8-2014

Linking DNA Polymorphisms and Populations' Evolutionary History

Ivan Juric
University of Tennessee - Knoxville, ijuric1@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Ivan Juric entitled "Linking DNA Polymorphisms and Populations' Evolutionary History." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Arts, with a major in Ecology and Evolutionary Biology.

Sergey Gavrilets, Major Professor

We have read this dissertation and recommend its acceptance:

Benjamin Fitzpatrick, Brian O'Meara, Arnold Saxton

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Linking DNA Polymorphisms and

Populations' Evolutionary History

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Ivan Juric

August 2014

To all my friends I met in Knoxville,

So Long, and Thanks for All the Fish

To my son Martin and his cows that say boo

Acknowledgements

First, I would like to thank my advisor, Dr. Sergey Gavrilets, for his help, support and guidance. I would also like to thank my committee, Drs. Brian O'Meara, Benjamin Fitzpatrick and Arnold Saxton for providing additional support and guidance. A big thank you goes to the members of Gavrilets lab and numerous friends I met while in grad school whose discussions and advice helped me a lot. I won't name you all, because if I try I'd forget someone, and that would be unfair. However, one person needs to be mentioned. Special thanks go to my wife, Katie Stuble, for supporting, encouraging and putting up with me during these years. Without your help, finishing my dissertation would have been much harder.

Abstract

This dissertation seeks to provide an understanding of how different evolutionary forces can affect the DNA polymorphism patterns. I use a combination of individual-based simulations and analytical to examine polymorphism patterns during divergence with gene flow, hybridization and territory expansion. In the first chapter, I show how during divergence with gene flow the appearance and maintenance of "Genomic Islands of Divergence" can be explained using standard population genetics terminology, thus removing some of the confusion recently introduced in that literature. In the second chapter I derive the expressions for the distribution of coalescent times and pairwise differences in a hybridization model with migration and show how those equations can be used to estimate model parameters. Finally, in third chapter, I consider the "Serial Founder" (SF) model. Previous work has shown that the SF model without migration can produce a pairwise F_{st} [fixation index] and heterozygosity patterns consistent to ones reported for human populations. Previous simulation results also suggest that including migration does not cause substantial departures from a model with no migration, but the lack of analytical result limits the ability to precisely describe the effects of migration on F_{st} and heterozygosity. I fill this void by showing analytically that a SF model with a historical migration can produce qualitatively different F_{st} and heterozygosity patterns from a model without migration, but not for parameters describing humans.

Table of Contents

List of Figures

Introduction

Genomes of all species are shaped by different evolutionary forces, such as mutation, recombination, population structure, random genetic drift and natural selection. Mathematical models and individual-based simulations provide a powerful and an invaluable tool to both guide our intuition as well as supplement and interpret empirical research. In this thesis I seek to understand and describe the connection between evolutionary forces and polymorphism patterns it produces across genomes.

First chapter deals with controversies regarding the existence of genetically diverged genomic regions, called “genomic islands of divergence”(GIDs). Multiple researchers have recently made claims regarding when, where and how GIDs appear and are maintained in the genome. We show that many GIDs features can be explained by previous theoretical work on barriers to the gene flow thus clarifying the role of natural selection, population size and recombination in creating and maintaining GIDs. Apart from that, we also point out to some unanswered questions regarding GIDs.

In the second chapter we use coalescent theory to study DNA polymorphism patterns produced by hybridization. During hybridization, individuals from two populations form a third, hybrid population. Detecting when hybridization happened, as well as estimating admixture coefficient is important for understanding genetic variation, as well as for conservation efforts. Our main result is the derivation of a closed-form analytical result for the distribution of coalescent times and pairwise differences in a

hybridization model that allows for the migration prior hybridization. Those results can be used as a foundation for developing methods for estimating hybridization time and admixture coefficients from genome scans.

In the final chapter, we study genomic patterns produced due to range expansion. We consider a "serial founder model" (SF) model in which a new population is formed by small number of migrants from adjacent one. SF model has been used to explain a general linear decrease in heterozygosity and increase in pairwise F_{st} as we sample populations farther from Africa. I expand a basic SF model to include historical migration and show that a model with historical migration can produce an increase in heterozygosity and decrease in F_{st} when basic SF model cannot. However, I also show that for parameters used to describe human conquest of the world, the model with migration produces very similar patterns as the model without migration, thus providing a theoretical justification to previous observation that migration might not affect the general patterns observed in the data.

Chapter 1

On genomic islands of divergence

Abstract

It is well established that divergent ecological selection in the presence of gene flow can result in the appearance of genomic islands of divergence (GIDs). Here, we illuminate the link between earlier and more recent work on GIDs. We use analytical approximation and individual-based simulations to show that the expected profiles of GIDs are well predicted by the standard population genetics theory. GIDs can be formed quickly and are stable in time rather than transient, but their features are subject to significant stochasticity. Our results suggest that that the presence of GIDs simplifies further divergence in weakly selected loci. We show that when one is using F_{ST} scans to compare GIDs in different species, larger GIDs do not necessarily imply stronger divergent selection.

Introduction

Lineages can diverge in spite of continuous gene flow if selection is strong enough and favors alternative alleles in different parts of the population's range (Allender et al., 2003, Schluter, 2009, Chapman et al., 2013, Gavrilets, 2004, Coyne and Orr, 2004, Price, 2007). When diverging lineages hybridize, gene introgression is less likely to occur near the loci subject to spatially variable selection. This causes heterogeneity in divergence levels across the genomes (Andolfatto, 2001, Nielsen, 2005, Storz, 2005, Turner et al., 2005, Harr, 2006, Hohenlohe et al., 2010, Ellegren et al., 2012, Martin et al., 2013).

To describe this heterogeneity, Turner et al. (2005) coined the metaphor "genomic islands of speciation" (also referred to as "genomic islands of divergence") in which highly diverged genomic regions stand above the regions of low divergence, like islands in a sea. Genomic islands of divergence (GIDs) have since received a great deal of attention and were recently declared a "metaphoric foundation on which the study of genomic architecture is currently based" (Nosil and Feder, 2012).

The presence of GIDs has been interpreted as evidence of local adaptation and/or ongoing ecological speciation (Via and West, 2008, Feder and Nosil, 2010). GIDs can be used to delimit locally adapted populations and potentially be used to improve conservation efforts for commercially important and exploited species (Bradbury et al., 2013). The size and the distribution of GIDs might help us understand the underlying genomic architecture (i.e., number and distribution of selected genes) of speciating populations (Nosil and Feder, 2012, Seehausen et al., 2014). For example, some theoretical work (Gavrilets et al., 2007, Gavrilets and Vose, 2007) argues that speciation happens the easiest if the number of loci controlling selected traits is small. However, empirical data suggest that the number and distribution of GIDs vary greatly during early stages of speciation (Turner et al., 2005, Via and West, 2008, Wood et al., 2008, Hohenlohe et al., 2010, Michel et al., 2010, Martin et al., 2013, Wang et al., 2014). GIDs have been argued to be a place where additional selected loci can diverge more easily resulting in clustering of selected genes in the genome (Via and West, 2008, Feder and Nosil, 2010). This however has been disputed recently on the basis of the results of simulations studies (Feder et al., 2012b, Yeaman, 2013).

Several verbal models have been proposed to explain how GIDs form and evolve during speciation (Wu, 2001, Via and West, 2008, Via, 2009, Feder et al., 2012a). In Wu's seminal paper (2001) on the genic view of speciation, a four-stage speciation model is introduced. Wu starts by recognizing two classes of loci: "speciation genes" (i.e., the loci that directly affect differential adaptation) and "marker genes" (i.e., all other loci such as allozymes, microsatellites, mitochondrial DNA, etc). During the first stage, gene flow across the genome is mostly unrestricted, with some reduction being limited to marker genes tightly linked to speciation genes of strong effect. Wu assumes that the number of loci causing reproductive isolation grows over time as new alleles arise by mutation. This decreases the gene flow (and increases divergence) at marker genes close to selected loci while the genome remains more "porous" at marker loci that are far from speciation genes (stage II). Eventually, the gene flow between two populations becomes very small (stage III) and then stops altogether (stage IV), at which point speciation is complete.

Several years after the publication of Wu (2001), Via (2009) proposed a two-stage model of ecological speciation with gene flow. During the first stage, the loci under strong selection diverge quickly. When this happens, the probability that a migrant survives in a new environment, mates with a resident, and produces a hybrid offspring decreases.

This causes an increase in the size of a "hitchhiking region" and enables loci of smaller effects to diverge among populations. In Via's model, a "hitchhiking region" is a part of the genome in which gene flow is substantially reduced due to the presence of

selected genes. At the end of the first stage, gene flow is mostly ceased, and the two populations evolve as if they are allopatric. During the second stage, additional postzygotic incompatibilities, such as Dobzhansky-Muller genetic incompatibilities (Dobzhansky, 1937, Muller, 1942, Gavrilets, 2004, Coyne and Orr, 2004) may accumulate in the genome. Similar to Wu (2001), Via (2009) predicts a "genetic mosaic of speciation", i.e. that some genomic regions will be more diverged than others. Via (2009) built on the "divergence hitchhiking" (DH) mechanism proposed in Via and West (2008). According to "divergence hitchhiking", the loci under divergent selection reduce successful interbreeding between subpopulations. Also reduced is the opportunity for recombination between chromosomes from different populations with the reduction being stronger for loci that are closer to selected loci. The "effective" recombination rate around loci under divergent selection is smaller than the rate based on physical distance and, in words of Via and West (2008), the populations become "protected from interrace recombination around loci under divergent selection during early speciation".

More recently, Feder et al. (2012a) proposed another four-phase model of speciation with gene flow. During the first phase, genetic divergence is mostly limited to loci experiencing direct selection. In the second phase, loci tightly linked to selected loci diverge due to a reduction in gene flow. In this model, divergence at linked loci is due to DH, which the authors define as "a process in which divergent selection on a locus can reduce the effective migration rate for physically linked gene regions and increase divergence in the surrounding region". During the third phase, multiple loci in the genome have diverged and effective migration rate is reduced across the whole genome.

At this point, genome-wide divergence is mostly due to “genome hitchhiking” (GH) which the authors define as the “process in which divergent selection reduces the average effective migration rate globally across the genome fostering increased divergence genome-wide” (Feder et al., 2012a). Finally, in the fourth phase, the genomes of the two species are highly diverged and introgression is greatly reduced.

While one can welcome new metaphors such as GIDs, “porous genome”, and “genetic mosaic of speciation” because they help us train our intuition about speciation process, new terminology can also introduce a lot of confusion into the field especially if the connection with earlier approaches is not clearly explained. For example, a number of recent publications treat divergence hitchhiking and genome hitchhiking as two processes whose relative importance needs to be studied (Feder et al., 2012b,a, Flaxman et al., 2013, Kronforst et al., 2013, Nosil and Feder, 2013).

These “hitchhiking” processes are also sometimes presented as something different from standard population-genetic descriptions of genetic divergence in the presence of gene flow (Via, 2012). As we show below, despite using a different vocabulary, all these verbal models actually describe the same process known from earlier studies by Barton, Bengtsson, and others (Barton, 1979b,a, Barton and Hewitt, 1983, Spirito et al., 1983a, Barton and Hewitt, 1985, Bengtsson, 1985, Barton and Bengtsson, 1986, Spirito, 1987, 1989, Barton and Bengtsson, 1986, Gavrilets, 1997, Gavrilets and Cruzan, 1998) as the evolution of genetic barriers to gene flow. The main insight from this earlier work, which we illustrate in the next section, is that selection on

some loci can serve as a barrier to gene flow at neutral loci, linked or not, to the loci under selection.

Our primary goal here is to illuminate the link between earlier and more recent work on GIDs and clarify their role in genetic divergence. We show that one can explain how GIDs evolve using well-established population genetics vocabulary of selection, migration, recombination, population size, and initial conditions. To that end, we use a combination of analytical approximations and individual-based simulations.

Results

Genetic barriers to gene flow and gene flow factor

We will illustrate the general approach using a simple model of a sexual diploid population with discrete nonoverlapping generations inhabiting two demes connected by migration. Each deme has effective size N. We focus on diallelic loci subject to symmetric mutation at rate μ. We assume adult migration (the probability that an individual moves from a deme where he was born) at rate m happening before mating which is random within the deme.

Neutral divergence and Fst .

If there is no selection, the population will reach a state of stochastic balance between mutation, migration, and random genetic drift in which individuals sampled from

different demes will, on average, be more different genetically than those sampled from the same deme. This effect can be described quantitatively by a coefficient F_{st} , defined as the correlation between gametes chosen randomly within demes relative to that between gametes chosen randomly from the whole population (Wright, 1969, p.294). For a diallelic locus, this is equivalent to an intraclass correlation coefficient: $F_{st} = \sigma_b^2/(\bar{p}(1-\bar{p})$, where σ_b^2 is the variance in allele frequency among demes and $\bar{p}$ is the mean allele frequency across the demes (Fu et al., 2003). In the model under consideration, the expected value of F_{st} is:

$$F_{st} = \frac{1}{1 + 16N(m + \mu)} \tag{1.1}$$

(Cockerham and Weir, 1987). This equilibrium is achieved very rapidly with a characteristic half-time being on the order of $1 / (2m+1/(2N) + \mu)$ (Crow and Aoki, 1984).

Gene flow factor.

Assume that the two demes are subject to spatially heterogeneous viability selection and have diverged in some selected loci. Now neutral alleles brought by immigrants will have a reduced probability of being incorporated in a local deme because initially they will typically be associated with locally deleterious selected alleles.

There are several ways to characterize this effect quantitatively (Barton, 1979b,a, Barton and Hewitt, 1983, Petry, 1983, Spirito et al., 1983b, Kobayashi et al., 2008, Fusco and

Uyenoyama, 2011b). The most intuitive is arguably Bengtsson's (1985) "gene flow factor" defined as the probability g that a neutral allele brought by immigrants is

incorporated in a local genetic background. Assume first that viability is controlled by a single diallelic locus with alleles **A** and **a**. Let allele **A** be advantageous in the focal deme and allele **a** in the other deme. If migration rate *m* is small, then most local genotypes will be homozygotes **AA** while most immigrants will be homozygotes **aa**. Assume that fitness of heterozygotes relative to that of the locally adaptive homozygotes is $v < 1$. Consider a diallelic neutral locus which can be linked or unlinked to the selected locus with the probability of recombination between the two loci being r ($0 < r < 0.5$). Then the gene flow factor is **Error! Bookmark not defined.**

$$\gamma = \frac{vr}{1 - v(1 - r)}$$

(1.2a)

(Bengtsson, 1985). Note that γ decreases as *v* becomes small (i.e. selection against heterozygotes increases) or *r* becomes small (i.e., the neutral locus gets more tightly linked to the selected locus). If the loci are unlinked ($r = 1/2$), the above expression reduces to

$$\gamma = \frac{v}{2 - v}$$

(1.2b)

Therefore γ can be very small even for an unlinked neutral locus provided selection is strong enough (i.e. v is small).

Genetic barriers to gene flow and gene flow factors were investigated in a number of different models including those with multiple selected genes, other fitness components such as fertility and mating success, and an unequal sex ratio (Barton and Hewitt, 1983, Barton and Bengtsson, 1986, Gavrilets, 1997, Gavrilets and Cruzan, 1998, Kobayashi et al., 2008, Kobayashi and Telschow, 2011, Fusco and Uyenoyama, 2011b,a). For example, assume that there are a number of unlinked loci interacting multiplicatively so that the fitness of the F_1 hybrid between the locally advantageous genotype and an immigrant is $v = \prod_i (1 - s_i)$ where s_i is the selection strength for locus i. Then the gene flow factor for a neutral locus unlinked to any selected locus is

$$\gamma = \prod_i \frac{(1 - s_i)}{(1 + s_i)} \approx v^2 \tag{1.2c}$$

where the approximation assumes that each individual s_i value is small (Bengtsson, 1985). Note that what matters here is the overall strength of selection against heterozygotes/hybrids characterized by parameter v rather than the strength of selection on each individual locus s_i. Note also that v^2 is the fitness of the least fit genotype (i.e. the homozygote with locally deleterious alleles at all loci).

Gene flow factor and F_{st} . A gene flow factor γ less than one implies that the neutral locus effectively experiences a reduced migration at rate $m_e = m\gamma$. If selection is sufficiently strong (i.e., v is small), γ will be small and the effective migration rate me can be very small even for neutral genes unlinked to the selected locus. In a sense, divergence in selected loci acts as a barrier to neutral gene flow (Barton, 1979b). Charlesworth et al.

(1997) used Bengtsson's result to approximate F_{st} in the presence of a genetic barrier to gene flow by substituting m for m_e in equation (1.1):

$$F_{st} = \frac{1}{1 + 16N(m\gamma + \mu)} \tag{1.3}$$

Since γ decreases with proximity to the selected locus, F_{st} at neutral loci close to the selected locus will be, on average, greater than that of more distant loci, and a "genomic island of divergence" will emerge. Via, Nosil, Feder, and their co-authors used Charlesworth et al. (1997) results to build their respective arguments.

Expected F_{st}.

As equations (1.1-1.3) show, the characteristics of GIDs depend on selection strength, migration, mutation rates as well as the population size. The results are qualitatively the same when multiple loci are under divergent selection, but the equations are more complicated (see the Appendix 1.1). To check the performance of analytical approximations, we computed F_{st} using individual-based simulations of the two-deme model with one, two, and three selected loci. We used multiplicative selection, keeping the fitness of the least fit genotype (i.e. the homozygous individual with locally deleterious alleles at every locus) the same regardless of the number of selected loci (see

the Appendix 1.1). By choosing such fitness scaling, we kept the range of fitness values independent of the number of selected loci. The fit between the analytical prediction and simulation results was generally good (Figures 1–3) with the fit being the best for intermediate selection strength. For weak selection (s = 0.1), analytical results overestimated the simulation results for neutral loci very close to selected locus. For very strong selection (s = 0.9) analytical results underestimate the level of divergence. A likely reason for this discrepancy is the violation of the assumption that local individuals are homozygous for advantageous alleles.

Variation in F_{st}.

Analytical methods predict the expected values of F_{st} . Whenever one measures F_{st} from empirical data, one expects stochastic deviations from the expectations. To study the variation in F_{st} , we computed its standard deviation numerically. Figure 1-4 shows that variation in F_{st} at each site is considerable and closely mimics the expectation of F_{st} . That is, the variation in F_{st} is highest at the neutral loci close to the selected loci and lowest at the loci from these selected loci. This means that while neutral loci close to the selected loci will on average have higher F_{st} values, we expect some to have low F_{st}. The reason for high variance at those loci is the inability of migration to homogenize the population due to a strong reduction of the effective migration rate me. If selection is strong enough, neutral loci close to selected loci evolve independently in two populations, and the dynamics of allele frequencies are influenced by drift and mutation, and not by migration. Via and West (2008) observed that in pea aphids some neutral markers situated close to

selected loci had low F_{st} values and suggested that this effect was due to ancestral polymorphisms. Our results offer different interpretation of their observation.

Dynamics of the size of GIDs.

Formation of GIDs takes time. To study how GIDs change in time, we used two different initial conditions: "the population split" (PS) and "the secondary contact" (SC). In the PS simulations, for the first 20,000 generations migration between two demes is unrestricted (m = 0.5) and both demes experience selection favoring the same alleles. 20,000 generations are enough for the population to reach a state of stochastic equilibrium between mutation, selection, and drift. At generation 20,000, migration is decreased to a specific rate m and selection in one deme changes to favor alternative alleles. In the SC simulations, initially there are two isolated populations (m = 0), with selection favoring different alleles in different demes. At generation 20,000, migration is increased to a specific rate m. For both the PS and the SC simulations, we consider up to three selected genes placed uniformly across the chromosome, keeping the total strength of selection the same (see above). We defined the GID size as the length of a chromosome region(s) that has F_{st} five or more times larger than that occurring in simulations with no selection (see the Appendix 1.1). In our simulations, the GID size often reaches a stochastic equilibrium in a couple of

thousand generations (Figure 1-5). When selection is strong, in small populations, GID size reaches a steady state more slowly during the PS scenario than in the SC scenario, but in large populations the differences in the time to reach an equilibrium are minimal. As expected, the GID size increases with increasing strength of selection and decreasing

migration rate. The size of GIDs increases with the number of loci (Figure 1-6). This happens because more neutral markers are close to selected loci when more loci are under selection.

Divergence in weakly selected loci.

It has been suggested that new selected alleles can establish more easily if they are close to a selected locus that has already diverged. That is, GIDs can serve as a place where the new loci experiencing divergent selection accumulate (Smadja et al., 2008, Via and West, 2008, Feder and Nosil, 2010).

To test this idea, we performed additional simulations. We modeled a single locus under strong divergent selection (with selection coefficient s_M) and eight loci under weak divergent selection (with selection coefficient s_m). The major locus was in the middle of the chromosome and the minor loci were uniformly spaced across the chromosome. Initial conditions were similar to the ”population split“ scenario. Mutation rate was set to $\mu = 10{-}4$ and deme sizes $N = 1000$. We compared F_{st} in three different cases: 1) both the minor and major loci are under selection ($s_M > 0$, $s_m > 0$), 2) only the minor loci are under selection ($s_M = 0$, $s_m > 0$), and 3) only the major locus is under selection ($s_M > 0$, $s_m = 0$). If the presence of a major locus is important for divergence at minor loci, F_{st} at minor loci in the first case should be significantly larger than in the other two cases. In our simulations, the minor loci diverged only when the major locus was under selection (Figure I.7). When divergence occurred, allele frequencies at minor loci reached an

equilibrium value within 10,000 generations from the onset of divergent selection. As expected, the F_{st} value at minor loci increases with increasing the strength of selection on the major locus and minor loci, and decreasing the recombinational distance from the major locus. For example, when s_M =0.9 and s_m =0.01, F_{st} at minor loci at distance $r \approx 0.3$ was 0.55, compared to approximately 0.1 for neutral loci at the same distance. In contrast, F_{st} stays close to zero when the major locus was not under selection. In this case, a minor locus would be considered an F_{st} outlier in genome scans when the major locus is under selection, but not in other cases. These results demonstrate that under some conditions major loci can indeed affect the divergence at nearby minor loci. However, if selection is very week and/or the major and minor loci are distant enough (e.g. the top row in Figure 1-7), it will be hard to distinguish minor selected loci from neutral ones on the basis of F_{st}. We simulated only one population size. Increasing the population size should lead to population divergence even at very weakly selected loci (provided the migration rate is small enough).

