for a model of isolation with migration from *H. cydno* into *H. m. melpomene* $IM_{c\rightarrow m}$) (Table 2). This model fits significantly better than a history of strict divergence. Similarly, a model of divergence with gene flow in the opposite direction ($IM_{m\rightarrow c}$) gives a much worse fit (and no migration

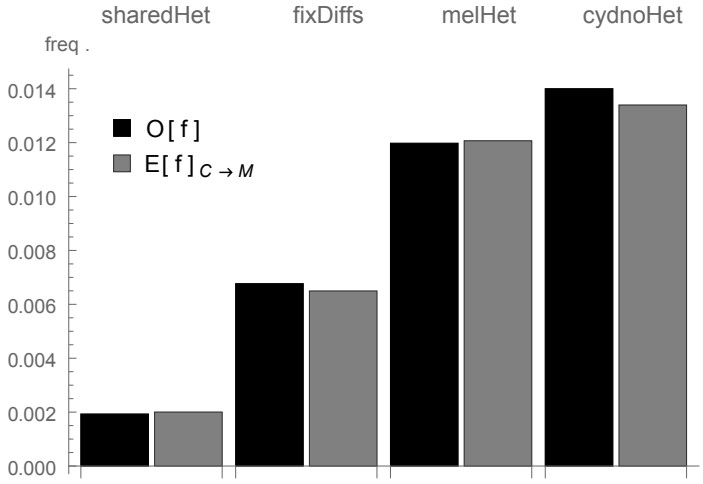
signal). Our results agree with earlier genomic analyses of this species pair that found support for post-divergence gene flow based on D-statistics (Martin *et al.*, 2013), model-based analyses of small numbers nuclear loci (Kronforst *et al.*, 2013) or genomewide SNP frequencies analysed using approximate Bayesian computation. Asymmetrical gene flow from *H. cydno* into *H. m. rosina* has also been reported previously, and could be explained by the fact that F1 hybrids resemble *H. m. rosina* more closely due to dominance relationships among wing patterning alleles, possibly making F1s more attractive to *H. melpomene* (Kronforst *et al.*, 2006, Martin et al. , *in prep.*).

A recent direct, genome-wide estimate of the mutation rate for *H. melpomene* (Keightley *et al.*, 2015) allows us to convert parameter estimates into absolute values. Assuming a spontaneous mutation rate of 2.9×10^{-9} per site and generation and using the ratio of divergence between *H. m. rosina* and the more distantly-related 'silvaniform' clade of *Heliconius* at synonymous and intergenic sites to estimate selective constraint on intergenic sites, gives an effective mutation rate of $\mu = 1.9 \times 10^{-9}$ (Martin et al. , *in prep.*). Applying this rate to our estimate of θ and assuming four generations per year, gives an N_e of 1.23×10^6 for *H. m. rosina* and the common ancestral population and 3.04×10^6 for *H. cydno*. Species divergence is estimated to have occurred 880 KY. Interestingly, this is considerably more recent than estimates obtained using approximate Bayesian computation and calibrations based on mitochondrial genealogies (Kronforst *et al.*, 2013, Martin et al. , *in prep.*). The IM history we estimated for the two *Heliconius* species fits the observed genome-wide frequency of the four mutation types (i.e. the folded, two population SFS) very well (Fig. 8).

Table 2: Top: Support (ΔlnL relative to the best model) for isolation with migration and strict divergence (Div) between *H. m. rosina* and *H. cydno*. Migration from *H. cydno* into *H. m. rosina* ($IM_{c\to m}$) fits better than migration in the other direction ($IM_{c\to m}$). Bottom: Maximum likelihood estimates of parameters under the $IM_{c\to m}$ model (scaled estimates in brackets).

