

# Hybridization as an invasion of the genome: Online Appendix 2.

## Multilocus methods for detecting hybridization and introgression between species

James Mallet

Hybridization is sometimes common enough to generate significant numbers of 'hybrids' (e.g.  $F_1$ ,  $F_2$ , backcrosses, etc.) in natural populations. If so, one may use multilocus genotypic information, typically provided by allozymes, microsatellites, AFLPs or SNPs scattered around the genome, to detect hybrids of different kinds (Table A2a). Within sexual, randomly mating populations, genotypes at unlinked genes are expected instantaneously to attain Hardy-Weinberg equilibrium, and rapidly to approach linkage equilibrium. In many cases, for example where there are fixed differences between species, hybrid genotypes are easy to detect by inspection [e.g. A2.1-2], or actual hybridizations can be observed directly in the field [A2.3]. Where the two species differ only moderately in frequency at the loci, more subtle techniques are used. Multilocus deviations from expected Hardy-Weinberg and linkage equilibria in natural populations allow 'partitioning' of genotypic samples into separate species or populations, and also sub-populations consisting of hybrids or intermediates [A2.4-5], within each of which unexpected Hardy-Weinberg and linkage disequilibria are minimized. This is usually done by estimating fractions of the populations consisting of hybrids using a Monte Carlo Markov Chain approach to explore maxima in Bayesian probability given the model of hybridization and data.

Hybrids are often so rare as to be virtually unsampleable in natural populations, but we might still suspect that alleles or haplotypes in one species come from related species by means of cryptic or occasional hybridization and introgression. Some recent methods employ coalescent theory to analyse this problem in samples of haplotype sequences from relevant species (Table A2b, Figure 1, main article).

Figure 1 (main article) displays some of the complexity inherent in even simple, neutral models of gene genealogies in three species. The true gene genealogy for a sample of non-recombining alleles (haplotypes) in each species (A-M) is shown as a red branching pattern, with neutral substitutions ('ticks of the clock') shown as rectangular boxes across the genealogy. Looking backwards in time, alleles found within a particular species may often coalesce after speciation if the effective population size is reasonably small, the time since speciation is reasonably long, and introgression is rare. Species 3 would have been such a species, with alleles L and M coalescing very recently, if introgression had not occurred from species 2. Hybridization can result in discordance between the gene genealogy and the species phylogeny, as in alleles H, I, J and K. However, it may often be the case if population sizes are large or the time since speciation is small that coalescence within species will precede speciation by a long time, even without hybridization. For instance, coalescence of the haplotypes sampled from species 1 (A-E) takes place way back in the ancestor of species 1, 2 and 3. Therefore phylogenetic discordance may result from 'ancestral polymorphism' as well as from introgression (haplotypes J and K). Distinguishing ancestral polymorphism from introgression is therefore tricky, because it is possible to explain 'foreign' alleles in a species by means of large effective population sizes and short times since speciation, in the absence of any other knowledge about these parameters.

As potential evidence for introgression, we typically only have samples of haplotype sequences from a small number of genes. The true phylogeny itself is uncertain, as are the parameters for population size ( $\theta$ ), time since lineage split, and introgression rates, required to understand the data. The topologies of the true genealogies are also unknown, and the pattern of substitutions among haplotypes, which is the only information we have on each genealogy, gives only a very incomplete picture. Most modern methods therefore use a Markov-Chain Monte-Carlo (MCMC) to sample the maxima of Bayesian probability or likelihood, given the neutral model and data, to infer a limited number of parameters (Table A2b). Such techniques can obtain estimates of the species parameters of interest by integrating over 'nuisance parameters' not of primary interest, such as those specifying estimated genealogies for each gene. Nonetheless, these methods are still in their infancy, and to answer the question: 'What is the level of introgression between a pair of species?' no method has yet been completely successful at incorporating all the relevant genealogical information, even for the simple three-species case given in Figure 1 (main article). In addition, the methods must assume very simple, neutral models. In nature, all parameters may vary, rather than being constant over time. In addition, natural selection may affect genealogies, recombination may occur; and many other complications are also likely. Our ability to detect introgression based on genealogical data, therefore, is very much in its infancy. Coalescent-based Bayesian analyses are, however, a growth area of research, and it seems likely that many of the current limitations can be overcome in the future.