Discussion

When populations are subject to divergent selection and gene flow, comparing genomes of individuals from different demes might reveal GIDs. Here, we have shown that features of GIDs can be explained using standard methods of population genetics and a well-established terminology, and does not require invoking new mechanisms, such as divergence hitchhiking (DH) or genome hitchhiking (GH). Below we comment on a number of additional related issues.

Is it really "hitchhiking"?

In population genetics the term hitchhiking was originally introduced to describe the effects of the substitution of a favorable mutation on linked loci (Smith and Haigh, 1974). This term typically implies 1) an important role of the physical linkage of genes, and 2) temporary/short-lived effects. For a reduction in the effective migration rate and an increase in F_{st} described above to occur whether or not the genes are physically linked is of secondary importance (compare eq. 2a and 2b). The predicted increase in F_{st} values is not transient but stable in time and represents a feature of the resulting migration-selection-mutation-drift equilibrium. Although some authors used the term hitchhiking more generally (e.g. to describe the "indirect effects of selection at one or more loci on the rest of the genome" (Barton, 2000)) in the case considered here this would not be justified. While initial hitchhiking of neural genes linked to selected loci might help the formation of GIDs, GIDs are also formed because of new mutations occurring after the onset of divergence (Figure I.5). In fact, the long term maintenance of GIDs occurs not because some neutral alleles quickly hitch a ride to high frequencies but on the contrary because neutral alleles carried by immigrants get bumped off of the ride by selection. Therefore the term "hitchhiking" is not really appropriate here. When one observes GIDs in genome scans, it is hardly possible to know whether they are due to initially segregating alleles hitchhiking to high frequencies or new mutations not able to overcome the genetic barrier.

"Divergence hitchhiking" vs. "multilocus migration/selection balance".

In a recent review, Via (2012) argued that DH and multilocus migration/selection balance represent "alternative visions of genomic divergence during speciation-with-gene-flow." We think this dichotomy is misleading. "Divergence hitchhiking" as a process of the formation of GIDs is a component of a multilocus migration-selection-mutation-drift balance. Via et al. (2012) and Via (2012) claimed that MM/SB is a mechanism which produces multiple small GIDs across the genome, while DH produces large GIDs. Our results show that contrary to these expectations, the GID size is the largest when the population has reached a migration-selection-mutation-drift balance (solid lines, Figure I.5). This happens due to new mutations accumulating after the onset of divergent selection. These mutations were not considered by Via.

Effects of the population size and migration rate on F_{st} .

One of the reasons stimulating Feder and Nosil (2010) to introduce the new term "genome hitchhiking" was the results of their large-scale numerical simulations that suggested that divergence hitchhiking cannot work in large populations subject to high migration. The effects of the deme size N and migration rate m on F_{st} can be easily evaluated from equation (1.3). Indeed increasing N and m will dramatically decrease F_{st} . However this equation also shows that these effects can be largely offset by decreasing the gene flow factor g which can be accomplished by increasing selection against hybrids.

"Divergence hitchhiking" vs. "genome hitchhiking".

Feder et al. (2012a) define GH as the "process in which divergent selection reduces the average effective migration rate globally across the genome fostering increased divergence genome-wise" (glossary, p324). Quantifying the contributions of DH and GH has been presented as the next important step in the field of research on divergence-with-gene flow Feder et al. (2012a). That is because those authors view DH and GH as acting during different stages of speciation. A similar sentiment is seen in a more recent review by Seehausen et al. (2014) who, when discussing GIDs say "The size of these regions would gradually increase through the process of divergence hitchhiking, and the effective migration rate would eventually decrease globally across the genome, which gives rise to genome-wide divergence (that is, genomic hitchhiking)".

Equation (2a) shows that gene flow factor g can be significantly reduced if either the neutral locus is very close to the selected locus (i.e., r is small) or selection against hybrids is very strong (i.e., v is small). The difference between DH and GH is that physical linkage of genes is necessary in the former but is irrelevant in the latter (provided selection is very strong). As should be clear from our previous discussion , any effects of linkage on GIDs are quantitative and not qualitative. Therefore treating DH and DG as different mechanisms of divergence acting during different stages of speciation is not justified. The real evolutionary mechanisms underlying the formation of GIDs are selection and linkage.

Feder and Nosil (2010) observed significant differences between the behavior of one-locus models of DH in which F_{st} was observed to increase only at short distances

from the selected locus and multi-locus models of GH in which gene flow was decreased across the whole genome. However, as realized already by Bengtsson (1985), “an increase in the number of factors building the genetic barrier does not - by itself - particularly influence the gene flow factor...” (p.36). The real reason for the differences between models studied by Feder and Nosil (2010) were vast differences in the strength of selection assumed in their models of DH and GH. Feder and Nosil (2010) used multiplicative selection (as assumed in eq. 2c), fixed the strength of selection s_i per locus, and then studied the effects of increasing the number of loci L. For example, in their approach a single selected locus model of DH with a 50% reduction of the fitness v of hybrids ($v = 1-s$ with $s = 0.5$) would be compared with a 10-locus model of GH with 1000-fold reduction ($v = (1-s)^L = 1/210$ with $s = 0.5$ and $L = 10$) of hybrid fitness. Therefore the comparisons of DH and GH performed by Feder and Nosil (2010) are not appropriate because they confound variation in number of selected loci with variation in the strength of selection.

A couple of other comments are in order. When interpreting results from different empirical studies one needs to be aware that larger GIDs do not necessarily imply stronger selection or reduced gene flow even if the same method for detecting GIDs is used in all studies. Because F_{st} depends inversely on the population size (see eq. 1.1), so does the cutoff value for identifying F_{st} outliers. Therefore, all else being equal, larger portions of the genome will have F_{st} above the cutoff in large populations compared to small ones. Figure 1-5 illustrates this effect. Large GIDs can also be a consequence of spatial subdivision, rather than the effects of selection. This can happen because the

genomic patterns of neutral variation depend on spatial subdivision. For example, F_{st} outliers have often been found in river organisms, but a recent simulation study showed that tools used for analyzing genomic data to detect F_{st} outliers have high false positive rate when population is spatially subdivided in river-like environments (Bierne et al., 2013, Fourcade et al., 2013). The reason is that in river-like environments, the variance of F_{st} is inflated compared to island models, due to strong correlation in co-ancestry between sampled individuals. Using a model that takes into account population subdivision to infer departure of F_{st} from neutrality can help to alleviate this problem.

Lastly, the existence of GIDs is not a condition for ongoing speciation as often mentioned in the literature. For example, polyploid speciation (Ramsey and Schemske, 1998) can produce different species with genomes that are not diverged, while a secondary contact after prolonged isolation can create diverged populations belonging to the same species. A number of conditions must be satisfied for local adaptation and genetic divergence to actually lead to speciation (Coyne and Orr, 2004, Gavrilets, 2004,Wolf et al., 2010, Butlin, 2012, Smajda and Butlin, 2011). Gavrilets (2004, Chap.4-5) explicitly studied how the gene flow factor affects the expected time to speciation in several models.

Many issues related to genomic patterns during population divergence and speciation remain open (Seehausen et al., 2014). We will just point to two issues of major interest. First, we still do not have an analytical theory describing the transient dynamics of GIDs even in simple models such as ones considered in this paper. In our individual-based simulations, the time span for GIDs to reach an equilibrium is on the order of

population size. However, lowering the mutation rate is expected to increase the time required for the size of GIDs to reach an equilibrium. Second, we do not have a full understanding of the effects of GIDs on non-neutral divergence. Recently, Yeaman and Otto (2011) found an expression for the ”probability of establishment” in a two-deme population, i.e., the probability that a mutant allele reaches a high frequency in a deme where it is favored. Their results were used by Feder et al. (2012b) to study the probability of establishment of new selected mutations linked to already diverged selected genes. They concluded that the selection coefficient of a new mutation is a more important predictor of the establishment of a new allele rather than its proximity to an already diverged locus. However, the “establishment” of an allele in a deme does not necessarily mean divergence between the demes. With recurrent mutation and weak selection, a locus in the two-deme model can be polymorphic in both populations (and thus “established”), but not diverged. Our numerical results show that divergence at minor loci can be substantial and rapid if they are close enough to an already diverged locus under strong enough selection. Having a better understanding of dynamics of GIDs and an expression for a critical migration rate at which populations diverge would be a valuable addition to the field.

References

Allender, C. J., Seehausen, O., Knight, M. E., Turner, G. F., and Maclean, N. (2003). Divergent selection during speciation of lake Malawi cichlid fishes inferred from parallel radiations in nuptial coloration. Proceedings of the National Academy of Sciences, 100(24):14074–14079.

Andolfatto, P. (2001). Adaptive hitchhiking effects on genome variability. Current Opinion in Genetics & Development, 11(6):635 – 641.

Barton, N. and Hewitt, G. (1983). Protein Polymorphism: Adaptive and Taxonomic Significance, chapter Hybrid zones as barriers to gene flow, pages 341–359. Oxford; Blackwells.

Barton, N. and Hewitt, G. (1985). Analysis of hybrid zones. Annual Review of Ecology and Systematics, 16:113–148.

Barton, N. H. (1979a). The dynamics of hybrid zones. Heredity, 43(43):341–359.

Barton, N. H. (1979b). Gene flow past a cline. Heredity, 43(43):333–339.

Barton, N. H. (2000). Genetic hitchhiking. Philosophical Transactions: Biological Sciences, 355(1403):1553–1562.

Barton, N. H. and Bengtsson, B. (1986). The barrier to genetic exchange between hybridizing populations. Heredity, 57(Part 3):357–376.

Bengtsson, B. (1985). Evolution. Essays in Honour of John Maynard Smith, chapter The flow of genes through a genetic barrier., pages 31–42. Cambridge University Press; Cambridge; UK.

Bierne, N., Roze, D., and Welch, J. J. (2013). Pervasive selection or is it...? why are fst outliers sometimes so frequent? Molecular Ecology, 22(8).

Bradbury, I. R., Hubert, S., Higgins, B., Bowman, S., Borza, T., Paterson, I. G., Snelgrove, P. V. R., Morris, C. J., Gregory, R. S., Hardie, D., Hutchings, J. A., Ruzzante, D. E., Taggart, C. T., and Bentzen, P. (2013). Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. Evolutionary Applications, 6(3):450–461.

Butlin, R. (2012). What do we need to know about speciation? Trends in Ecology & Evolution, 27(1):27 – 39.

Chapman,M. A., Hiscock, S. J., and Filatov, D. A. (2013). Genomic divergence during speciation driven by adaptation to altitude. Molecular Biology and Evolution, 30(12):2553–2567.

Charlesworth, B., M., N., and D., C. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetical Research, 70:155–174.

Cockerham, C. C. and Weir, B. S. (1987). Correlations, descent measures: Drift with migration and mutation. Proceedings of the National Academy of Sciences of the United States of America, 84(23):8512–8514.

Coyne, J. and Orr, H. (2004). Speciation. Sinauer Associates, Inc, Sunderland.

Crow, J. F. and Aoki, K. (1984). Group selection for a polygenic behavioral trait: Estimating the degree of population subdivision. Proceedings of the National Academy of Sciences of the United States of America, 81(19):6073–6077.

Dobzhansky, T. (1937). Genetics and the Origin of Species. Columbia University, New York.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backstrom, N., Kawakami, T., Kunstner, A., Makinen, H.,Nadachowska-Brzyska, K., Qvarnstrom, A., Uebbing, S., and Wolf, J. B. W. (2012). The genomic landscape of species divergence in Ficedula flycatchers. Nature, 491(7426):756–760.

Feder, J. L., Egan, S. P., and Nosil, P. (2012a). The genomics of speciation-with-gene-flow. Trends in Genetics,28(7):342–350.

Feder, J. L., Gejji, R., Yeaman, S., and Nosil, P. (2012b). Establishment of new mutations under divergence and genome hitchhiking. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1587):461–474.

Feder, J. L. and Nosil, P. (2010). The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution, 64(6):1729–1747.

Flaxman, S. M., Feder, J. L., and Nosil, P. (2013). Genetic hitchhiking and the dynamic buildup of genomic divergence during speciaiton with gene flow. Evolution, 67(9):2577–2591.

Fourcade, Y., Chaput-Bardy, A., Secondi, J., Fleurant, C., and Lemaire, C. (2013). Is local selection so widespread in river organisms? fractal geometry of river networks leads to high bias in outlier detection. Molecular Ecology, 22(8):2065–2073.

Fu, R., Gelfand, A. E., and Holsinger, K. E. (2003). Exact moment calculations for genetic models with migration, mutation, and drift. Theoretical Population Biology, 63(3):231 – 243.

Fusco, D. and Uyenoyama, M. K. (2011a). Effects of polymorphism for locally adapted genes on rates of neutral introgression in structured populations. Theoretical Population Biology, 80(2):121 – 131.

Fusco, D. and Uyenoyama, M. K. (2011b). Sex-specific incompatibility generates locus-specific rates of introgression between species. Genetics, 189(1):267–288.

Gavrilets, S. (1997). Hybrid zones with Dobzhansky-type epistatic selection. Evolution, 51(4):1027–1035.

Gavrilets, S. (2004). Fitness landscapes and the origin of species. Princeton University Press; Princeton & London.

Gavrilets, S. and Cruzan, M. B. (1998). Neutral gene flow across single locus clines. Evolution, 52(5):1277–1284.

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. Genome Research, 16:730–737.

Hohenlohe, P., Bassham, S., Etter, P., Stiffler, N., Johnson, E., and et.al (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. PLoS Genetics, 6:e1000862.

Kobayashi, Y., Hammerstein, P., and Telschow, A. (2008). The neutral effective migration rate in a mainland-island context. Theoretical Population Biology, 74(1):84 - 92.

Kobayashi, Y. and Telschow, A. (2011). The concept of effective recombination rate and its application in speciation theory. Evolution, 65(3):617–628.

Kronforst, M. R., Hansen, M. E. B., Crawford, N. G., Gallant, J. R., Zhang, W., Kulathinal, R. J., Kapan, D. D., and Mullen, S. (2013). Hybridization reveals the evolving genomic architecture of speciation. Cell Reports, 5(3):666–677.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in heliconius butterflies. Genome Research, 23(11):1817–1828.

Michel, A. P., Sim, S., Powell, T. H. Q., Taylor, M. S., Nosil, P., and Feder, J. L. (2010). Widespread genomic divergence during sympatric speciation. Proceedings of the National Academy of Sciences, 107(21):9724–9729.

Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. Biology Symposium, 6:71–125.

Nielsen, R. (2005). Molecular signatures of natural selection. Annual Review of Genetics, 39:197–218.

Nosil, P. and Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1587):332–342.

Nosil, P. and Feder, J. L. (2013). Genome evolution and speciation: Toward quantitative descriptions of pattern and process. Evolution, 67(9):2461–2467.

Petry, D. (1983). The effect on neutral gene flow of selection at a linked locus. Theoretical population biology, 23(3):300–313.

Price, T. (2007). Speciation in birds. Roberts & Company Publishers, Greenwood Village, Colorado.

Ramsey, J. and Schemske, D. (1998). Pathways, mechanisms, and rates of polyploidy formation in flowering plants. Annual Review Of Ecology and Systematics, 29:467–501.

Roesti, M., Hendry, A. P., Salzburger, W., and Berner, D. (2012a). Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. Molecular Ecology, 21(12):2852–2862.

Roesti, M., Salzburger, W., and Berner, D. (2012b). Uninformative polymorphisms bias genome scans for signatures of selection. BMC Evolutionary Biology, 12(1):94.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. Science, 323(5915):737–741.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Braennstroem, A., Brelsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., Lindholm, A. K., Lucek, K., Maan, M. E., Marques, D. A.,

Martin, S. H., Matthews, B., Meier, J. I., Moest, M., Nachman, M. W., Nonaka, E., Rennison, D. J., Schwarzer, J., Watson, E. T., Westram, A. M., and Widmer, A. (2014). Genomics and the origin of species. Nature Reviews Genetics, 15(3):176–192.

Smadja, C., Galindo, J., and Butlin, R. (2008). Hitching a lift on the road to speciation. Molecular Ecology, 17(19):4177–4180.

Smajda, C. M. and Butlin, R. K. (2011). A framework for comparing processes of speciation in the presence of gene flow. Molecular Ecology, 20(24):5123–5140.

Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genetical Research, 23(1):23–35.

Spirito, F. (1987). The reduction of gene exchange due to a prezygotic isolating mechanism with monogenic inheritance. Theoretical Population Biology, 32(2):216 – 239.

Spirito, F. (1989). Neutral gene flow in the presence of a selected gene with random or assortative mating. Theoretical Population Biology, 35(3):295 – 306.

Spirito, F., Rossi, C., and Rizzoni, M. (1983a). Reduction of gene flow due to the partial sterility of heterozygotes for a chromosome mutation. I. Studies on a ’neutral’ gene not

linked to the chromosome mutation in a two population model. Evolution, 37(4):785–797.

Spirito, F., Rossi, C., and Rizzoni, M. (1983b). Reduction of gene flow due to the partial sterility of heterozygotes for a chromosome mutation. I. Studies on a ’neutral’ gene not linked to the chromosome mutation in a two population model. Evolution, 37(4):785–797.

Storz, J. F. (2005). Invited review: Using genome scans of DNA polymorphism to infer adaptive population divergence. Molecular Ecology, 14(3):671–688.

Turner, T., Hahn, M., and Nuzhdin, S. (2005). Genomic islands of speciation in Anopheles gambiae. Plos BIology, 3(9):1572–1578.

Via, S. (2009). Natural selection in action during speciation. Proceedings of the National Academy of Sciences of the United States of America, 106:9939–9946.

Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-geneflow. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1587, SI):451–460.

Via, S., Conte, G., Mason-Foley, C., and Mills, K. (2012). Localizing Fst outliers on a QTL map reveals evidence for large genomic regions of reduced gene exchange during speciation-with-gene-flow. Molecular Ecology, 21(22):5546–5560.

Via, S. andWest, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. Molecular Ecology, 17(19):4334–4345.

Wang, J., Abbott, R. J., Ingvarsson, P. K., and Liu, J. (2014). Increased genetic divergence between two closely related fir species in areas of range overlap. Ecology and Evolution, pages n/a–n/a.

Wolf, J. B. W., Lindell, J., and Backstrom, N. (2010). Speciation genetics: current status and evolving approaches. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1547):1717–1733.

Wood, H. M., Grahame, J.W., Humphray, S., Rogers, J., and Butlin, R. K. (2008). Sequence differentiation in regions identified by a genome scan for local adaptation. Molecular Ecology, 17(13):3123–3135.

Wright, S. (1969). Evolution and the Genetics of Populations. Vol. 2,The Theory of Gene Frequencies. University of Chicago Press; Chicago, IL.

Wu, C. (2001). The genic view of the process of speciation. Journal of evolutionary biology, 14(6):851–865.

Yeaman, S. and Otto, S. (2011). Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. Evolution, 67(5):2123–2129.

Appendix

Calculating the gene flow factor Bengtsson's method for calculating the gene flow factor γ is described in the appendix of his paper. His method requires one to specify viabilities of genotypes with no more than one "foreign" allele per locus. For example, in the case of two selected diallelic loci, the four relevant genotypes are *AB/AB*; *aB/AB*; *Ab/AB* and *ab/AB* where we assume that "local" and "foreign" alleles are given by the upper-case and lower-case letters, respectively. Let viabilities of these four genotypes be 1; 1-a; 1-b, and 1-s, respectively. Assume that the order of the loci on the chromosome is MAB, where M represents the neutral locus. Let r_1 be the recombinational distance between M and A locus, and r_2 be the distance between loci *A* and *B*. Then the gene flow factor is

$$\gamma_{MAB} = (1-s)r_1 \frac{(r_1-1)^2 r_2[(1-r_2)(1-b)-1]-[(1-r_1)(1-a)-1][r_2 b+(1-r_2)(1-b)-1]}{[1-(1-r_2)(1-a)][1-(1-r_1)(1-b)][1-(1-r_1)(1-a)(1-r_2)(1-s)]}. \tag{1.4}$$

Assume that the order of the loci on the chromosome is AMB. Let r_1 be the distance between A and M and r_2 be the distance between M and B. Then

$$\gamma_{AMB} = (1-s)r_1 r_2 \frac{1-(1-r_1)(1-r_2)(1-a)(1-b)}{[1-(1-r_2)(1-a)][1-(1-r_1)(1-b)][1-(1-r_1)(1-a)(1-r_2)(1-s)]}. \tag{1.5}$$

With three loci under selection, there are 8 relevant genetic backgrounds and four different positions a marker locus can occupy with respect to selected loci (MABC,

AMBC, ABMC, ABCM). The analytical expressions for γ are cumbersome, so we do not present them here.

Individual-based simulations

We consider a population of diploid individuals with discrete non-overlapping generations subdivided in two demes of equal size N. Each individual has two chromosomes with 1024 diallelic loci, of which some are under selection and some are neutral. Mutation rate per locus is m and is equal to 10^{-5} unless stated otherwise. Each generation begins with offspring production by random mating of their parents. Parents are chosen randomly within a deme. Each mating pair produces a random number of offspring chosen from a Poisson distribution with mean $B = 4$, which means that on average the offspring population is twice as large as the parent population. Offspring migrate to the other deme with probability m. After migration, the number of offspring in each deme is reduced to N by viability selection. Viability selection in each deme was implemented by drawing N individuals (without replacement) from the deme's offspring population. The probability that an individual i with fitness w_i is picked during draw j is $w_i / \sum_{k=1}^{N_{OP}-j} w_k$, where N_{OP} is the size of offspring population in deme.