Div	$IM_{m\to c}$	$IM_{c\to m}$	
-42.6	-8.4	0	
θ (N_e)	θ_C (N_e)	T	M
1.40 (1.23×10^6)	θ (3.04×10^6)	1.44 (0.88MY)	1.39

Figure 8: The genome wide frequencies of the four site types used in the blockwise likelihood computation: i) heterozygous sites unique to either *H. m. melpomene* or ii) *H. cydno* iii) shared in both species and iv) fixed differences. The exptected frequencies under the IM history estimated from the data (Table 2) was computed using eq. 11 (grey).



Discussion

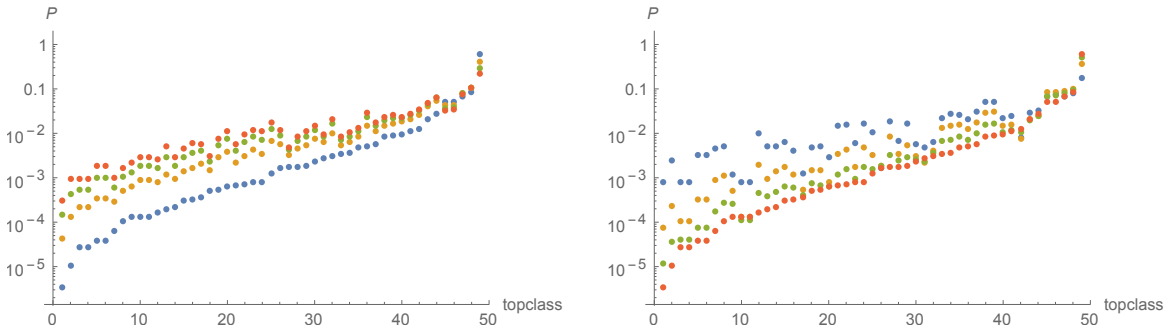
We have shown that combining equivalent topologies and branches gives an efficient strategy for exploiting the symmetries inherent in the coalescent. In the GF framework, this combinatorial partitioning can be automated using symbolic software such as *Mathematica*. Starting with the full GF for a particular model and sampling scheme, one picks out terms that contribute to one representative genealogy in each equivalence class. An alternative strategy would be to condition the GF on a particular class representative in the first instance. However, since generating the GF recursion is extremely fast, it seems that this is unlikely to bring

any substantial computational saving.

A related partitioning of the GF can be used to find approximations for models that include reversible events, in particular migration between populations and recombination between discrete loci. These strategies make it possible to solve the GF and derive the likelihood for surprisingly large samples and biologically interesting models. Despite this, full likelihood calculations will rarely be feasible for samples > 6 given the rapid increase in the number of equivalence classes and mutational configurations per class with sample size (Table 1). However, given that a separation of timescales exists for many models of geographic and genetic structure (Wakeley, 1998, 2009), full likelihood solutions for moderate ($n < 6$) samples may still be sufficient for computing likelihoods for much larger samples if these contribute mainly very short branches with no mutations in the scattering phase.

While our initial motivation for studying the GF for unrooted and unphased data was to deal with the absence of outgroup and phase information in practice, the results are encouraging more generally for two reasons. Firstly, and perhaps surprisingly, at least in the case of the IM model expressions for the total length of branches contributing to unphased and unpolarized mutation types are much simpler than those of the underlying rooted branches, which suggests that it may be possible to find general results. Secondly, combining branches connected to equivalent sets of leaf-nodes halves the number of mutation types: there are $n - 1$ polarized site frequencies, but $2(n - 1)$ genealogical branches, which in turn drastically reduces the number of possible mutational configurations. As Bunnefeld et al. *in press.* have shown for models of bottlenecks in a single population, this extension of the site frequency spectrum (SFS) to block-wise site frequency counts gives a promising strategy for summarising polymorphism information. Like the SFS, block-wise site frequency counts can be extended to multiple populations. For a sample comprised of n_i individuals from population i and assuming a global maximum number of mutations k_m for all frequency categories, there are $\prod_i((n_i - 1)^{(k_m+2)})$ possible mutational configurations defined in terms of block-wise

Figure 9: The (rooted) topology spectrum for a sample of $n = 6$ from a two population IM model with asymmetric migration and $T = 1.5$. Shown are topological probabilities of all 49 equivalence classes for a range of migration rates: $M = (0.2, 0.4, 0.8, 1.6)$ (bottom, up, left) and divergence times; $T = (0.5, 1, 1.5, 2)$ (top to bottom, right).



frequency counts.