### References

*Hybridization and invasion of the genome: Online Appendix 2 – J. Mallet*

- A2.1 Mallet, J. *et al.* (1998) Estimating the mating behavior of a pair of hybridizing *Heliconius* species in the wild, *Evolution* 52, 503–510
- A2.2 Pfennig, K.S. (2003) A test of alternative hypotheses for the evolution of reproductive isolation between spadefoot toads: support for the reinforcement hypothesis, *Evolution* 57, 2842–2851
- A2.3 Grant, P.R. *et al.* (2003) Inbreeding and interbreeding in Darwin's finches, *Evolution* 57, 2911–2916
- A2.4 Pritchard, J.K. *et al.* (2000) Inference of population structure using multilocus genotype data, *Genetics* 155, 945–959
- A2.5 Anderson, E.C. and Thompson, E.A. (2002) A model-based method for identifying species hybrids using multilocus genetic data, *Genetics* 160, 1217–1229
- A2.6 Hudson, R.R. *et al.* (1987) A test of neutral molecular evolution based on nucleotide data, *Genetics* 116, 153–159
- A2.7 Wang, R.L. *et al.* (1997) Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives, *Genetics* 147, 1091–1106
- A2.8 Beerli, P. and Felsenstein, J. (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach, *Proc. Natl. Acad. Sci., USA* 98, 4563–4568
- A2.9 Hey, J. and Nielsen, R. (2004) Multilocus methods for estimating population sizes, migration rates and divergence times, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*, *Genetics* 167, 747–760
- A2.10 Rannala, B. and Yang, Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci, *Genetics* 164, 1645–1656

Table A2 for Online Appendix 2.

Marker-based genetic methods useful or potentially useful for investigating hybridization and introgression between diverging populations or species

Programme or method and reference	Website	Optimization Method	Examples of restrictions
<b>a. Multilocus genotypic methods (e.g. allozymes, microsatellites, AFLP, SNP)</b>			
Structure [A2.4]	<a href="http://pritch.bsd.uchicago.edu/">http://pritch.bsd.uchicago.edu/</a>	MCMC/Bayesian	1. 'Admixture' populations are assumed to be in Hardy-Weinberg and linkage equilibrium and so does not deal with ongoing hybridization (nb: method is to perform population assignments, rather than intended to estimate the fraction of hybrids)
NewHybrids [A2.5]	<a href="http://ib.berkeley.edu/labs/slatkin/eriq/">http://ib.berkeley.edu/labs/slatkin/eriq/</a>	MCMC/Bayesian	1. Method is used to estimate the fractions of hybrids of arbitrary complexity in a single mixed population
<b>b. Coalescent DNA sequence-based methods</b>			
HKA test [A2.6]	Several implementations, e.g. <a href="http://lifesci.rutgers.edu/~heylab/">http://lifesci.rutgers.edu/~heylab/</a>	Based on analytical moments for levels of polymorphism and divergence times	1. Assumes ancestral $\theta_{12}$ is average of derived $\theta_1, \theta_2$ 2. Site-based, no use of full genealogical information 3. Only two species and an outgroup treated (nb: intended to test for selection, not introgression)
WH [A2.7]	<a href="http://lifesci.rutgers.edu/~heylab/">http://lifesci.rutgers.edu/~heylab/</a>	Coalescent simulation to obtain null hypothesis, fit is by least deviation	1. Infinite sites model (no homoplasy allowed) 2. Site-based, as above 3. Only a pair of species treated
Migrate, LAMARC [A2.8]	<a href="http://evolution.genetics.washington.edu/lamarc.html">http://evolution.genetics.washington.edu/lamarc.html</a>	MCMC/Likelihood	1. Takes no account of copy number 2. Assumes same migration at each locus 3. Assumes equilibrium between gene flow and drift
IM [A2.9]	<a href="http://lifesci.rutgers.edu/~heylab/">http://lifesci.rutgers.edu/~heylab/</a>	MCMC/Bayesian	1. Assumes no recombination 2. Only a pair of species/populations treated
MCMCCoal [A2.10]	<a href="http://abacus.gene.ucl.ac.uk/">http://abacus.gene.ucl.ac.uk/</a>	MCMC/Bayesian	1. Takes no account of substitution rate ( $\mu$ ) variation among loci 2. Takes no account of the effect of genomic copy number (e.g. mitochondria, sex chromosomes, autosomes) on $\theta$ 3. Recombination assumed absent 4. True phylogeny assumed known (nb: intended to estimate times of divergence and $\theta$ parameters in the absence of introgression)