Defining DIG size

To compute the size of GIDs, we first split the chromosome into 128 bins, each containing eight neighboring loci and then calculate the mean F_{st} value for each bin. If the mean F_{st} value of a bin is larger than a predefined cut-off (we chose 5 times the mean F_{st} under neutrality), we say that the bin is part of the GID. To obtain F_{st} under neutrality, we ran simulations under the same parameters as above but with selection coefficients set to zero. The mean GID size was measured as the sum of lengths of all bins that were part of the GID. We performed summation because we are interested in cumulative effects of selected loci on the amount of divergence across the genome. Since we fix the total strength of selection in simulations, as the number of loci increases, each individual locus experiences weaker selection and the region of elevated F_{st} around each selected locus is smaller (compare Figures 1-1, 1-2 and 1-3). Splitting the chromosome in bins containing multiple loci and calculating the average divergence index of a region is a common practice in empirical work studying patterns of divergence (Turner et al., 2005, Hohenlohe et al., 2010, Roesti et al., 2012a,b), but see Via (2009), Via et al. (2012). While empirical studies use more sophisticated statistical methods to determine whether bins have an elevated F_{st} , the number of loci in our simulation is relatively small (1024), and we believe that the simple method we use is good enough to show how the GID size behaves without complicating the analysis.

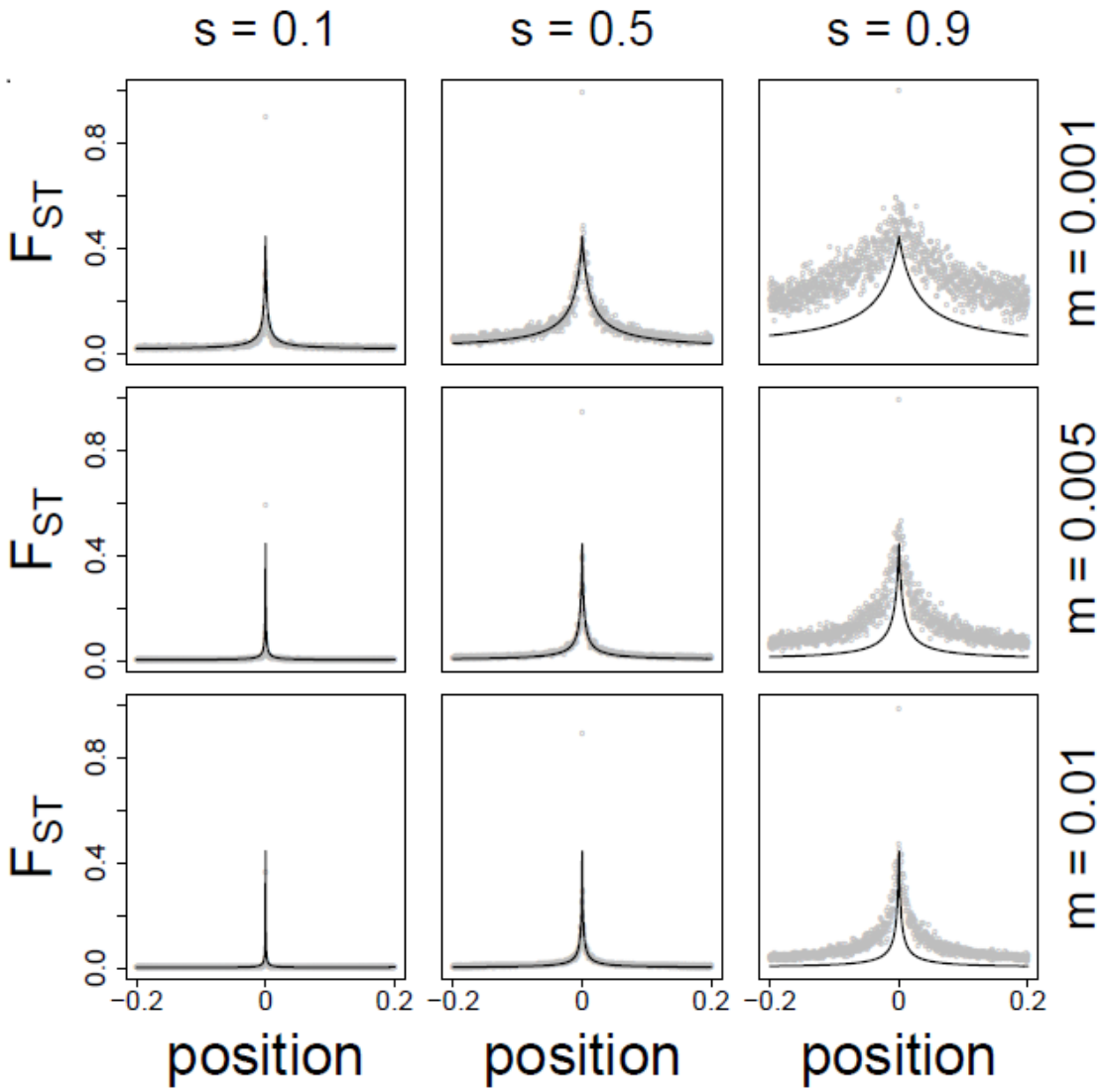

Figure 1-1. F_{st} values for loci across the chromosome with one locus under selection. Black line: analytical predictions, grey dots: mean values from simulation results. N = 4000. The absolute value of position represents the recombinational distance from the center of the chromosome.

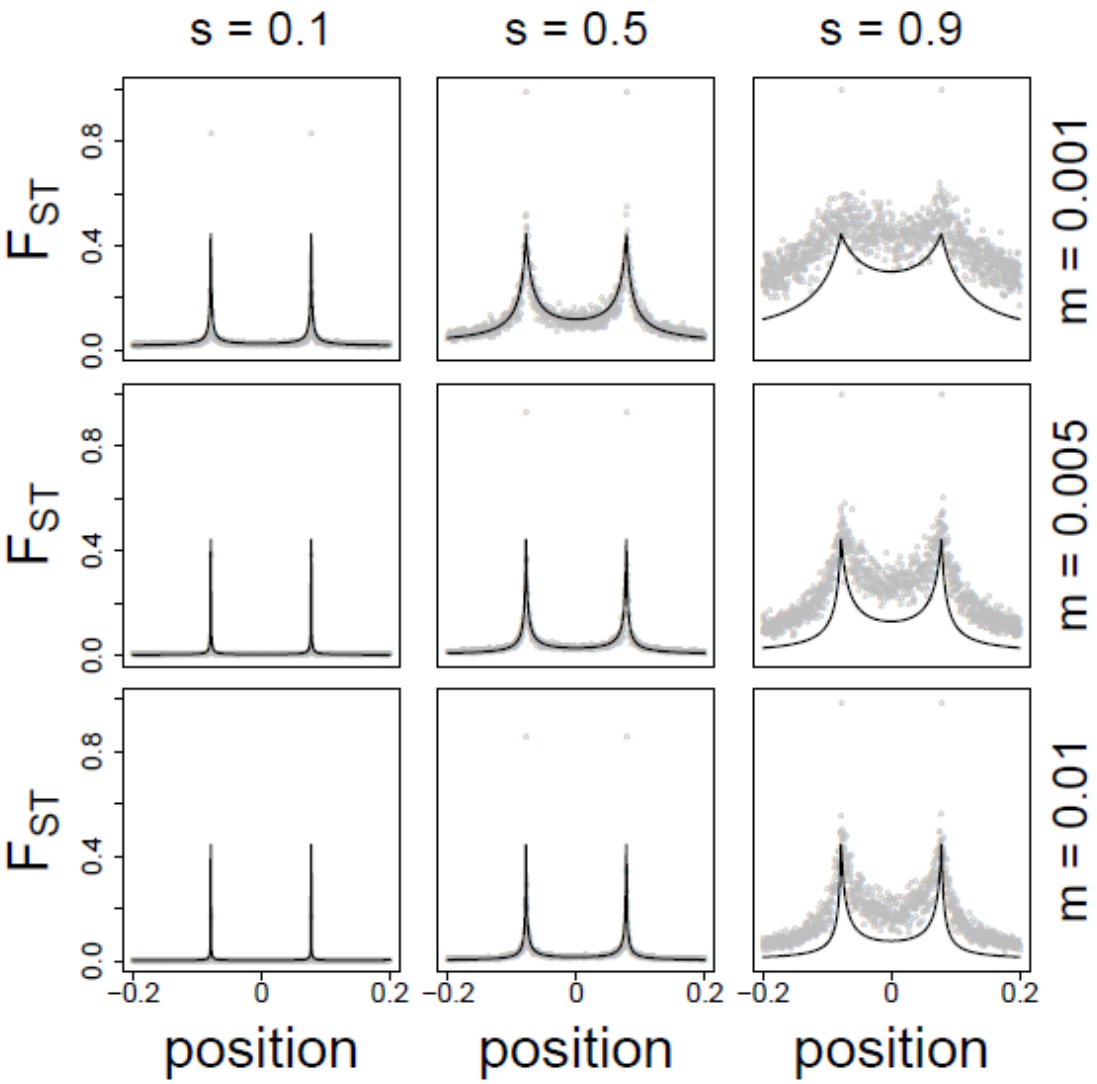

Figure 1-2. F_{st} values for loci across the chromosome with two loci under selection. Black line: analytical predictions, grey dots: mean values from simulation results. N = 4000. The absolute value of position represents the recombinational distance from the center of the chromosome.

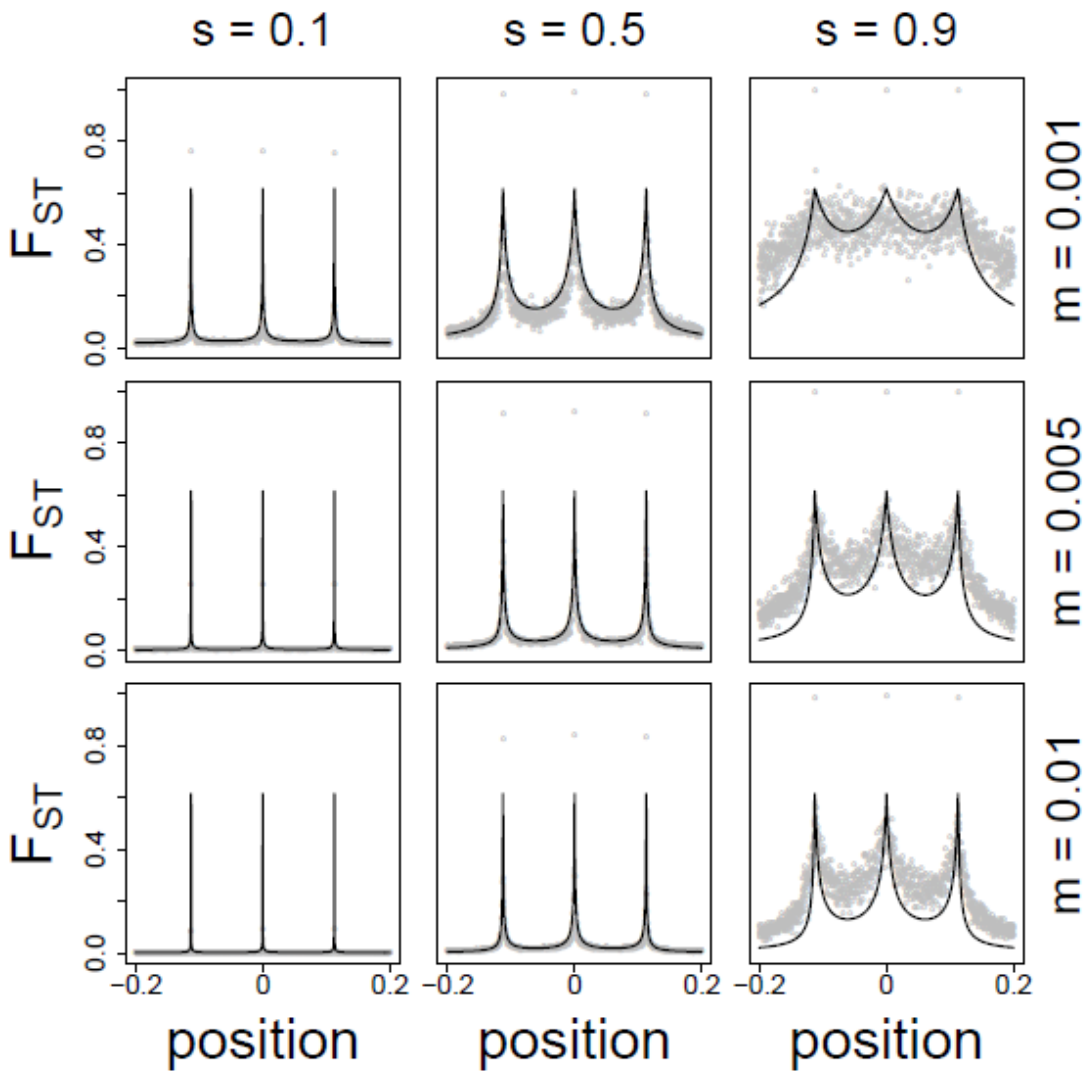

Figure 1-3. Fst values for loci across the chromosome with three loci under selection. Black line: analytical predictions, grey dots: mean values from simulation results. N = 4000. The absolute value of position represents the recombinational distance from the cente

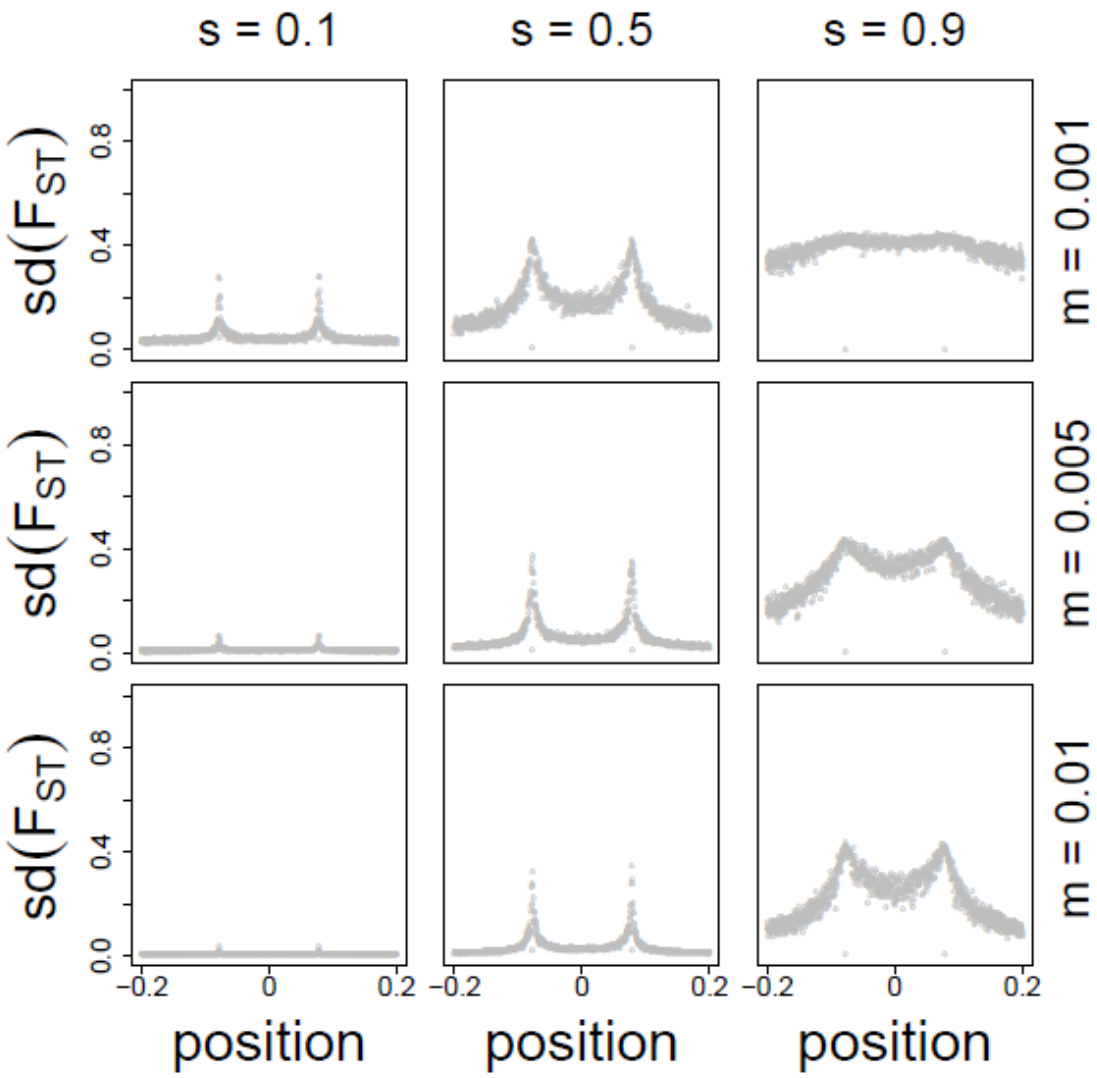

Figure 1-4. Standard error of F_{st} with two loci under selection.

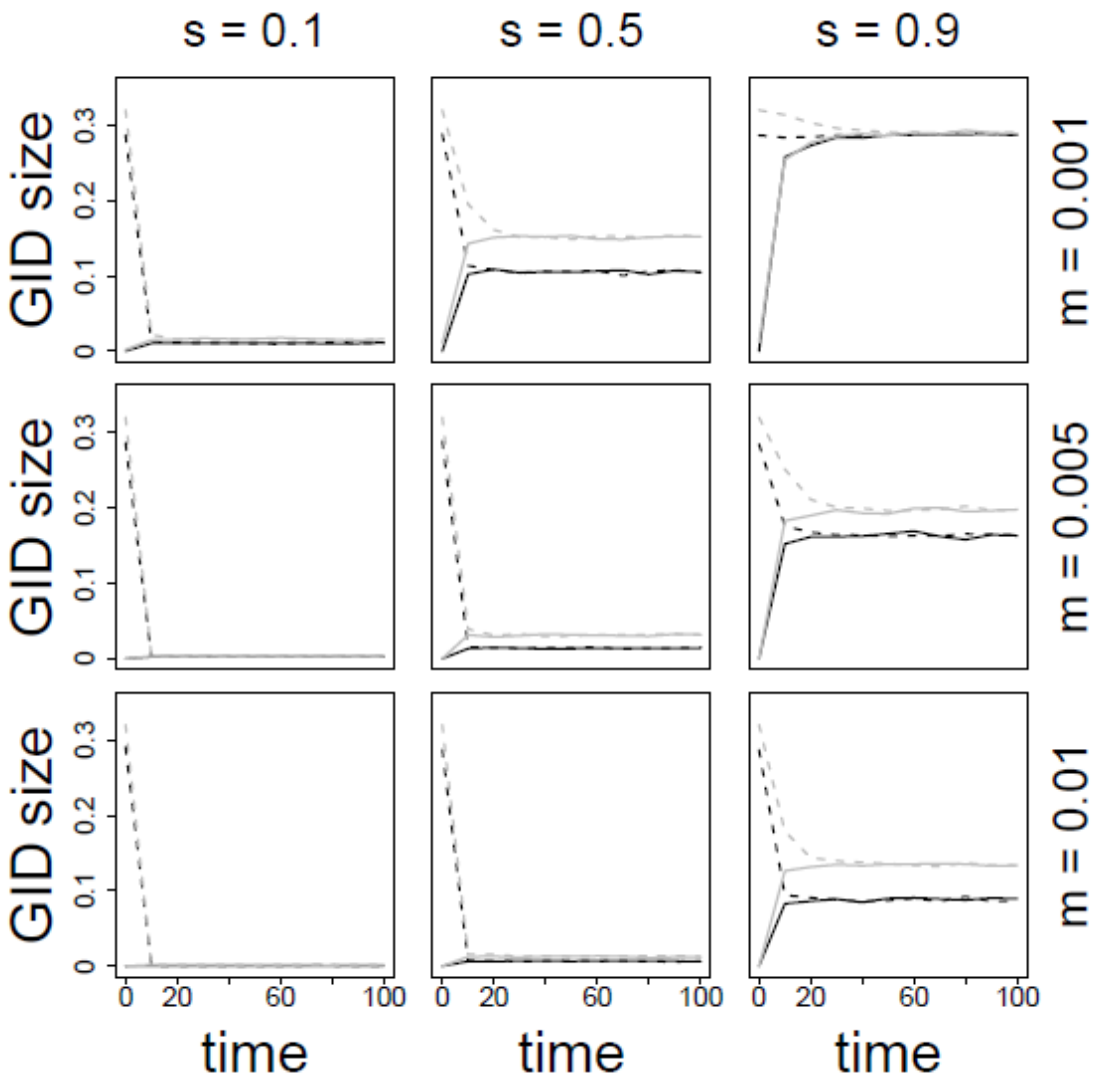

Figure 1-5. Dynamics of the mean GID size for different initial conditions and parameters. Secondary contact (dashed line). Population split (solid line). N = 4000 (black) N = 2000 grey. Each time unit represents 1000 generations.