Another obvious strategy to summarize block-wise data is to focus exclusively on the topology information contained in the presence and absence of diagnostic mutations. Setting all ω_h in the GF for a particular equivalence class to zero and multiplying with n_h, gives the total probability of this equivalence class. We can think of this set of probabilities as the "topology spectrum". Figure 9 shows the topology spectrum under the IM model for a relatively recent split ($T = 1.5$). In this case, most equivalence classes are extremely unlikely, e.g. the least likely class has probability < 0.0001. The most likely equivalence class consists of reciprocally monophyletic topologies, i.e. $(((a, a), a)), ((b, b), b))))$. As expected, the probability of reciprocal monophyly decreases with M and increases with T. In order to use the topology spectrum for inference, we would have to include the probability of seeing at least one mutation per branch. In practice, short blocks (small θ) rarely contain enough mutations to fully resolve their topology. However, the probabilities of unrooted and partially resolved topologies can be computed from the GF. Each unresolved node of the genealogy is a trifurcation, so the GF for a partially resolved genealogy is a sum over three equivalence classes.

In general, the GF framework makes it possible to derive the distribution of any summary statistics that can be defined as a combination of genealogical branches and understand its properties for simple models and small n. Although explicit likelihood calculations based on such summaries will not be feasible for large n, such summary statistics may still have wide applicability for fitting data to complex models, for example using composite likelihood or likelihood-free methods, or simply as a way to visualize how genealogies vary along the genome.

Acknowledgements

This work was supported by funding from the UK Natural Environment Research Council to KL (NE/I020288/1) and a grant from the European Research Council (250152) to NB.

References

Chen, H. (2012). The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theoretical Population Biology*, 81(2), 179 – 195. doi:http://dx.doi.org/10.1016/j.tpb.2011.11.004.

Davey, J.W. & Blaxter, M.L. (2011). RADseq: next-generation population genetics. *Briefings in Functional Genomics*, 9, 416–423. ISSN 1558-5646. doi:10.1093/bfgp/elq031.

Edwards, A.W.F. (1970). Estimation of the branch points of a branching diffusion process (with discussion). *J. R. Stat. Soc. B.*, 32, 155–174.

Excoffier, L., Dupanloup, I., Huerta-SÃąnchez, E., Sousa, V.C. & Foll, M. (2013). Robust demographic inference from genomic and snp data. *PLoS Genet*, 9(10), e1003905. doi:10.1371/journal.pgen.1003905.

Felsenstein, J. (1978). The number of evolutionary trees. *Molecular Phylogenetics and Evolution*, 27(1), 27–33.

Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annu Rev Genet*, 22, 521–565.

Felsenstein, J. (2003). *Inferring phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. & Pääbo, S. (2010). A draft sequence of the Neanderthal genome. *Science*, 328(5979), 710–722.

Griffiths, R. & Tavaré, S. (1998). The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models*, 14(1-2), 273–295. doi:10.1080/15326349808807471.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & D., B.C. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoSGenetics*, 5(10), e1000695.

Harris, K. & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*, 9(6), e1003521. doi:10.1371/journal.pgen.1003521.

Hearn, J., Stone, G.N., Bunnefeld, L., Nicholls, J.A., Barton, N.H. & Lohse, K. (2014). Likelihood-based inference of population history from low-coverage de novo genome assemblies. *Molecular Ecology*, 23(1), 198–211. ISSN 1365-294X. doi:10.1111/mec.12578.