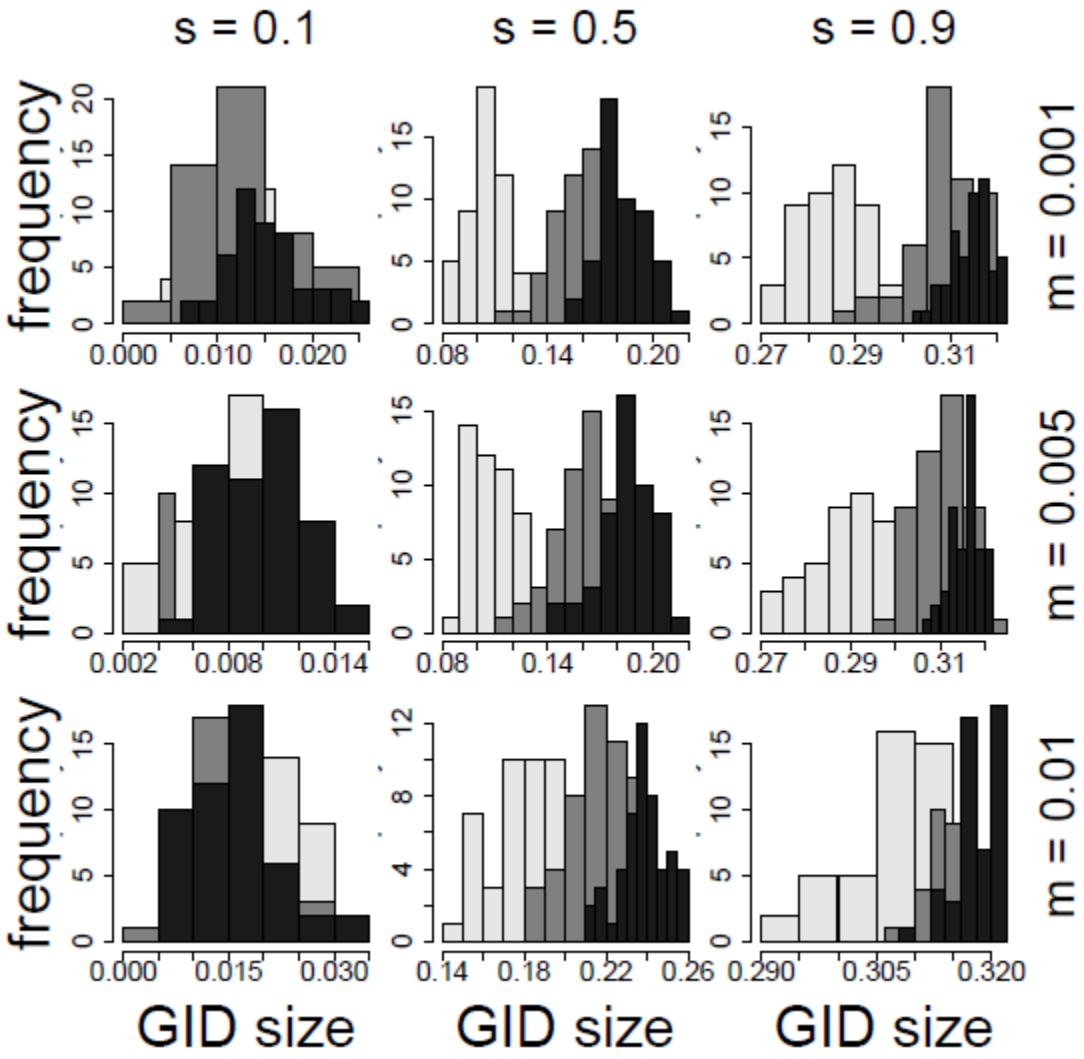

Figure 1-6. Distribution of the GID size for different migration rates m and selection coefficients s. One (light grey), two (intermediate grey), or three (black) loci under selection of the same total strength. N = 4000. Histograms were constructed from 50 samples, each taken 100,000 generations after the start of a simulation.

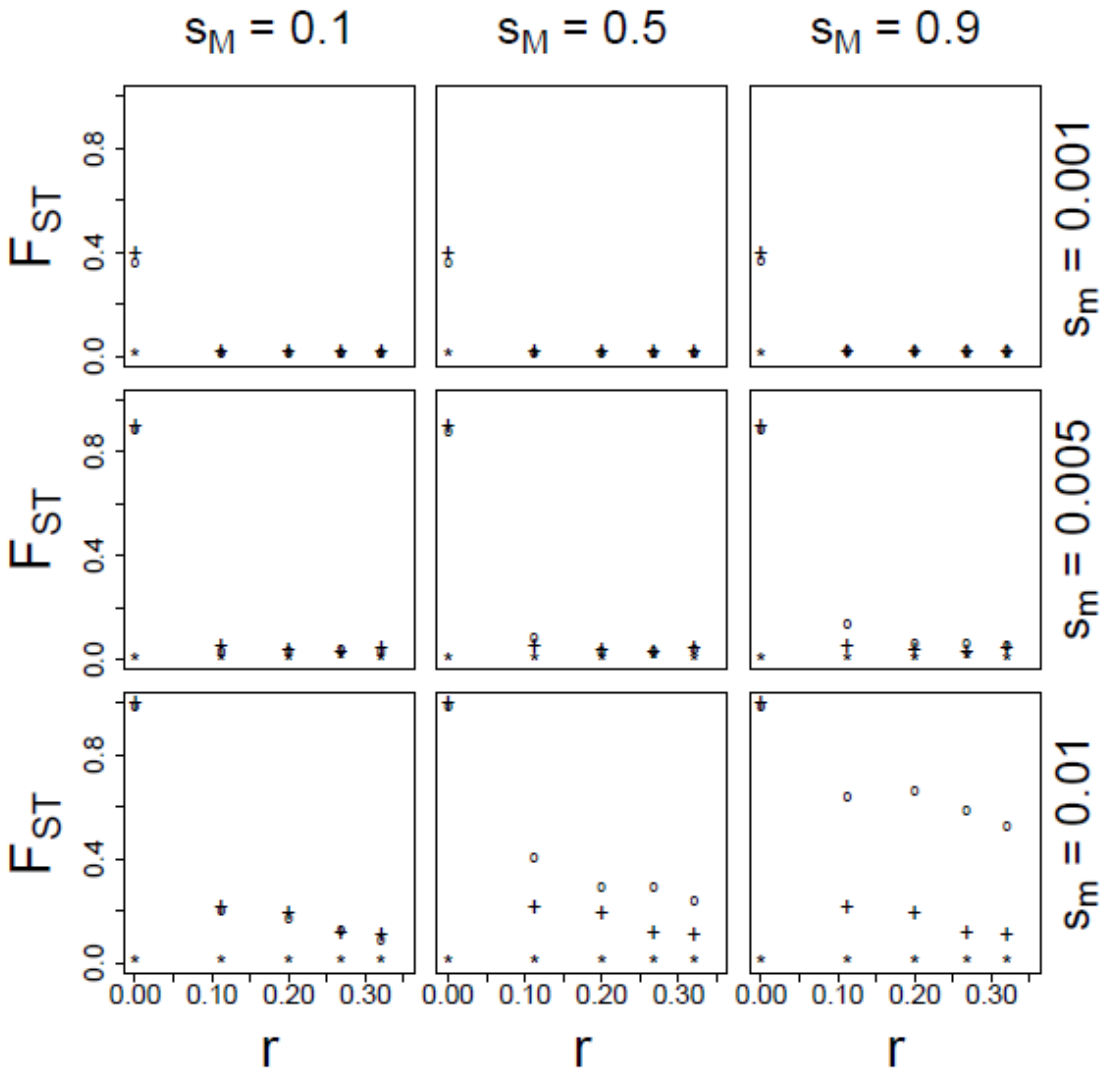

Figure 1-7. Effects of a major selected locus on divergence of minor loci. Shown are F_{st} values at minor loci at different distances r from a single major locus (which is at position r =0). Different symbols correspond to: only major locus is under selection (+), all loci are under selection (o), and only minor loci are under selection (*). The selection strength at major and minor loci is s_M and s_m respectively. N = 1000, m = 0.01.

Chapter 2

Distribution of coalescent times and number of pairwise differences in models of hybridization

Abstract

We study the coalescent process of two genes in a hybridization model that includes population size change and ancestral migration. We obtain the analytical results for the distribution of coalescent times and pairwise differences under infinite site model and symmetrical migration rates and link these results to previously studied "Isolation with initial migration" model. Lastly, we show how to infer model parameters from whole genome scans using our results.

Introduction

Hybridization is an important source of diversity and can arise as a result of numerous mechanisms including environmental change, introduction of new competitors or predators, secondary contact, or reduced selection at low population densities (Hudson et al., 2013, Ward and Blum, 2012, Taylor et al., 2006, Ropiquet and Hassanin, 2006, Seehausen, 2004, Dowling and Secor, 1997). Hybridization increases biological diversity by creating genetic variation, novel traits and new species, and this newly derived diversity can have important ecological and evolutionary consequences (Stebbins, 1959, Mallet, 1995, Arnold, 1997, Vollmer and Palumbi, 2002, Rieseberg et al., 2003, Seehausen, 2004, Dittrich-Reed and Fitzpatrick, 2014).

In sunflowers, ancient hybrids between *H.annuus* and *H.petiolaris* species perform well in novel environments which are not readily available to parent species (Lexer et al., 2003). Novel coloration patterns on the wings of *Heliconius heurippa*, a butterfly species believed to be a hybrid of *H. cydno* and *H. melpomene*, serve as

important anti-predatory and mate recognition signals (Mavarez et al., 2006, Brower, 2011). Many adaptive radiations appear in regions of secondary contact and admixture between previously allopatric lineages, providing further evidence for the potential importance of hybridization in generating diversity (Arnold et al., 2012). In humans, studying population admixture is important to help us to describe our history (Lipson et al., 2013, Novembre and Ramachandran, 2011, Lohmueller et al., 2010), but also to identify genes linked to diseases (Patterson et al., 2004, Shriner et al., 2011).

However, despite widespread interest in hybridization and population admixture, reconstructing the history of hybrid/admixed populations is still a major challenge. Hybridization can be detected from DNA data via several methods such as comparing distribution patterns of mitochondrial and/or nuclear haplotypes across multiple populations, using phylogenetic methods, and fitting data to explicit population models (Barton and Hewitt, 1985, Arnold, 1993, Bertorelle and Excoffier, 1998, Chikhi et al., 2001, Anderson and Thompson, 2002, Wang, 2003, Wilson and Rannala, 2003, Manel et al., 2005, DiCandia and Routman, 2007, Hubisz et al., 2009, Alexander et al., 2009). The genome-wide distribution of single nucleotide polymorphisms (SNP) may also provide us with a means to decipher the history of hybrid populations, however fitting the SNP distribution to a particular demographics model can be challenging and is often performed by means of computations methods, which can be time-consuming and imprecise.

The distribution of pairwise differences is a useful summary statistic which, in principle, can also be used to reconstruct a population's history (Wakeley, 2008,Wakeley

et al., 2012, Huff et al., 2011, Wang and Hey, 2010, Wilkinson-Herbots, 2012). While the distribution of pairwise differences considers only two individuals and thus ignores available data, in some cases it is possible to obtain closed form analytical solutions which can then be used as a rapid means of estimating model parameters when only a few genome-wide scans are available. Apart from that, analytical solutions provide a deeper insight into how population history shapes DNA polymorphism patterns.

Here, we use coalescent theory to obtain solutions for the distribution of pairwise differences in a hybridization (or population admixture) models with complex histories, such as population size change and migration. Coalescent models are a subset of population genetics models that examine DNA polymorphism patterns by tracing the ancestry of the sample as they coalesce back in time until the most common recent ancestor is found (Takahata, 1995, Wakeley, 1996, Rannala, 1997, Excoffier, 2004, Wakeley, 2008). For a general hybridization model, we create a set of equations from which the distribution of pairwise differences can be obtained by numerical methods. We derive the closed form analytical solutions in the case of symmetrical migration rates.

We use those results to gain insight about the importance of sampling genes from hybrid population and to explain the limits of parameter estimation using genome-wide distribution of pairwise differences.

Model

Modelling assumptions

We describe a coalescent process for two genes in a general hybridization model. We assume that genealogies of two genes can be described in terms of Kingman's or structured coalescent (Kingman 1982a,b, Notohara, 1990). By gene we mean a selectively neutral sequence of non-recombining DNA which mutates according to infinite site mutation model (Watterson, 1975).

General model

In the general model, a population splits T_2 generations ago into two populations P_1 and P_2 which we call "parent" populations because they will give rise to hybrid population. Parent populations exchange migrants at rates $m_{1,2}$ and $m_{2,1}$ until T_a generations ago, after which migration stops. T_1 generations ago a third, hybrid (*H*), population is formed by an admixture of two parent populations. To keep model general, we also allow all existing populations to change sizes at T_1, T_a and T_2 (Figure 2-1).

We rescale model parameters by *2N* and we define $\tau_1 = 2NT_1$, $\tau_a = 2NT_a$, and $\tau_2 = 2NT_2$. Let p_1 be the probability that an ancestral lineage of one gene was in population P_1 at τ_1. Then $p_2 = 1 - p_1$ is the probability that an ancestral lineage is in the other parent population. In natural systems p_1 can depend on numerous factors, but we treat it as a parameter that can be between 0 and 1 without going into details about mechanisms that define its value.

There are six ways to sample two genes from three populations when order in which genes are picked does not matter. In deriving the distribution of coalescent times for each of those cases, we will encounter three different situations: lineages in the same isolated population, lineages in different isolated populations and lineages in population(s) that are exchanging migrants.

When lineages are in the same isolated population, distribution of coalescent time is exponentially distributed with mean proportional to population size. When two lineages are in different populations, they cannot coalesce.

We use the formalism of structured coalescent to derive the distribution of coalescent times when lineages are in population(s) exchanging migrants (Notohara, 1990). When population size goes to infinity and the product of population size and migration rate converges to constant, after scaling by population size, the coalescent process can be described as a continuous Markov process with rate matrix Q. For two genes and two populations of sizes $2Nx_1$ and $2Nx_2$ exchanging migrants with backward migration rates $m_{1,2}$ (migration from $2Nx_1$ to $2Nx_2$) and $m_{2,1}$ (migration from $2Nx_2$ to $2Nx_1$), Q has five states: two lineages in the first population, one lineage in each population, two lineages in the second population and two lineages coalesced in first and second population respectively. Matrix Q is then given by:

$$Q = \begin{pmatrix} -1/x_1 - M_{1,2} & M_{1,2} & 0 & 1/x_1 & 0 \\ M_{2,1}/2 & -(M_{1,2}+M_{2,1})/2 & M_{2,1}/2 & 0 & 0 \\ 0 & M_{2,1} & -1/x_2 - M_{2,1} & 0 & 1/x_2 \\ 0 & 0 & 0 & -M_{1,2}/2 & M_{1,2}/2 \\ 0 & 0 & 0 & M_{2,1}/2 & -M_{2,1}/2 \end{pmatrix} \tag{2.1}$$

Where $M_{i,j} = 4Nm_{i,j}$, $i, j = 1,2$, $i \neq j$ is scaled migration size. We obtain the probability of being in state *j* after some time *t* given it started in state i using standard continuous Markov chain methods by calculating matrix exponent of Q,

$$A(x_1, x_2, M_{1,2}, M_{2,1}, t) = e^{Qt} = \sum_{k=0}^{\infty} (Qt)^k / k!.$$

Distribution of coalescent times, expected coalescent time and the distribution of pairwise differences

First case we consider is when two genes are sampled from the same extant population P_j. Before τ_a lineages are in the same isolated population P_j, that changed size at τ_1. Coalescent time follows exponential distribution with mean d_j prior to τ_1 and, given that lineages did not coalesce (probability $exp(-\tau_1 / d_j)$), c_j between τ_1 and τ_a. Assuming no coalescent happened, at τ_a two lineages will both in population of size b_j ,which is a population exchanging migrants. Since lineages can coalesce only if they are in the same population, only entries of e^{Qt} corresponding to those cases will be used to describe coalescent between τ_a and τ_2. Lastly, if coalescent did not happen by τ_2,

lineages will find themselves in the same (isolated) population and will coalesce with rate *a*.

Therefore, the probability density function (p.d.f) of coalescent times in hybridization model can be written as a combination of different stages, with coalescent in each stage described by appropriate exponentially distributed random variables multiplied by the probability of coalescence not happening prior to the stage:

$$f_{T_s^{P_j}}(t) = \begin{cases} \frac{1}{d_j} e^{-\frac{t}{d_j}} & 0 \le \mathrm{t} < \tau_1 \\ \frac{1}{c_j} e^{-\frac{\tau_1}{d_j} - \frac{t-\tau_1}{c_j}} & \tau_1 \le \mathrm{t} < \tau_\mathrm{a} \\ e^{-\frac{\tau_1}{d_j} - \frac{\tau_a - \tau_1}{c_j}} \left(\frac{1}{b_1} A_{s_1,s_1}(t-\tau_a) + \frac{1}{b_2} A_{s_1,s_2}(t-\tau_a)\right) & \tau_\mathrm{a} \le \mathrm{t} < \tau_2 \\ \frac{1}{a} e^{-\frac{\tau_1}{d_j} - \frac{\tau_a - \tau_1}{c_j} - \frac{t-\tau_2}{a}} \left(A_{s_1,s_1}(\tau_2-\tau_a) + A_{s_1,2}(\tau_2-\tau_a) + A_{s_1,s_2}(\tau_2-\tau_a)\right) & \tau_2 \le \mathrm{t} \end{cases} \tag{2.2}$$

Where $s_1 = 1$ and $s_2 = 3$ if $j = 1$ and $s_1 = 3, s_2 = 1$ if $j = 2$.

Similarly, when two genes are sampled from different extant parent population, coalescent is possible only after τ_a, at which time two lineages are in different populations exchanging migrants. The p.d.f. of coalescent times is:

$$f_{T_d^P}(t) = \begin{cases} 0 & 0 \le t < \tau_a \\ \frac{1}{b_1} A_{2,1}(t-\tau_a) + \frac{1}{b_2} A_{2,3}(t-\tau_a) & \tau_{\mathrm{a}} \le \mathrm{t} < \tau_2 \\ \frac{1}{a} e^{-\frac{t-\tau_2}{a}} \sum_{l=1}^{3} A_{2,l}(\tau_2 - \tau_a) & \tau_2 \le \mathrm{t} \end{cases}$$

(2.3)

When one gene is sampled from an extant population P_j and the other from the hybrid population H, with probability $1-p_j$ ancestral lineages will be in different populations between τ_1 and τ_a. The coalescent process is then the same as when two genes are sampled from different extant parent populations. With probability p_j, ancestral lineages will be in the same population, and the distribution of coalescent times will be given by $f_{T_s^{P_jH}}$:

$$f_{T_s^{P_jH}}(t) = \begin{cases} 0 & 0 \le \mathrm{t} \le \tau_1 \\ \frac{1}{c_j} e^{-\frac{t-\tau_1}{c_j}} & \tau_{\mathrm{a}} \le \mathrm{t} \le \tau_1 \\ e^{-\frac{\tau_a-\tau_1}{c_j}} (\frac{1}{b_1} A_{1,k_1}(t-\tau_a) + \frac{1}{b_2} A_{1,k_2}(t-\tau_a)) & \tau_{\mathrm{a}} < \mathrm{t} \le \tau_2 \\ \frac{1}{a} e^{-\frac{\tau_a-\tau_1}{c_j} - \frac{t-\tau_2}{a}} (\sum_{l=1}^{3} A_{1,l}(\tau_2 - \tau_a)) & \tau_2 < \mathrm{t} \end{cases}$$

(2.4)

The distribution of coalescent times in when one gene is sampled from population P_j, and the other from H can be written as:

$$f_{T^{P_jH}} = p_j f_{T_s^{P_jH}} + (1-p_j) f_{T_d^P}$$

(2.5)

Lastly, for two genes are sampled from hybrid population, we can write:

$$f_{T^H} = p_1^2 f_{T_{s1}^H} + p_2^2 f_{T_{s2}^H} + 2p_1 p_2 f_{T_d^H}$$

(2.6)

Terms on the right side correspond to cases when after τ_1 two ancestral lineages are in P_1, P_2 and in different populations. Expressions for $f_{T_{s1}^H}$ and $f_{T_{s2}^H}$ are obtained from $f_{T_s^{P_j}}$ by replacing d_j in equation (2.2) with d_h. When two lineages are in different populations between τ_a and τ_1, coalescent is not possible. Therefore, term for $f_{T_d^H}$ is:

$$f_{T_d^H}(t) = \begin{cases} \frac{1}{d_h} e^{\frac{-t}{d_h}} & 0 \le \mathrm{t} \le \tau_1 \\ 0 & \tau_\mathrm{a} \le \mathrm{t} \le \tau_1 \\ e^{\frac{-\tau_1}{d_h}} (\frac{1}{b_1} A_{2,1}(t-\tau_a) + \frac{1}{b_2} A_{2,3}(t-\tau_a)) & \tau_\mathrm{a} < \mathrm{t} \le \tau_2 \\ \frac{1}{a} e^{\frac{-\tau_1}{d_h} - \frac{t-\tau_2}{a}} (\sum_{l=1}^{3} A_{2,l}(\tau_2 - \tau_a)) & \tau_2 < \mathrm{t} \end{cases}$$

(2.7)

The expected coalescent time of a random variable T with p.d.f f_T is:

$$E[T] = \int_0^\infty t f_T(t) dt$$

(2.8)

Since under infinite site model each mutation produces one new pairwise difference and mutations accumulate over time independently across lineages according to Poisson process, the probability of k pairwise differences is a function of the distribution of

coalescent time, and for a random coalescent time variable T with p.d.f $f_T(t)$, the distribution of pairwise differences is given by the integral:

$$S(T) = P(T = k) = \int_0^\infty \frac{(\theta t)^k e^{-\theta t}}{k!} f_T(t) dt \tag{2.9}$$

Unfortunately, there is no simple expression for this expression in our general hybridization model, but it can be evaluated numerically using many readily available computer programs. However, we can obtain exact analytical solution when migration is symmetrical.

Model with symmetric migration

To derive closed-form results, we assume that migration between the two populations is symmetric and equal to m. To keep the population sizes constant during migration time, we also assume that parent populations' sizes are the same ($b_1 = b_2 = b$). For simplicity, we set *b = 1*, but as long as the population sizes are the same, we can easily obtain equivalent values for b different than 1 if we rescale time by *2Nb*.

We found the expressions for relevant entries of e^{Qt} "general model" section (see appendix), but instead of using them directly, we rewrite them in a way similar to (Wilkinson-Herobots, 2012). We then obtain:

$$e_{1,1}^{Qt} = \frac{1}{2}\left(\sum_{r=1}^{2} A_{0r}\lambda_r e^{-\lambda_r t} + e^{-(M+1)t}\right)$$

(2.10a)

$$e_{1,2}^{Qt} = \sum_{r=1}^{2} A_{1r}\lambda_r e^{-\lambda_r t}$$

(2.10b)

$$e_{1,3}^{Qt} = \frac{1}{2}(\sum_{r=1}^{2} A_{0r}\lambda_r e^{-\lambda_r t} - e^{-(M+1)t})$$

(2.10c)

$$e_{2,1}^{Qt} = e_{2,3}^{Qt} = \frac{1}{2} e_{1,2}^{Qt}$$

(2.10d)

$$e_{2,2}^{Qt} = A_{01}\lambda_1 e^{-\lambda_2 t} + A_{02}\lambda_2 e^{-\lambda_1 t}$$

(2.10e)

where:

$\lambda_1 = \frac{M+1/2-\sqrt{D}}{2}$, $\lambda_2 = \frac{M+1/2+\sqrt{D}}{2}$, $D = (2M+1)^2 - 4M$, $A_{01} = \frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}$ and

$A_{02} = \frac{1-\lambda_1}{\lambda_2 - \lambda_1}$.

From equation (2.2) and (2.10a-c), we obtain the probability density function (p.d.f.) of coalescent times for two genes sampled from extant parent population *j*:

$$f_{T_s^{P_j}}(t)=\begin{cases}\frac{1}{d_j}e^{-(1/d_j)t} & 0\le \mathrm{t}<\tau_1\\ \frac{1}{c_j}e^{-\tau_1/d_j-(t-\tau_1)/c_j} & \tau_1\le \mathrm{t}<\tau_\mathrm{a}\\ e^{-\tau_1/d_j-(\tau_a-\tau_1)/c_j}\sum_{r=1}^{2}A_{0r}\lambda_r e^{-\lambda_r(t-\tau_a)} & \tau_\mathrm{a}\le \mathrm{t}<\tau_2\\ \frac{1}{a}e^{-\tau_1/d_j-(\tau_a-\tau_1)/c_j-(t-\tau_2)/\mathrm{a}}\sum_{r=1}^{2}A_{0r}e^{-\lambda_r(\tau_2-\tau_a)} & \tau_2\le \mathrm{t}\end{cases}$$

(2.11)

In appendix, we show that the expected value of $T_s^{P_j}$ is:

$$E[T_s^{P_j}]=d_j+e^{-\tau_1/d_j}(c_j-d_j+e^{-(\tau_a-\tau_1)/c_j}(2-c_j+\sum_{r=1}^{2}A_{0r}(a-1/\lambda_r)e^{-\lambda_r(\tau_2-\tau_a)}))$$

(2.12)

Similarly, we show in appendix that the probability of observing k pairwise differences is:

$$P(S_s^{P_j}=k)=F_1(d_j)+e^{-\tau_1/d_j}(F_2(c_j,\tau_1)-F_2(d_j,\tau_1)+e^{-(\tau_a-\tau_1)/c_j}(\sum_{r=1}^{2}A_{0r}F_2(1/\lambda_r,\tau_a)-F_2(c_j,\tau_a)$$
$$+\sum_{r=1}^{2}A_{0r}e^{-\lambda_r(\tau_2-\tau_a)}(F_2(a,\tau_2)-F_2(1/\lambda_r,\tau_2)))$$

(2.13)

where:

$$F_1(x)=(x\theta)^l/(1+x\theta)^{l+1}$$

(2.14)

$$F_2(x,\tau)=\frac{e^{\theta\tau_i}(x\theta)^l}{(1+x\theta)^{l+1}}\sum_{m=0}^{l}(1/x+\theta)^m\tau^m/m!$$

(2.15)

and $\theta = 4N\mu$ is the scaled mutation rate. Equations (2.14) and (2.15) are explained in Appendix.

The expressions above are equivalent to the two-population "Isolation with initial migration" (*IIM*) model of (Wilkinson-Herbots, 2012) with population size change during isolation time.

When two genes are sampled from different extant parent populations, $f_{T_d^P}$ from equation (2.3) is given by:

$$f_{T_d^P}(t) = \begin{cases} 0 & 0 \le \mathrm{t} \le \tau_1 \\ \sum_{r=1}^{2} A_{1r}\lambda_r e^{-\lambda_r(t-\tau_1)} & \tau_1 < \mathrm{t} \le \tau_2 \\ \frac{1}{a} e^{-\frac{1}{a}(t-\tau_2)} \sum_{r=1}^{2} A_{1r} e^{-\lambda_r(\tau_2-\tau_1)} & \tau_2 < \mathrm{t} \end{cases} \tag{2.16}$$

where $A_{11} = \dfrac{\lambda_2}{\lambda_2 - \lambda_1}$ and $A_{12} = \dfrac{-\lambda_1}{\lambda_2 - \lambda_1}$.

This case is when two genes are sampled from different populations in IMM model, so the expected value of T_d^P is given by equation (25) in (Wilkinson-Herbots, 2012):

$$E[T_d^P] = \tau_a + 2 + 1/M + \sum_{r=1}^{2} A_{1r}(a - 1/\lambda_r)e^{-\lambda_r(\tau_2-\tau_a)} \tag{2.17}$$

The probability of observing k pairwise differences is then:

$$P(S_d^{P_j} = k) = \sum_{r=1}^{2} A_{1r}(F_2(1/\lambda_r, \tau_a) - e^{-\lambda_r(\tau_2-\tau_a)}(F_2(a,\tau_2) - F_2(1/\lambda_r, \tau_2)))$$

(2.18)

When one gene is sampled from hybrid population and the other from extant parent population $f_{T_s^{HP_j}}$ (equation 2.4) is given by:

$$f_{T_s^{HP_j}} = \begin{cases} 0 & 0 \le t < \tau_1 \\ (1/c_j)e^{-(t-\tau_1)/c_j} & \tau_1 \le t < \tau_a \\ e^{-(\tau_a-\tau_1)/c_j}(\sum_{r=1}^{2} A_{0r}\lambda_r e^{-\lambda_r(t-\tau_a)}) & \tau_a \le t < \tau_2 \\ e^{-(\tau_a-\tau_1)/c_j-(t-\tau_2)/a}\sum_{r=1}^{2} A_{0r}e^{-\lambda_r(\tau_2-\tau_a)} & \tau_2 \le t \end{cases}$$

(2.19)

From equation (2.5) we derive the expression for the expected value of T^{HP_j} in a similar way as equation (2.12) as:

$$E[T^{HP_j}] = p_j(\mathrm{c}_j + \tau_1 + e^{-(\tau_a-\tau_1)/c_j}(2 - c_j + \sum_{r=1}^{2} A_{0r}(a - 1/\lambda_r)e^{-\lambda_r(\tau_2-\tau_a)}))$$
$$+(1-p_j)(\tau_1 + 2 - 1/M + \sum_{r=1}^{2} A_{1r}(a - 1/\lambda_r)e^{-\lambda_r(\tau_2-\tau_a)})$$

(2.20)

The probability of observing k pairwise differences is then:

$$P(S^{HP_j} = k) = p_j(F_2(c_j,\tau_1) + e^{-(\tau_a-\tau_1)/c_j}(\sum_{r=1}^{2} A_{0r}F_2(1/\lambda_r,\tau_a) - F_2(c_j,\tau_a)$$
$$+\sum_{r=1}^{2} A_{0r}e^{-\lambda_r(\tau_2-\tau_a)}(F_2(a,\tau_2) - F_2(1/\lambda_r,\tau_2)))$$
$$+(1-p_j)(\sum_{r=1}^{2} A_{1r}(F_2(1/\lambda_r,\tau_a) - e^{-\lambda_r(\tau_2-\tau_a)}(F_2(a,\tau_2) - F_2(1/\lambda_r,\tau_2))))$$

(2.21)

Lastly, when two genes are sampled from the hybrid population the distribution of coalescent times is given by equation (2.7), and $f_{T_d^H}$ is given by:

$$f_{T_d^H} = \begin{cases} (1/d_h)e^{-(t-\tau_1)/c_h} & 0 \le t < \tau_1 \\ 0 & \tau_1 \le t < \tau_a \\ e^{-\tau_1/\mathrm{d}_h}(\sum_{r=1}^{2} A_{0r}\lambda_r e^{-\lambda_r(t-\tau_a)}) & \tau_a \le t < \tau_2 \\ e^{-\tau_1/\mathrm{d}_h-(t-\tau_2)/a}\sum_{r=1}^{2} A_{0r}e^{-\lambda_r(\tau_2-\tau_a)} & \tau_2 \le t \end{cases}$$

(2.22)

The expressions for $E[T^H]$ and $S(T^H)$ are:

$$E[T^H] = \sum_{j=1}^{2} p_j^{\,2}(d_h + e^{-\tau_1/d_j}(c_j - d_h + e^{-(\tau_a-\tau_1)/c_j}(2 - c_j + \sum_{r=1}^{2} A_{0r}(a - 1/\lambda_r)e^{-\lambda_r(\tau_2-\tau_a)}))) \\ +2p_1p_2(d_h + e^{-\tau_1/d_j}(2 + 1/M + \tau_a - d_h - \tau_1) + \sum_{r=1}^{2} A_{1r}(a - 1/\lambda_r)e^{-\lambda_r(\tau_2-\tau_a)})$$

(2.23)

$$P(S^H = k) = \sum_{j=1}^{2} p_j(F_1(d_h) + e^{-\tau_1/d_j}(F_2(c_j,\tau_1) - F_2(d_h,\tau_1) + e^{-(\tau_a-\tau_1)/c_j}(\sum_{r=1}^{2} A_{0r}F_2(1/\lambda_r,\tau_a) - F_2(c_j,\tau_a) \\ + \sum_{r=1}^{2} A_{0r}e^{-\lambda_r(\tau_2-\tau_a)}(F_2(a,\tau_2) - F_2(1/\lambda_r,\tau_2)))) + 2p_1p_2(F_1(d_h) + e^{-\tau_1/d_j}(\sum_{r=1}^{2} A_{1r}F_2(1/\lambda_r,\tau_a) \\ - F_2(d_h,\tau_1) + \sum_{r=1}^{2} A_{1r}e^{-\lambda_r(\tau_2-\tau_a)}(F_2(a,\tau_2) - F_2(1/\lambda_r,\tau_2)))$$

(2.24)

Discussion

Wang and Hey (2010) used the distribution of pairwise differences from a large number of genes of *D.melanogaster* and *D.simulans* to estimate parameters of the "Isolation with Migration" model (Hey and Nielsen, 2004). They inferred nonzero migration from

D.simulans to *D.melanogaster*, showing the usefulness of a "few individuals but many loci" approach. Hobolth et al. (2011) have shown how to compute the distribution of coalescent times of two genes in "Isolation with Migration" model using properties of continuous time Markov chain (also see (Notohara, 1990) and chapter 4 in (Wakeley, 2008) for using the same approach on different migration models). Wilkinson-Herbots (2012) has found the distribution of coalescent times and pairwise differences in "Isolation with Initial Migration" (*IIM*) model with symmetric migration. In *IIM* model, after the initial split, two populations share migrants for some time, but eventually stop and evolve in isolation.

Compared to models describing population divergence, there are fewer analogous analytical results for hybridization models. Bertorelle and Excoffier (1998) developed statistics based on mean coalescent times to estimate admixture coefficient in a hybridization model with no migration and equal population sizes. A more general hybridization model which allows change of parent population sizes and migration after hybridization was considered in (Wang, 2003), but focus of that paper was developing a numerical method for parameter estimation.

In this paper we considered a hybridization model with complex parent population history using the approach equivalent to that described in Hobolth et al. (2011). In the case of symmetric migration rate, we found the closed form expressions for the distribution of coalescent times and pairwise differences when two genes are sampled from each of 6 different population combinations.

Distribution of pairwise differences

The distributions of coalescent times in the hybridization model can for the most part be expressed in terms of modified *IIM* divergence model (Wilkinson-Herobots, 2012).

When two genes are sampled from extant parent populations, expressions we obtained are equivalent to ones in "Isolation with initial migration" (*IIM*) (Wilkinson-Herbots, 2012) if populations change sizes during period of isolation. Therefore the distribution of coalescent times $T_s^{P_j}$ and T_d^P can be continuous or discontinuous, and the distribution of pairwise differences can be unimodal or multimodal, depending on parameters (Wilkinson-Herbots, 2012).

When one gene is sampled from a hybrid population and the other from an extant parent population j the resulting distributions are a weighted average of two cases: 1) ancestral lineages in the same population and 2) ancestral lineages in different populations at a time preceding hybridization (after τ_1). Weights are p_j and $1-p_j$ respectively. The first case is mathematically equivalent to sampling two genes from extant parent population of size d_h prior to τ_1. The second case is equivalent to sampling two genes from different extent parent populations. Therefore, the shape of the distribution of pairwise differences also depends on p_j.

The coalescent process of two genes sampled from a hybrid population consists of three cases, two of which (each occurs with probability p_j^2) can be described in terms of a modified *IIM* model. In the third case, which occurs with probability $2p_1p_2$, ancestral lineages are in different populations just after τ_1. This case is specific for hybridization model. Similarly as in the case when one gene is sampled from extant parent and the other from hybrid population, the shape of the distribution of pairwise differences of two genes sampled from hybrid population can vary depending on the admixture parameter.

Migration and change in population size can have the same effect on the distribution of pairwise differences. To show that, we consider two special cases, one in which two parent populations are of constant sizes and exchange migrants continuously until hybridization $(c_1 = c_2 = b_1 = b_2 = 1, \tau_a = \tau_1)$ and the other in which parent populations changes size but do not exchange migrants $(M = 0)$. For some parameter sets, both models produce the same distribution of pairwise differences when one or both genes are sampled from hybrid population (Figure 2-2). This result is of particular interest because it shows that when one parent population cannot be sampled (because it went extinct for example), we can't uniquely distinguish between change in parent population size and migration based on the distribution of pairwise differences alone. Even with all populations available, it may be difficult to distinguish between population size change and migration as different parameter combination can result in same distributions of pairwise differences (Figure 2-3).

More work is needed to explore conditions when migration and the change in population size result in the same or very similar distributions of pairwise differences. However, since calculating likelihoods can be done reasonably fast, a person interested in data analysis could fit multiple different models and perform model comparison (for example likelihood-ratio test when possible or AICc (Hurvich and Tsail, 1989)) to test how well the model fits the data.

Parameter estimation

To understand how the distribution of pairwise differences relates particular model parameters, and test whether model parameters can be estimated, we used the ms program (Hudson, 2002) to simulate 5000 pairs of genes sampled from each possible pair of extant populations assuming infinite site mutation model. The resulting six sets of 5000 numbers were used to calculate likelihood function. For simplicity, we assumed that all population sizes are the same, migration is continuous between τ_1 and τ_a and that θ is known. For each of 5 parameters in this simplified model, we considered 10 or more different values to calculate the likelihood of observed data. Given that the pair of genes is sampled from populations i and j (where i and j can be P_1, P_2 and H) the number of pairwise differences (Data) given model parameters is just ($P_{i,j}(Data \mid parameters)$) using equations (2.13), (2.18), (2.21) and (2.24). Since all pairs of genes are independent, the likelihood of all the number of pairwise differences for all gene pairs is a product of

likelihoods of each gene pair (Takahata et al. 1995). To calculate the marginal likelihood of a particular model parameter, we integrate out other model parameters.

Depending on the parameters, we could estimate some parameters better than others. Figure 2-4 shows the case when migration rate was estimated poorly compared to other parameters. That is because the distribution of pairwise differences for that particular parameter set does not depend as strong on migration rate as on other parameters (compare Figure 2-5 to Figure 2-6). When polymorphism is low (which happens for small θ, small population sizes, high migration, recent hybridization or parent population split) different parameter combinations will produce same distribution of pairwise differences resulting in flat posterior distribution.

The main focus of this paper was describing a coalescent process in a hybridization model and understanding how it connects to other models. The parameter estimation approach illustrated here, while encouraging, might not be applicable for some empirical studies for several reasons. First, we are assuming no recombination between genes. It might be hard to find enough appropriate genes if chromosome is small or if recombination rates across the chromosomes are unequal. A possible way to mitigate this problem is to use short DNA segments separated by longer stretches and to avoid genes in parts of genome with low recombination (recombinational coldspots or inversions for example). Wang and Hey (2010) used 500 bp long DNA segments separated by at least 2000 bp for their analysis of *Drosophila* species. To calculate the likelihood of data, we are assuming that genes are independent. This means that different genes need to be used for building a distribution of pairwise differences for each of 6 different population pairs,

causing a problem with using this method on small genomes. Parameter estimation method presented here relies on comparing two genes. Expanding model to consider more than two sequences is possible, but obtaining analytical result is harder. Also, the number of ways to sample genes from different populations is larger. For example, with 3 genes there are 9 different ways to choose genes from 3 populations. Given that genes cannot be reused in different population pairs, it is unclear how to sample genes most efficiently. In this paper we are assuming infinite site mutation model, but different mutation models can be included. For example, under Jukes-Cantor mutation model, the probability that a nucleotide is the same after time t is $1/4+(3/4)e^{-4t/3}$, so we can use the same approach as the one outlined in appendix to derive the probability of homozygosity under this mutation model. We were able to obtain analytical results for a model with symmetric migration. For asymmetrical migration, we need to rely on numerical methods to find the expressions for the exponent of Q matrix. A model with symmetrical migration already has 11 parameters (6 population sizes, 3 times, migration rate, admixture coefficient and scaled mutation rate) and we did not explore how well all parameters can be estimated. However, based on the results of a simplified model, we would not be surprised if multiple different parameter combinations might result in equally good fit to data. With asymmetrical migration and population sizes, the number of parameters will increase. Future work will focus on developing ways to deal with asymmetrical migration, different population sizes and other issues we mentioned.

Conclusion

We described the coalescent process for the sample of two genes in a hybridization model which allows for the complex population histories. We obtained analytical results for the distribution of coalescent times and pairwise differences under infinite site model in the case of symmetrical migration rate and equal population sizes. We have shown how this model relates to previously studied “Isolation with initial migration” models. Lastly, we have shown that at least in some cases model parameters can be inferred from data, however more work is needed to better understand when accurate parameter estimation is possible.

References

Anderson, E. C. and Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. Genetics, 160(3):1217–1229.

Arnold, J. (1993). Cytonuclear disequilibria in hybrid zones. Annual Review of Ecology and Systematics, 24:521–554.

Arnold, M., Hamlin, J., Brothers, A., and Ballerini, E. (2012). Rapidly Evolving Genes and Genetic Systems, chapter Natural hybridization as a catalyst of rapid evolutionary change., pages 256–265. Oxford University Press, Oxford; Oxford; UK.

Arnold, M. L. (1997). Natural Hybridization and Evolution. Oxford University Press; New York.

Barton, N. and Hewitt, G. (1985). Analysis of hybrid zones. Annual Review of Ecology and Systematics, 16:113–148.

Bertorelle, G. and Excoffier, L. (1998). Inferring admixture proportions from molecular data. Molecular Biology & Evolution, 15(10):1298–1311.

Brower, A. V. Z. (2011). Hybrid speciation in heliconius butterflies? a review and critique of the evidence. Genetica, 139(5):589–609.

DiCandia, M. and Routman, E. (2007). Cytonuclear discordance across a leopard frog contact zone. Molecular Phylogenetics and Evolution, 45(2):564–575.

Dittrich-Reed D.R. and Fitzpatrick B.M (2014). Transgressive Hybrids as Hopeful Monsters. Evolutionary Biology 40(2):310-315

Dowling, T. and Secor, C. (1997). The role of hybridization and introgression in the diversification of animals. Annual Review of Ecology and Systematics, 28:593–619.

Excoffier, L. (2004). Patterns of dna sequence diversity and genetic structure after a range expansion Lessons from the infinite-island model. Molecular Ecology, 13(4):853 - 865.

Hubisz, M., Falush, D., Stephens, M., and Pritchard, J. (2009). Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources, 9(5):1322–1332.

Hudson, A. G., Vonlanthen, P., Bezault, E., and Seehausen, O. (2013). Genomic signatures of relaxed disruptive selection associated with speciation reversal in whitefish. BMC Evolutionary Biology, 13.

Huff, C. D., Witherspoon, D. J., Simonson, T. S., Xing, J., Watkins, W. S., Zhang, Y., Tuohy, T. M., Neklason, D. W., Burt, R. W., Guthery, S. L.,Woodward, S. R., and Jorde, L. B. (2011). Maximum-likelihood estimation of recent shared ancestry (ersa). Genome Research, 21(5):768–774.

Kingman, J.F.C., (1982a). On the genealogy of large populations. Journal of Applied Probability 19, 27–43.

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

Lexer, C., Welch, M., Raymond, O., and Rieseberg, L. (2003). The origins of ecological divergence in helianthus paradoxus asteraceae Selection on transgressive characters in a novel hybrid habitat. Evolution, 57(9):1989–2000.

Lipson, M., Loh, P.-R., Levin, A., Reich, D., Patterson, N., and Berger, B. (2013). Efficient moment-based inference of admixture parameters and sources of gene flow. Molecular Biology and Evolution, 30(8):1788–1802.