Hey, J. & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 167(2), 747–760.

Hobolth, A., Andersen, L.N. & Mailund, T. (2011). On computing the coalescent time density in an isolation-with-migration model with few samples. *Genetics*, 187, 1241–1243.

Hudson, R.R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37, 203–217.

Keightley, P.D., Pinharanda, A., Ness, R.W., Simpson, F., Dasmahapatra, K.K., Mallet, J., Davey, J.W. & Jiggins, C.D. (2015). Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Molecular Biology and Evolution*, 32(1), 239–243. doi:10.1093/molbev/msu302.

Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and their Applications*, 13, 235–248.

Kronforst, M.R., Young, L.G., Blume, L.M. & Gilbert, L.E. (2006). Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. *Evolution*, 60(6), 1254–1268. ISSN 1558-5646. doi:10.1111/j.0014-3820.2006.tb01203.x.

Kronforst, M., Hansen, M., Crawford, N., Gallant, J., Zhang, W., Kulathinal, R., Kapan, D. & Mullen, S. (2013). Hybridization reveals the evolving genomic architecture of speciation. *Cell Reports*, 5(3), 666–677. doi:10.1016/j.celrep.2013.09.042.

Li, H. & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357), 493–6.

Lohse, K., Barton, N.H., Melika, N. & Stone, G.N. (2012). A likelihood-based comparison of population histories in a parasitoid guild. *Molecular Ecology*, 49(3), 832–842.

Lohse, K. & Frantz, L.A.F. (2014). Neandertal admixture in eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics*, 196(4), 1241–1251. doi:10.1534/genetics.114.162396.

Lohse, K., Harrison, R.J. & Barton, N.H. (2011). A general method for calculating likelihoods under the coalescent process. *Genetics*, 58(189), 977–987.

Mailund, T., Halager, A.E., Westergaard, M., Dutheil, J.Y., Munch, K., Andersen, L.N., Lunter, G., Püfer, K., Scally, A., Hobolth, A. & Schierup, M.H. (2012). A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genetics*, 8(12), e1003125. doi:10.1371/journal.pgen.1003125.

Mallet, J., Beltran, M., Neukirchen, W. & Linares, M. (2007). Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology*, 7(1), 28. ISSN 1471-2148. doi:10.1186/1471-2148-7-28.

Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., Blaxter, M., Manica, A., Mallet, J. & Jiggins, C.D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*.

Nee, S., Holmes, E.C., Rambaut, A. & Harvey, P.H. (1995). Inferring population history from molecular phylogenies. *Philosophical Transactions of the Royal Society of London Series B*, 349(25-31).

Pybus, O.G., Rambaut, A., Holmes, E.C. & Harvey, P.H. (2002). New inferences from tree shape: numbers of missing taxa and population growth rates. *Systematic Biology*, 51(6), 881–888.

Schiffels, S. & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8), 919 – 925.

Wakeley, J. (1998). Segregating sites in Wright's island model. *Theoretical Population Biology*, 53(2), 166–174.

Wakeley, J. (2009). *Coalescent theory*. Roberts & Company Publishers, Greenwood Village, Colorado.

Wakeley, J. & Hey, J. (1997). Estimating ancestral population parameters. *Genetics*, 145(3), 847–855.

Wang, Y. & Hey, J. (2010). Estimating divergence parameters with small samples from a large number of loci. *Genetics*, 184, 363–373.

Wilkinson-Herbots, H. (2012). The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. *Theoretical Population Biology*, 82, 92–108.

Wilkinson-Herbots, H.M. (2008). The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. *Theoretical Population Biology*, 73(2), 277–288.

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4), 1811–1823.

Zhu, T. & Yang, Z. (2012). Maximum likelihood implementation if an isolation-with-migration model with three species for testing speciation with gene flow. *Molecular Biology and Evolution*, 49(3), 832–842.