Lohmueller, K. E., Bustamante, C. D., and Clark, A. G. (2010). The effect of recent admixture on inference of ancient human population history. Genetics, 185(2):611–622. 23

Mallet, J. (1995). A species definition for the modern synthesis. Trends in Ecology & Evolution, 10(7):294 – 299.

Manel, S., Gaggiotti, O., and Waples, R. (2005). Assignment methods: matching biological questions with appropriate techniques. Trends in Ecology & Evolution, 20(3):136 – 142.

Mavarez, J., Salazar, C., Bermingham, E., Salcedo, C., Jiggins, C., and Linares, M. (2006). Speciation by hybridization in heliconius butterflies. Nature, 441(7095):868–871.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. Journal of Mathematical Biology, 29(1):59–75.

Novembre, J. and Ramachandran, S. (2011). Perspectives on human population structure at the cusp of the sequencing era. Annual Review of Genomics and

Human Genetics, 12:245–274.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J., and Reich, D. (2004). Methods for high-density admixture mapping of disease genes. The American Journal of Human Genetics, 74(5):979– 1000.

Rannala, B. (1997). Gene genealogy in a population of variable size. Heredity, 78(4):417–423.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J. L., Schwarzbach, A. E., Donovan, L. A., and Lexer, C. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. Science, 301(5637):1211–1216.

Ropiquet, A. and Hassanin, A. (2006). Hybrid origin of the pliocene ancestor of wild goats. Molecular Phylogenetics and Evolution, 41(2):395 – 404.

Seehausen, O. (2004). Hybridization and adaptive radiation. Trends in Ecology & Evolution, 19(4):198–207.

Shriner, D., Adeyemo, A., Ramos, E., Chen, G., and Rotimi, C. (2011). Mapping of disease-associated variants in admixed populations. Genome Biology, 12(5):223.

Stebbins, G. (1959). The role of hybridization in evolution. Proceedings of the American Philosophical Society, 103(2):231 – 251.

Takahata, N. (1995). A genetic perspective on the origin and history of humans. Annual Review of Ecology and Systematics, 26:343–372.

Taylor, E., Boughman, J., Groenenboom, M., Sniatynski, M., Schluter, D., and Gow, J. (2006). Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (gasterosteus aculeatus) species pair. Molecular Ecology, 15(2):343–355.

Vollmer, S. V. and Palumbi, S. R. (2002). Hybridization and the evolution of reef coral diversity. Science, 296(5575):2023–2025.

Vrijenhoek, R. (2006). Polyploid hybrids Multiple origins of a treefrog species. Current Biology, 16(7):R245 – R247.

Wakeley, J. (1996). Pairwise differences under a general model of population subdivision. Journal of Genetics, 75(1):81–89.

Wakeley, J. (2008). Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village, Colorado.

Wakeley, J., King, L., Low, B. S., and Ramachandran, S. (2012). Gene genealogies within a fixed pedigree, and the robustness of kingman's coalescent. Genetics.

Wang, Y. and Hey, J. (2010). Estimating divergence parameters with small samples from a large number of loci. Genetics, 184(2):363–379.

Ward, J. L. and Blum, M. J. (2012). Exposure to an environmental estrogen breaks down sexual isolation between native and invasive species. Evolutionary applications, 5(8):901–912.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. Theoretical Population Biology, 7(2):256–276.

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.

Wilkinson-Herbots, H. M. (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(2):92-108.

Wilson, G. and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. Genetics, 163(3):1177–1191.

Witte, F., Seehausen, O., Wanink, J., Kishe-Machumu, M., Rensing, M., and Goldschmidt, T. (2013). Cichlid species diversity in naturally and anthropogenically turbid habitats of lake Victoria, east Africa. Aquatic Sciences, 75(2):169–183.

Appendix

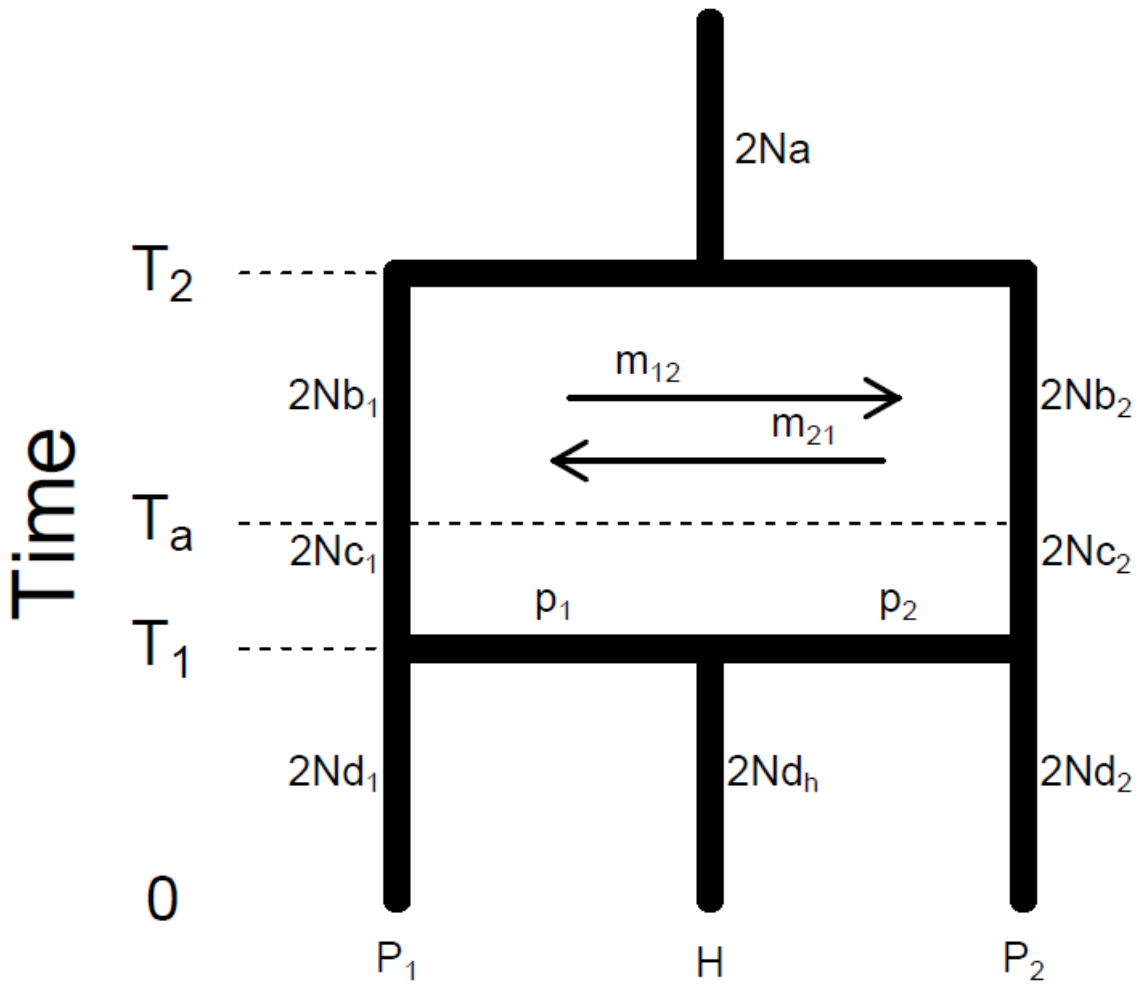

Figure 2-1. A general model considered in this paper. Ancestral population of size 2Na haploid individuals splits T_1 generations ago in two populations which differ in sizes. Two populations evolve in isolation until T_a generations ago when they start sharing migrants with different migration rates $m_{1,2}$ and $m_{2,1}$. At T_1, migration stops and a hybrid population is formed.

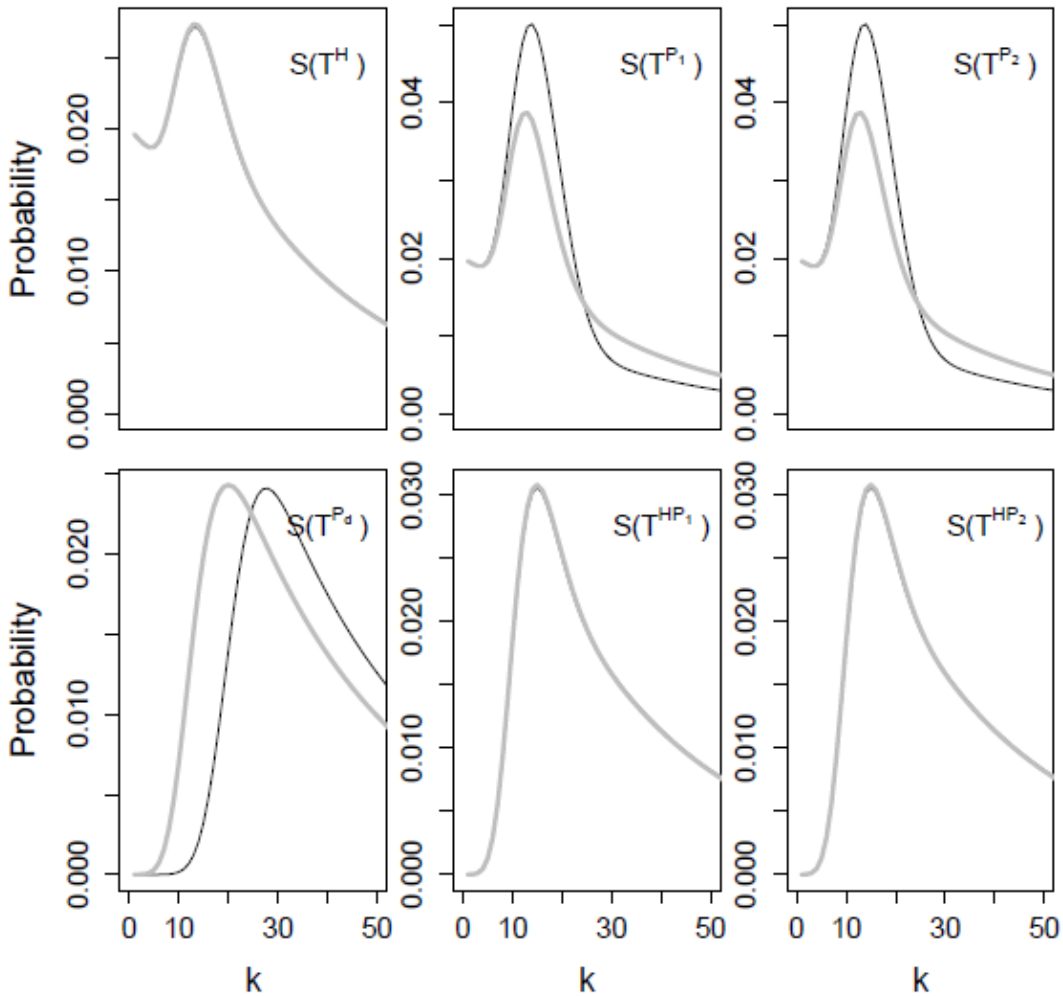

Figure 2-2 Sampling both parent populations is necessary to distinguish migration and population growth before hybridization. Both events can produce the same distribution of pairwise distributions for all pairs of genes involving hybrid population, as shown in this example. On the other hand, $S(T^{P_j})$ *and* $S(T_d^P)$ *do not depend on p and are thus different. Parameters: model with no migration (black):* $c_1 = c_2 = 1, b_1 = b_2 = 0.75, \tau_a = 1.5$, *model with migration(grey):* $M = 1$, *same in both models:* $\theta = 10, a = 3, d_1 = d_2 = 5, \tau_1 = 1, \tau_2 = 2, p = 0.5$.

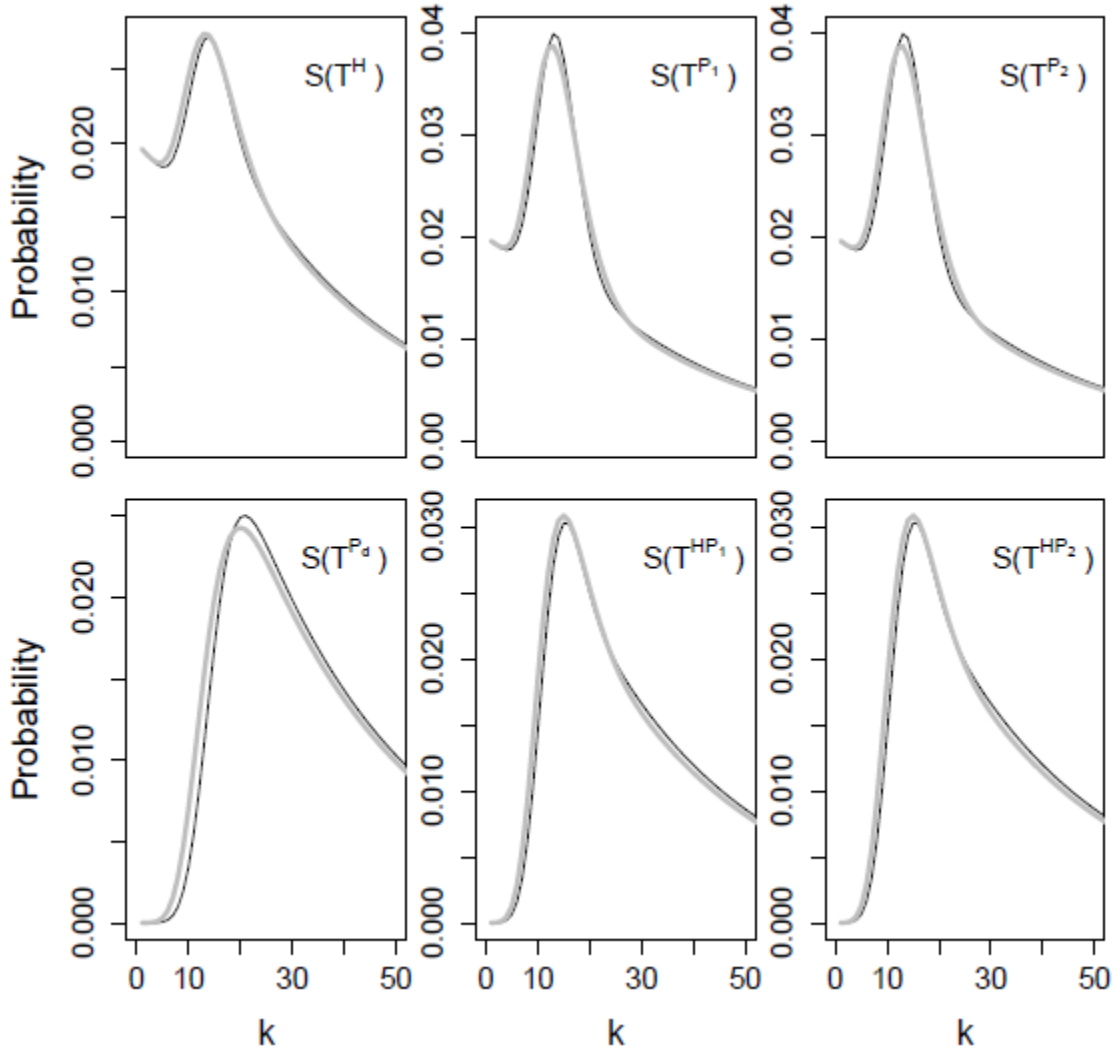

Figure 2-3 Even when all three populations are available distinguishing between migration and population change might be hard since both effects can result in similar distribution of pairwise differences. Parameters: No migration (black) model: $b_1 = 0.5, b_2 = 0.5, c_1 = 1.5, c_2 = 1.5, \tau_1 = 1.1, \tau_a = 1.25, \tau_2 = 1.4$, *migration model (grey),:* $\tau_1 = \tau_a = 1, \tau_2 = 2, M = 1$, *same in both models:* $\theta = 10, a = 3.0, p = 0.5, d_1 = 5, d_2 = 5, d_h = 5$

.

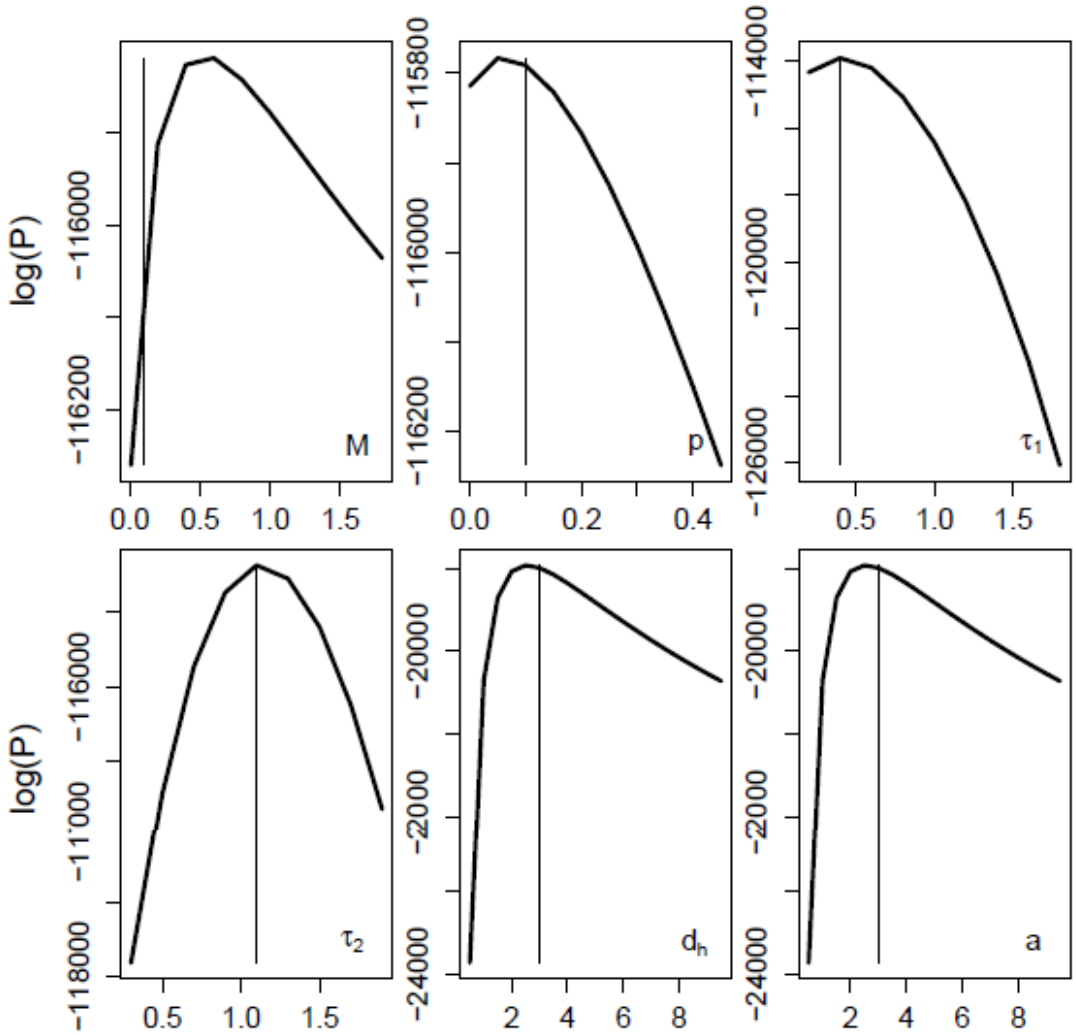

Figure 2-4 Marginal log likelihood functions. Model parameters $\theta = 5, \tau_1 = \tau_a = 0.4, \tau_2 = 1.1, a = 3, d_1 = 3, d_2 = 3, d_h = 3, p = 0.1, M = 0.1$. *Migration rate cannot be estimated precisely because the distribution of pairwise differences does not change much with changing M for this set of parameters.*

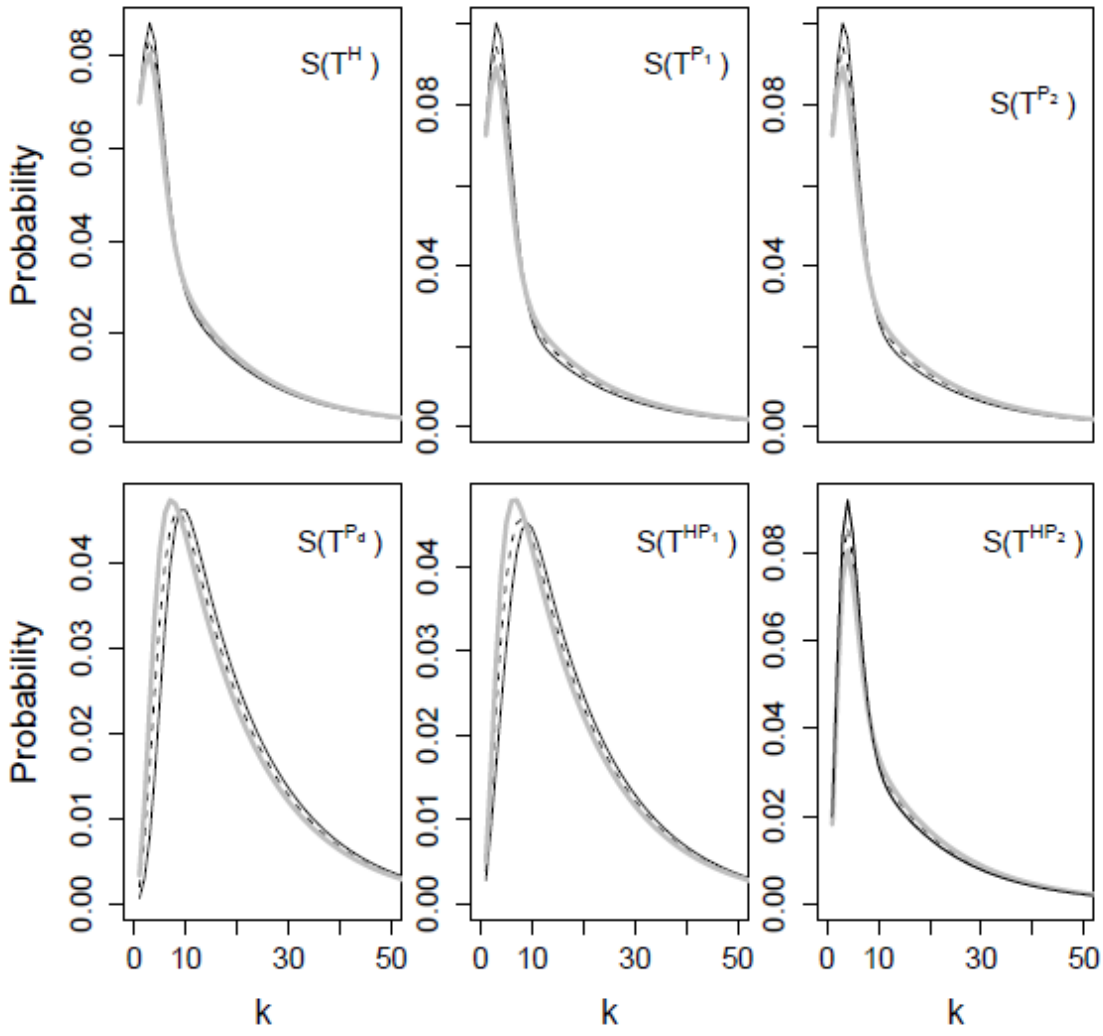

Figure 2-5. Effect of changing migration on the distribution of pairwise differences. For this parameter set, changing migration does not affect the distribution of coalescent times much. Model parameters $M = 0.1$ *(full line),* $M = 0.5$ *(dashed line),* $M = 1$ *(grey line).Other parameters:* $\theta = 5, \tau_1 = \tau_a = 0.4, \tau_2 = 1.1, a = 3, d_1 = 3, d_2 = 3, d_h = 3, p = 0.1$.

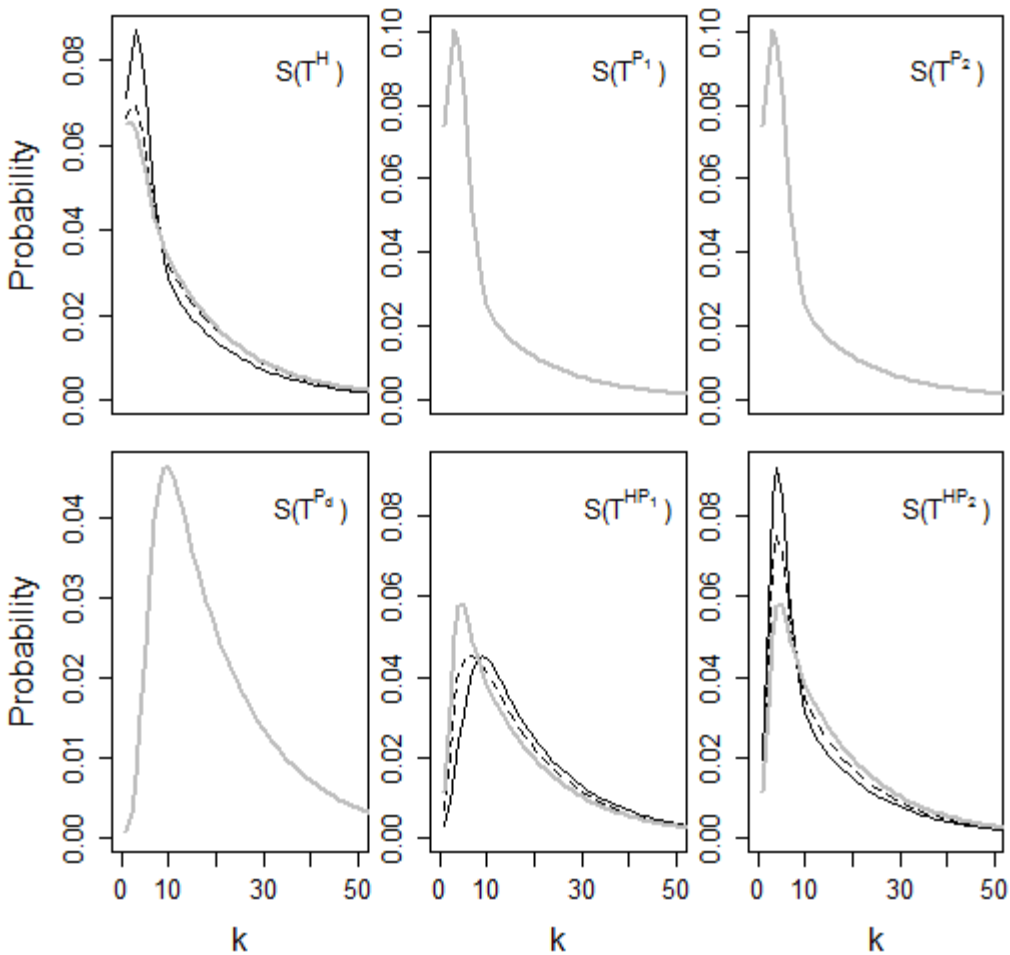

Figure 2-6.Effect of changing admixture coefficient on the distribution of pairwise differences. Changing the admixture coefficient changes the distribution of pairwise differences. Model parameters $p_1 = 0.1$ *(full line),* $p_1 = 0.3$ *(dashed line),* $p_1 = 0.5$ *(grey line). Distribution of pairwise differences does not depend on* p *when genes are sampled from parent populations which causes the three lines to overlap. Other parameters:* $\theta = 5, \tau_1 = \tau_a = 0.4, \tau_2 = 1.1, a = 3, d_1 = 3, d_2 = 3, d_h = 3, M = 0.1$.

Calculating expected values and their functions

Consider a case when two genes are sampled from parental population P_j.Then coalescent process is same as in *IMM* model of (Wilkinson-Herobots, 2012) with population change during time of isolation. Let *T* be a random variable denoting the coalescent time of two lineages. *T* can be written as a mixture of exponentially distributed random variables, X_j, W_j, Y_r and Z with means d_j, c_j, $1/\lambda_r$ and a respectively, and we can write:

$$fT_s^{P_j} = \sum_{r=1}^{2} A_{0r} f_{T_r} \tag{A2.1}$$

Where:

$$T_r = \begin{cases} X_j & \mathrm{X}_j \le \tau_1 \\ \tau_1 + W_j & \mathrm{X}_j > \tau_1 \text{ and } \tau_1 + \mathrm{W}_j \le \tau_\mathrm{a} \\ \tau_a + Y_r & \mathrm{X}_j > \tau_1 \text{ and } \tau_1 + \mathrm{W}_j > \tau_\mathrm{a} \text{ and } \tau_\mathrm{a} + \mathrm{Y}_r \le \tau_2 \\ \tau_2 + Z & \mathrm{X}_j > \tau_1 \text{ and } \tau_1 + \mathrm{W}_j > \tau_\mathrm{a} \text{ and } \tau_\mathrm{a} + \mathrm{Y}_r > \tau_2 \end{cases} \tag{A2.2}$$

Any function g of T_r is then:

$$g(T_r) = g(X_j) + I_{\{X>\tau_1\}}[g(\tau_1 + W_j) - g(X_j)] + I_{\{X>\tau_1, \tau_1+W>\tau_a\}}[g(\tau_a + Y_r) - g(\tau_1 + W_j)] + I_{\{X>\tau_1, \tau_1+W>\tau_a, \tau_a+Y>\tau_2\}}[g(\tau_2 + Z) - g(\tau_a + Y_r)] \tag{A2.3}$$

where I_A is an indicator variable that has value 1 if the event A occurred and 0 if it did not. Since X_j, W_j, Y_r and Z are exponentially distributed and independent for the expectation of T_r is:

$$
\begin{aligned}
E[g(T_r)] &= E[g(X_j)] + P(X_j > \tau_1)(E[g(\tau_1 + W_j)] - E[g(X_j) \mid X_j > \tau_1]) \\
&+ P(X_j > \tau_1)P(\tau_1 + W_j > \tau_a)(E[g(\tau_a + Y_r)] - E[g(\tau_1 + W_j) \mid \tau_1 + W_j > \tau_a]) \\
&+ (X_j > \tau_1)P(\tau_1 + W_j > \tau_a)P(\tau_a + Y_r > \tau_2)(E[g(\tau_2 + Z)] - E[g(\tau_a + Y_r) \mid \tau_a + Y_r > \tau_2]) \\
&= E[g(X_j)] + P(X_j > \tau_1)(E[g(\tau_1 + W_j)] - E[g(X_j + \tau_1)]) \\
&+ P(X_j > \tau_1)P(\tau_1 + W_j > \tau_a)(E[g(\tau_a + Y_r)] - E[g(\tau_a + W_j)]) \\
&+ (X_j > \tau_1)P(\tau_1 + W_j > \tau_a)P(\tau_a + Y_r > \tau_2)(E[g(\tau_2 + Z)] - E[g(\tau_2 + Y_r)])
\end{aligned}
$$

(A2.4)

We used the memoryless property of exponential random variable to obtain the last equality. Lastly, to obtain the expected value of a function of $T_s^{P_j}$ we use the relation:

$$E[g(T_s^{P_j})] = \sum_{r=1}^{2} A_{0r} E[g(T_r)]$$

(A2.5)

We can obtain the expressions for the probability of observing *l* pairwise differences for two genes by setting function $g_l(x) = e^{-\theta x}(\theta x)^l / l!$ in the equations (A2.4) and then using the equation (A2.5). Then, for an exponentially distributed random variable X with mean a we have:

$$F_1(a) = E[g_l(X)] = (a\theta)^l / (1 + a\theta)^{l+1}$$

(A2.6)

where $\theta = 4N\mu$. This result also follows from Watterson, (1975) who has shown that the distribution of pairwise differences in panmictic population of size *2N* is geometric with mean $1/(1+\theta)$. The probability of observing l pairwise differences during time $X + \tau$ is:

$$\mathrm{F}_2(a,t) = E[g_l(X+\tau)] = \frac{e^{\theta\tau_i}(a\theta)^l}{(1+a\theta)^{l+1}} \sum_{m=0}^{l} (1/a+\theta)^m \tau^m / m! \tag{A2.7}$$

Equation (A2.7) has been derived in (Wilkinson-Herbots, 2012). By applying equations (A2.6) and (A2.7) to (A2.5) we obtain equation (2.13).

Expected values and the distribution of pairwise differences for other 5 population pairs can be obtained in a similar way.

The expression for elements of e^{Qt} *:*

When migration is symmetric and population sizes are the same, matrix Q is:

$$Q = \begin{pmatrix} -1-M & M & 0 & 1 & 0 \\ M/2 & -M & M/2 & 0 & 0 \\ 0 & & -1-M & 0 & 1 \\ 0 & 0 & 0 & -M/2 & M/2 \\ 0 & 0 & 0 & M/2 & -M/2 \end{pmatrix} \tag{A2.8}$$

Then, the elements of matrix exponent e^{Qt} (equations 2.10a-e) are:

$$e_{1,1}^{Qt} = (1/4)(2e^{-(M+1)t}\sqrt{4M^2+1} + e^{-(1/2)(2M+1+\sqrt{4M^2+1})t}\sqrt{4M^2+1} + e^{-(1/2)(2M+1-\sqrt{4M^2+1})t}\sqrt{4M^2+1} + e^{-(1/2)(2M+1+\sqrt{4M^2+1})t} - e^{-(1/2)*(2M+1-\sqrt{4M^2+1})t)}) / \sqrt{4M^2+1} \tag{A2.9a}$$

$$e_{1,2}^{Qt} = M(e^{-(1/2)(2M+1-\sqrt{4M^2+1})t} - e^{-(1/2)(2M+1+\sqrt{4M^2+1})t}) / \sqrt{4M^2+1} \tag{A2.9b}$$

$$e_{1,3}^{Qt} = (1/4)(-2e^{-(M+1)t}\sqrt{4M^2+1} + e^{-(1/2)(2M+1+\sqrt{4M^2+1})t}\sqrt{4M^2+1} + e^{-(1/2)(2M+1-\sqrt{4M^2+1})t}\sqrt{4M^2+1} + e^{-(1/2)(2M+1+\sqrt{4M^2+1})t} - e^{-(1/2)*(2M+1-\sqrt{4M^2+1})t)})/\sqrt{4M^2+1}$$

(A2.9c)

$$e_{2,1}^{Qt} = M(e^{-(1/2)(2M+1-\sqrt{4M^2+1})t} - e^{-(1/2)(2M+1+\sqrt{4M^2+1})t})/2\sqrt{4M^2+1}$$

(A2.9d)

$$e_{2,2}^{Qt} = (1/2)(e^{-(1/2)(2M+1-\sqrt{4M^2+1})t}\sqrt{4M^2+1} + e^{-(1/2)(2M+1+\sqrt{4M^2+1})t}\sqrt{4M^2+1} + e^{-(1/2)(2M+1-\sqrt{4M^2+1})t} - e^{-(1/2)(2M+1+\sqrt{4M^2+1})t})/\sqrt{4M^2+1}$$

(A2.9e)

$$e_{2,3}^{Qt} = e_{2,1}^{Qt}$$

(A2.9f)

Chapter 3

Serial Founder Model with historical migration

Abstract

Recently, DeGiorgio et al.(2011) have obtained a closed form expression for the distribution of coalescent times and several related statistics for "serial founder model" (SF model). Their model does not include migration and analytical results concerning migration in SF models are lacking. Here we study the effects of historical migration in SF models. We derive a closed form expression for the distribution of coalescent times and the distribution of pairwise differences under infinite site mutation models. We find that coalescent times for two genes sampled from the same population are longer when migration is incorporated into the model. Longer coalescent times cause slower decay of heterozygosity in a migration model. Heterozygosity can even increase with distance from the source population. Additionally, the pairwise F_{st} can decrease with distance from the oldest population.

Introduction

The "serial founder model" (SF) is a nonequilibrium population models used to describe the spread of humans from Africa across the world (Ramachandran et al., 2005). In this model, a small number of individuals from an initial (source) population move to a new geographic region and form a second population that grows to carrying capacity. A group of individuals from this second population then moves to a new geographic region

forming a third population. This process is repeated until *n* populations are formed. Each new population passes through a genetic bottleneck during its formation.

Several variations of SF and related territory expansion models have been studied extensively using simulations (Ramachandran et al., 2005, Deshpande et al., 2009, DeGiorgio et al., 2009, Hunley et al., 2009) and analytical approaches (Austerlitz et al., 1997, Liu et al., 2006, Excoffier and Ray, 2008, Slatkin and Excoffier, 2012, Nullmeier and Hallatschek, 2013).

Recently, DeGiorgio et al. (2011) have studied a SF model and found closed-form expressions for the distribution of coalescent times, expected coalescent time, expected heterozygosity and pairwise F_{st}. SF model of DeGiorgio et al. (2011) produced linear decay in heterozygosity with respect to geographical distance from the source population and increase in F_{st} between distant populations. With some exceptions, patterns produced by their model are consistent with patterns observed in human data.

However, this model by DeGiorgio et al. (2011) is limited by its lack of continued migration between populations. Although simulation results suggest that small to moderate migration does not affect the patterns produced by SF model (DeGiorgio et al., 2009), analytical results concerning migration in SF models are lacking. Here, we seek to understand how migration affects patterns of heterozygosity and pairwise F_{st} in SF model. To that end, we incorporate historical migration in the SF model of (DeGiorgio et al., 2009) and calculate the distribution of pairwise differences. Since gene identity and heterozygosity are special cases when the number of differences between

two genes is equal to zero or non-zero respectively, we also expand SF model with no migration by finding an expression for the distribution of pairwise differences (mismatch distribution).

Model

Our main goal is to derive the distribution of coalescent times and pairwise differences in the serial founder model with historical migration (figure 3.1) and compare it to a model with no migration. To make comparison easier, we also consider a serial founder model with no migration to obtain simpler expressions than those in (DeGiorgio et al, 2011). Through the paper, we assume that genealogies of two genes can be described in terms of Kingman's or structured coalescent (Kingman 1982a,b, Notohara, 1990). By gene we mean a selectively neutral sequence of non-recombining DNA which mutates according to infinite site mutation model (Watterson, 1975).

No migration

In this model all migration rates are equal to zero. We scale model parameters by population size *2N*. One time unit now corresponds to *2N* generations and $a_1, a_2, ..., a_{2n}$ correspond to relative population sizes (figure 3.1). There are *n* populations in a model, each of which changes size once. However, it is more convenient to think of a model as consisting of *2n* populations of constant size because then a population j has size a_j .

Furthermore, every odd-numbered population *k* cannot be sampled and it exists only between τ_k and τ_{k-1}.

From remaining *n* even-numbered populations, a pair of genes can be sampled in $n(n+1)/2$ different ways if sampling order does not matter. To fully describe the coalescent we only need to distinguish two different ways in which genes can be sampled. Two genes can be sampled from the same population and or from different populations. Then, the coalescent process for a pair of genes in serial founder model with no migration can be modeled as a modification of "complete isolation" model of (Takahata, 1995), where the modification is population size change after the period of isolation.

Wilkinson-Herbots (2012) has shown how to use indicator variable to obtain the expression from which the distribution of coalescent times and pairwise differences can be easily calculated in complex models. Following that method, we write a random variable T_{jj} denoting the coalescent time of two genes sampled from the population j, as a combination of j random exponentially distributed variables with means $a_j, a_{j-1}, a_{j-2}, ..., a_1$. That follows because ancestral lineages of two genes sampled from population *j* can coalesce in each population preceding population *j*. We can write T_{jj} as:

$$T_{jj} = \begin{cases} X_j & \text{if } \mathrm{X_j} \le \tau_{\text{j-1}} \\ X_{j-1} + \tau_{j-1} & \text{if } \tau_{\text{j-1}} \le \mathrm{X_j} \text{ and } \mathrm{X_{j-1}} + \tau_{\text{j-1}} \le \tau_{\text{j-2}} \\ ... & \\ X_1 + \tau_1 & \text{if } \tau_{\text{j-1}} \le \mathrm{X_j} \text{ and } \tau_{\text{j-2}} \le \mathrm{X_{j-1}} + \tau_{\text{j-1}} ... \text{ and } \tau_1 \le \mathrm{X_2} + \tau_2 \end{cases}$$

(3.1)

We use indicator variable I_A, to obtain the expression for any function g of T_{jj}. Let I_A have a value 1 if the event happened and 0 if the event did not happen. Then, given equation (3.1) we can write:

$$g(T_{jj}) = g(X_j) + I_{X_j > \tau_{j\text{-}1}}(g(X_{j\text{-}1} + \tau_{j\text{-}1}) - g(X_j)) + I_{X_j > \tau_{j\text{-}1}, X_{j\text{-}1} > \tau_{i\text{-}2}}(g(X_{j\text{-}2} + \tau_{j\text{-}2}) - g(X_{j\text{-}1} + \tau_{j\text{-}1}))$$
$$+ \ldots + I_{X_j > \tau_{j\text{-}1}, X_{j\text{-}1} > \tau_{i\text{-}2}, \ldots, X_2 > \tau_1}(g(X_2 - \tau_2) - g(X_1 - \tau_1))$$

(3.2)

Using the fact that variables are independent and that exponentially distributed random variable has a memoryless property, we obtain the expectation of $g(T_{jj})$ as:

$$E[g(T_{jj})] = E[g(X_j)] + P(X_j > \tau_{j-1})(E[g(X_{j-1} + \tau_{j-1})] - E[g(X_j + \tau_{j-1})])$$
$$+ P(X_j > \tau_{j-1})P(X_{j-1} > \tau_{j-2})(E[g(X_{j-2} + \tau_{j-2})] - E[g(X_{j-1} + \tau_{j-2})])$$
$$+ \ldots + \prod_{i=2}^{j} P(X_i > \tau_{i-1})(E[g(X_1 + \tau_1)] - E[g(X_2 + \tau_1)])$$

(3.3)

From the equation above we can get the expression for the expected coalescent time for two genes from the same population:

$$E[T_{jj}] = a_j + e^{-\tau_{j-1}/a_j}(a_{j-1} - a_j) + e^{-\tau_{j-1}/a_j}\sum_{k=2}^{j-1}(a_{k-1} - a_k)e^{-\sum_{l=k}^{j-1}(\tau_{l-1} - \tau_l)/a_l}$$

(3.4)

$E[T_{jj}]$ can therefore be written as the expectation under complete isolation model (first two terms on the right side of equation (3.4)) and the summation term that represents the effect of repeated bottlenecks.

When two genes are sampled from different populations, say population j and $k, j < k$, we can represent the coalescent time T_{jk} as a series of j exponentially distributed random variables similarly as we did for T_{jj}. We obtain:

$$E[g(T_{jk})] = E[X_j + \tau_j] + P(X_j + \tau_j > \tau_{j-1})(E[g(X_{j-1} + \tau_{i-1})] - E[g(X_j + \tau_{j-1})]) + \ldots + \prod_{i=2}^{j} P(X_i + \tau_i > \tau_{i-1})(E[g(X_1 + \tau_1)] - E[g(X_2 + \tau_1)])$$

(3.5)

The expected coalescent time for two genes sampled from different populations, $E[T_{jk}]$, $j < k$, is then:

$$E[T_{jk}] = \tau_j + a_j + \sum_{l=2}^{j} (a_{l-1} - a_l) e^{-\sum_{m=l}^{j} (\tau_{m-1} - \tau_m)/a_m}$$

(3.6)

$E[T_{jk}]$ can also be written in terms of "complete isolation" model and a bottleneck term. $E[T_{jk}]$ does not depend on population k, because the two lineages can coalesce only when in the same population, which happens at and after τ_j for all k's.

Next, we derive the distribution of pairwise differences in the serial founder model assuming an infinite site mutation model (Watterson, 1975). In infinite site mutation model, the appearance of new mutations follows Poisson distribution with mean μ, and each mutation produces a new polymorphism. We define the function $g_l(x) = e^{-\theta x}(\theta x)^l / l!$. For an exponentially distributed random variable X_i with mean a_i we have:

$$A(a_i, l, \theta) = E[g_l(X_i)] = (a_i\theta)^l / (1 + a_i\theta)^{l+1} \tag{3.7}$$

where $\theta = 4N\mu$. We also find that the probability of having l pairwise differences during time $X_i + \tau_i$ is:

$$B(a_i, l, \theta, t_j) = E[g_l(X_i + \tau_j)] = \frac{e^{\theta\tau_i}(a_i\theta)^l}{(1 + a_i\theta)^{l+1}} \sum_{m=0}^{l} (1/a_i + \theta)^m \tau_j{}^m / m! \tag{3.8}$$

Equation (3.7) follows from Watterson, (1975) who has shown that the distribution of pairwise differences in panmictic population of size *2N* is geometric with mean $1/(1+\theta)$, while equation (3.8) can be found in (Wilkinson-Herbots, 2012).

We can now write the equation for the probability that two genes sampled from population j are different in l sites $P(S_{jj}=l)$ from equation (3.3) as:

$$E[g_l(T_{jj})] = P(S_{jj}=l) = A(a_j,l,\theta) + e^{\tau_{j-1}/a_j}(B(a_{j-1},l,\theta,t_{j-1}) - B(a_j,l,\theta,t_{j-1}))$$
$$+e^{-\tau_{j-1}/a_j}\sum_{k=2}^{j-1}(B(a_{k-1},l,\theta,t_{k-1}) - B(a_k,l,\theta,t_{k-1}))e^{-\sum_{m=k}^{j-1}(\tau_{l-1}-\tau_l)/a_l}$$

(3.9)

Similarly, from (3.5) we obtain the following equation:

$$E[g_l(T_{jk})] = P(S_{jk}=l) = B(a_j,l,\theta,t_j) +$$
$$\sum_{k=2}^{j}(B(a_{k-1},l,\theta,t_{k-1}) - B(a_k,l,\theta,t_{k-1}))e^{-\sum_{m=k}^{j}(\tau_{m-1}-\tau_m)/a_m}$$

(3.10)

Historical migration

Migration can be incorporated in serial founder model in different ways. The simplest and analytically tractable way to introduce migration is by considering a historical migration model (figure 3.1). In this model, populations j and $j+2$ share migrants between τ_{j+2} and τ_{j+1}. This model might roughly correspond to case in which the loss of contact with old population results in the formation of a new population.

For simplicity, we assume that all non-bottleneck populations are of the same sizes *2N* and all bottleneck populations of size $2Nb, b < 1$.

Before describing the distribution of coalescent times in this model with migration, we need to derive certain results concerning coalescent with migration. Notohara (1990) described the coalescent process for a sample of genes from populations exchanging migrants as a continuous Markov process with rate matrix Q (also see chapter 4 in Wakeley (2008)).

In our case matrix Q has 5 states: two lineages in the first population, one lineage in each population, two lineages in the second population and one lineage in first or second population after coalescent. When population sizes are equal to *2N* and migration is symmetrical and equal to m, Q is:

$$Q = \begin{pmatrix} -1-M & M & 0 & 1 & 0 \\ M/2 & -M & M/2 & 0 & 0 \\ 0 & & -1-M & 0 & 1 \\ 0 & 0 & 0 & -M/2 & M/2 \\ 0 & 0 & 0 & M/2 & -M/2 \end{pmatrix} \tag{3.11}$$

where $M = 4Nm$.

By calculating a matrix exponent of Q, $e^{Qt} = \sum_{k=0}^{\infty} (Qt)^k / k!$, we obtain the probability of system being in state j after time t given it started in state i. The relevant entries of e^{Qt} can be written as (Juric, unpublished results, also see Wilkinson-Herbots(2012)):

$$e_{1,1}^{Qt} = \frac{1}{2}(\sum_{r=1}^{2} A_{0r}\lambda_r e^{-\lambda_r t} + e^{-(M+1)t})$$

(3.12a)

$$e_{1,2}^{Qt} = \sum_{r=1}^{2} A_{1r}\lambda_r e^{-\lambda_r t}$$

(3.12b)

$$e_{1,3}^{Qt} = \frac{1}{2}(\sum_{r=1}^{2} A_{0r}\lambda_r e^{-\lambda_r t} - e^{-(M+1)t})$$

(3.12c)

$$e_{2,1}^{Qt} = e_{2,3}^{Qt} = \frac{1}{2} e_{1,2}^{Qt}$$

(3.12d)

$$e_{2,2}^{Qt} = A_{01}\lambda_1 e^{-\lambda_2 t} + A_{02}\lambda_2 e^{-\lambda_1 t}$$

(3.12e)

where: $\lambda_1 = \frac{M + 1/2 - \sqrt{D}}{2}$, $\lambda_2 = \frac{M + 1/2 + \sqrt{D}}{2}$, $D = 4M^2 + 1$ and $A_{01} = \frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}$,

$A_{02} = \frac{1 - \lambda_1}{\lambda_2 - \lambda_1}$, $A_{11} = \frac{\lambda_2}{\lambda_2 - \lambda_1}$ and $A_{12} = \frac{-\lambda_1}{\lambda_2 - \lambda_1}$.

Consider two genes sampled from population j. Up to τ_{j+2} both lineages are in the population j and because the population size is *2N*, they coalesce with rate 1. If lineages do not coalesce, they enter first migration-bottleneck block. A migration-bottleneck block is a time period in which two adjacent populations share migrants followed by a time period during which one of the populations is experiencing bottleneck. Looking forward in time, migration-bottleneck block is a bottleneck period during founding of a new population followed by migration from adjacent population after a new population grew in size.

When two lineages enter the first migration-bottleneck block, migration can move lineages between populations j and $j+2$ from time τ_{j+2} to τ_{j+1}. Assuming no coalescent, three mutually exclusive outcomes are possible at τ_{j+1}. 1) both lineages stay in population j, 2) both lineages move to population $j+2$ between τ_{j+2} to τ_{j+1}, and are now in the bottleneck population $j+1$ 3) one lineage is in the population $j+1$ while the other stays in j. If both lineages enter the bottleneck population they coalesce with rate *b*. If lineages remain in the population *j*, they coalesce with rate 1. If lineages are in the different populations, coalescent is not possible until the exit from migration-bottleneck block at time τ_j.

We can write the distribution of coalescent times until the end of the first migration-bottleneck block as:

$$f_{T_{jj}^{M}}(t)=\begin{cases} e^{-t} & 0\le t\le \tau_{j+2} \\ e^{-\tau_{j+2}}(e_{1,1}^{Q(t-\tau_{j+2})}+e_{1,3}^{Q(t-\tau_{j+2})}) & \tau_{j+2}<t\le \tau_{j+1} \\ e^{-\tau_{j+2}}(e_{1,1}^{Q\tau_M}e^{-(t-\tau_{j+1})}+e_{1,3}^{Q\tau_M}e^{-(t-\tau_{j+1})/b}/b) & \tau_{j+1}<t\le \tau_{j} \end{cases}$$

(3.13)

where $\tau_M=\tau_{j+1}-\tau_{j+2}$ is the duration of migration period.

The probability that coalescent did not happen by the time the first migration-bottleneck block ended is:

$$X_1=e^{-\tau_{j+2}}(e_{1,1}^{Q\tau_M}e^{-\tau_b}+e_{1,3}^{Q\tau_M}e^{-\tau_b/b}+e_{1,2}^{Q\tau_M})$$

(3.14)

where $\tau_b=\tau_j-\tau_{j+1}$ is the duration of bottleneck period.

In the time between τ_j and τ_2 there will be another ($j/2-1$) migration-bottleneck blocks. Unlike the first block, in each of the following blocks lineages enter a bottleneck population unless they move during migration period. The probability of no coalescent during one of the remaining $(j/2-1)$ migration-bottleneck blocks is:

$$Y=e_{1,1}^{Q\tau_M}e^{-\tau_b/b}+e_{1,3}^{Q\tau_M}e^{-\tau_b}+e_{1,2}^{Q\tau_M}$$

(3.15)

We can write the distribution of coalescent times between τ_j and τ_2 as:

$$f_{T_{jj}^M}(t) = \begin{cases} X_1 Y^{l/2}(e_{1,1}^{Q(t-\tau_{j-l})} + e_{1,3}^{Q(t-\tau_{j-l})}) & \tau_{j\text{-}l} \le t \le \tau_{j\text{-}l\text{-}1} \\ X_1 Y^{l/2}(e_{1,1}^{Q\tau_M} e^{-(t-\tau_{j-l-1})} + e_{1,3}^{Q\tau_M} e^{-(t-\tau_{j-l-1})/b} / b) & \tau_{j\text{-}l\text{-}1} < t \le \tau_{j\text{-}l\text{-}2} \end{cases}$$

(3.16)

where $l = 0, 2, 4, 6, 8, 10 ... j-4$.

Lastly, there is one last bottleneck between τ_2 and τ_1:

$$f_{T_{jj}^M}(t) = \begin{cases} X_1 Y^{j/2-1} e^{-(t-\tau_2)} & \tau_2 \le t \le \tau_1 \\ X_1 Y^{j/2-1}(1/b) e^{-(\tau_1-\tau_2)} e^{-(t-\tau_1)/b} & \tau_l < t \end{cases}$$

(3.17)

The first migration-bottleneck block is also different from the others when two genes are sampled from populations j and k, $j < k$. Two lineages enter the first migration-bottleneck block from different populations, while all subsequent from the same population. Therefore, $f_{T_{jk}^M}$ up until the end of first migration-bottleneck block can be written as:

$$f_{T_{jk}^M}(t) = \begin{cases} 0 & 0 \le t \le \tau_{j+2} \\ e_{2,1}^{Q(t-\tau_{j+2})} + e_{2,3}^{Q(t-\tau_{j+2})} & \tau_{j+2} < t \le \tau_{j+1} \\ e_{2,1}^{Q\tau_M} e^{-(t-\tau_{j+1})/a} / a + e_{2,3}^{Q\tau_M} e^{-(t-\tau_{j+1})/b} / b & \tau_{j+1} < t \le \tau_j \end{cases}$$

(3.18)

The probability of no coalescent till τ_j is:

$$X_2 = e_{2,1}^{Q\tau_M} e^{-\tau_b/a} + e_{2,3}^{Q\tau_M} e^{-\tau_b/b} + e_{2,2}^{Q\tau_M}$$

(3.19)

After τ_j, coalescent is the same as when two genes are sampled from the same population, and is described by equations (3.15) and (3. 16) and replacing X_1 with X_2.

In a way similar to the one for model with no migration, we obtain a general expression for an expectation of a function g of coalescent times when two genes are sampled from the same population as:

$$g(E[T_{jj}^M]) = A_E + e^{-\tau_{j+2}}(B_{E1}(\tau_{j+2},\tau_{j+1}) + C_{E1}(\tau_{j+1},\tau_j) + X_1(\sum_l Y^{l/2}(B_E(\tau_{j-l},\tau_{j-l-1}) + D_E(\tau_{j-l-1},\tau_{j-l-2}))$$
$$+Y_E^{j/2-1}(E_E(\tau_2,\tau_1) + F_E(\tau_1)))$$

(3.20)

where $l = 0,2,4,6,8,10...j-4$. Expected coalescent times and the probability of l pairwise differences can be obtain by replacing terms on the right side with appropriate expressions listed in appendix .

Similarly, when two genes are sampled from different populations we obtain:

$$g(E[T_{jk}^M]) = B_{E2}(\tau_{j+2},\tau_{j+1}) + C_{E2}(\tau_{j+1},\tau_j) + X_2(\sum_l Y^{l/2}(B_E(\tau_{j-l},\tau_{j-l-1}) + D_E(\tau_{j-l-1},\tau_{j-l-2}))$$
$$+Y_E^{j/2-1}(E_E(\tau_2,\tau_1) + F_E(\tau_1))$$

(3.21)

Discussion

Distribution of coalescent times

Probability density function of coalescent times in the model with historical migration can have more complex shape compared to model with no migration (figure 3.2). In the example on figure 3.2, the density of coalescent times when one gene is sampled from the oldest population and other from a different population, ($f_{T^M_{2,k}}$) increases monotonically between T = 4 and 5 in a model with migration. In a model without migration, there are no time periods between which the density of coalescent times increases monotonically. The explanation of this difference between models with and without migration is following. When migration started at T = 4, lineages were in different populations. Lineages need to be in the same population to coalesce, the probability of which increases with time. In a model with no migration, coalescent between is described by Kingman's coalescent, which means that the density of coalescent times is follows exponential distribution, therefore it is decreasing with time. We also note that with the exception of first migration period, during each subsequent migration-bottleneck block, two lineages will initially be in the same population, and the migration will move them apart, thus decreasing the probability of coalescent. That is the reason why only during first migration period $f_{T^M_{j,k}}$ grows continuously.

Migration can mitigate the effect of bottleneck in two ways. Lineages can be in non-bottleneck population during bottleneck time, in which case they will coalesce with slower rate. Lineages can also be in different populations during the bottleneck time in

which case they cannot coalesce. Both ways have the effect of decreasing probability of coalescent during bottlenecks. This effect is better seen when comparing $f_{T^M_{j,j}}$ for two models (figure 3.3) when migration moved lineages between different population so much that $f_{T^M_{2,2}}$ and $f_{T^M_{6,6}}$ look almost undistinguishable.

Expected coalescent times

In the model with no migration, differences in population sizes, duration and the number of bottlenecks affect the expected coalescent times. Large population size increases coalescent time while bottlenecks decrease it. When genes are sampled from populations distant to the source population (population 2), their ancestral lineages will have more opportunities to experience bottlenecks causing shorter coalescent times compared to when two lineages are sampled from populations close to the source population.

With only two different population sizes (bottleneck and post-bottleneck sizes of sizes a_2 and a_1 respectively), as is in simulations of DeGiorgio et al., (2011), equation (3.4) tells us that the expected coalescent time of two genes sampled from younger population will always be shorter than when genes are sampled from older population due to the effects of multiple bottlenecks.

In the model with historical migration, the expected coalescent time of two genes sampled from distant population will be larger than in model with no migration. That happens because migration can split lineages to different populations thus delaying

coalesce. This effect can be so strong that for some model parameters, genes sampled from distant populations will have larger expected coalescent times than genes from source population. For example, when the duration of bottleneck is $\tau_b = 0.1$, duration of migration period is $\tau_M = 0.5$, migration rate $M = 1$ and the bottleneck population size $b = 0.1$, $E[T_{2,2}^M] = 0.85$ and $E[T_{4,4}^M] = 0.92$.

On the other hand, the expected coalescent time of two genes sampled from different populations in migration model is shorter than in model with no migration. That can easily be seen since with migration two lineages can coalesce after τ_{j+2}, while when $M = 0$, coalescent is possible only after τ_j.

Heterozygosity

By setting l to the zero in equations (3.6) we obtain the expressions for gene identity (J_{jj}) in model with no migration.

$$P(S_{jj} = 0) = J_{jj} = \frac{1}{1 + a_j\theta} + e^{\tau_{j-1}/a_j}\left(\frac{e^{-\theta\tau_{j-1}}}{1 + a_{j-1}} - \frac{e^{-\theta\tau_{j-1}}}{1 + a_j}\right) + e^{\tau_{j-1}/a_j}\sum_{k=2}^{j-1}\left(\frac{e^{-\theta\tau_k}}{1 + a_{k-1}} - \frac{e^{-\theta\tau_k}}{1 + a_k}\right)e^{\sum_{l=k}^{j-1}(T_{l-1} - T_l)/a_l} \tag{3.22}$$

Heterozygosity is calculated as one minus gene identity. With only two different population sizes (bottleneck and post-bottleneck sizes of sizes a_2 and a_1 respectively), second and third terms on the right side of equation (3.22) are positive, since $1/(1+a_2) < 1/(1+a_1)$. By comparing equation (3.22) to (3.5) we can see that gene identity decreases with increasing expected coalescent time. Therefore, all conclusions about expected coalescent times translate to heterozygosity. Namely, distant populations will have lower heterozygosity compared to ones close to the source. Whether the heterozygosity decrease is linear depends on the population sizes and times and duration of bottlenecks.

If bottlenecks happened long time ago, such that the first term in equation (3.22) dominates, heterozygosity will entirely be defined by the scaled mutation rate ($\theta = 4N\mu$). Then for all populations $H \approx \theta/(1+\theta)$, which is the same as in unstructured population (Watterson, 1975, also see figure 9E in DeGiorgio et al. 2011).

In a model with migration, heterozygosity can decrease or increase with the distance from the source population depending on the relative strength of migration and bottlenecks effects (figure 3.5). For the top plot on figure (3.5), we used the parameters as in DeGiorgio et al. (2011), while for the bottom plot we extended the time between bottlenecks from 19 generations to 200 generations (corresponding to changing τ_M from 0.00095 to 0.01) . When the timing between bottlenecks is short, heterozygosity patterns are similar in both models. However, when migration lasts longer, distant populations are more diverse than ones close to the source.

Pairwise F_{st}

Understanding how $E[T_{jj}]$ and $E[T_{jk}]$ compare between models allow us to understand spatial patterns of pairwise F_{ST}. Pairwise F_{ST} between populations j and k is defined as (Slatkin, 1991, DeGiorgio et al., 2011):

$$F_{ST} = \frac{E[T_{jk}] - 0.5(E[T_{jj}] + E[T_{kk}])}{E[T_{jk}] + 0.5(E[T_{jj}] + E[T_{kk}])} \tag{3.23}$$

In a model with no migration and two different population sizes F_{st} increases when j is kept constant and k increases (figure 3.5). That is because $E[T_{kk}]$ decreases while all other terms in equation (3.23) remain the same. The result is the decrease of F_{st} with distance from population *j*. When j increases, F_{st} decreases because τ_j decreases faster than the bottleneck term increases in equation (3.4) causing F_{st} between a pair of distant populations to be smaller than between populations closer to the source.

Since F_{st} is a function of expected coalescent times, it is not surprising that in a model with migration it can increase or decrease with distance from the source population (figure 3.5). Again, this is because with migration $E[T_{kk}]$ can be larger than $E[T_{jj}]$.

Conclusion

We have examined the effects of historical migration on serial founder model. We derived the expressions for the distribution of coalescent times and pairwise differences, as well as the expression for the expected coalescent times. We used those expressions to understand the effects of migration on patterns of heterozygosity and F_{st}.

Migration can offset the effects of repeated bottlenecks by increasing coalescent times for two genes sampled from the same population and decreasing coalescent time for two genes sampled from different populations. Longer coalescent times of genes sampled from the same populations causes slower heterozygosity decay in a migration model. In fact, heterozygosity can increase with distance from the source population. Another consequence of altered coalescent times in migration model is the smaller pairwise F_{st} compared to the model with no migration. In a model with migration F_{st} can decrease with distance from the source population. However, increasing heterozygosity or decreasing F_{st} are not a unique signature of historical migration in serial founder model because they can be obtained in a model with no migration when the distant populations' sizes are bigger than population sizes closer to the source population.

Model with no migration has been used to describe human spread around the globe (DeGiorgio et al., 2011). For parameters considered by those researchers, introducing historical migration produces qualitatively same results (top of figures 3.4 and 3.5) even when migration is high, suggesting that it might be hard if not impossible to detect historical migration as humans conquered the world based on patterns of heterozygosity

and pairwise F_{st}. In theory, it might be possible to use equations (3.9, 3.10, 3.20, 3.21) to calculate the likelihood of pairwise differences based on whole genome scans under different models and compare different models. However, given the large number of parameters, we are bit skeptical about the results of such analysis.

References

Austerlitz, F., Jung-Muller, B., Godelle, B., and Gouyon, P.-H. (1997). Evolution of coalescence times, genetic diversity and structure during colonization. Theoretical Population Biology,51(2):148 – 164.

DeGiorgio, M., Degnan, J., and Rosenborg, N. (2011). Coalescence-time distributions in a serial founder model of human evolutionary history. Genetics, 189(2):579–593.

DeGiorgio, M., Jakobsson, M., and Rosenberg, N. A. (2009). Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proceedings of the National Academy of Sciences, 106(38):16057–16062.

Deshpande, O., Batzoglou, S., Feldman, M. W., and Luca Cavalli-Sforza, L. (2009). A serial founder effect model for human settlement out of Africa. Proceedings of the Royal Society B: Biological Sciences, 276(1655):291–300.

Excoffier, L. and Ray, N. (2008). Surfing during population expansions promotes genetic revolutions and structuration. Trends in Ecology & Evolution, 23(7):347 – 351.

Hunley, K. L., Healy, M. E., and Long, J. C. (2009). The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: Implications for biological race. American Journal of Physical Anthropology, 139:35–46.

Liu, H., Prugnolle, F., Manica, A., and Balloux, F. (2006). A geographically explicit genetic model of worldwide human-settlement history. The American Journal of Human Genetics, 79(2):230– 237.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. Journal of Mathematical Biology, 29(1):59–75.

Nullmeier, J. and Hallatschek, O. (2013). The coalescent in boundary-limited range expansions. Evolution, 67(5).

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences, 102(44):15942–15947.

Slatkin, M., (1991). Inbreeding coefficients and coalescence times. Genetical Research 58: 167–175.

Slatkin, M. and Excoffier, L. (2012). Serial founder effects during range expansion: A spatial analog of genetic drift. Genetics, 191(1):171–181.

Takahata, N. (1995). A genetic perspective on the origin and history of humans. Annual Review of Ecology and Systematics, 26:343–372.

Wakeley, J. (2008). Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village, Colorado.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. Theoretical Population Biology, 7(2):256–276.

Wilkinson-Herbots, H. M. (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(2):92–108.

Appendix

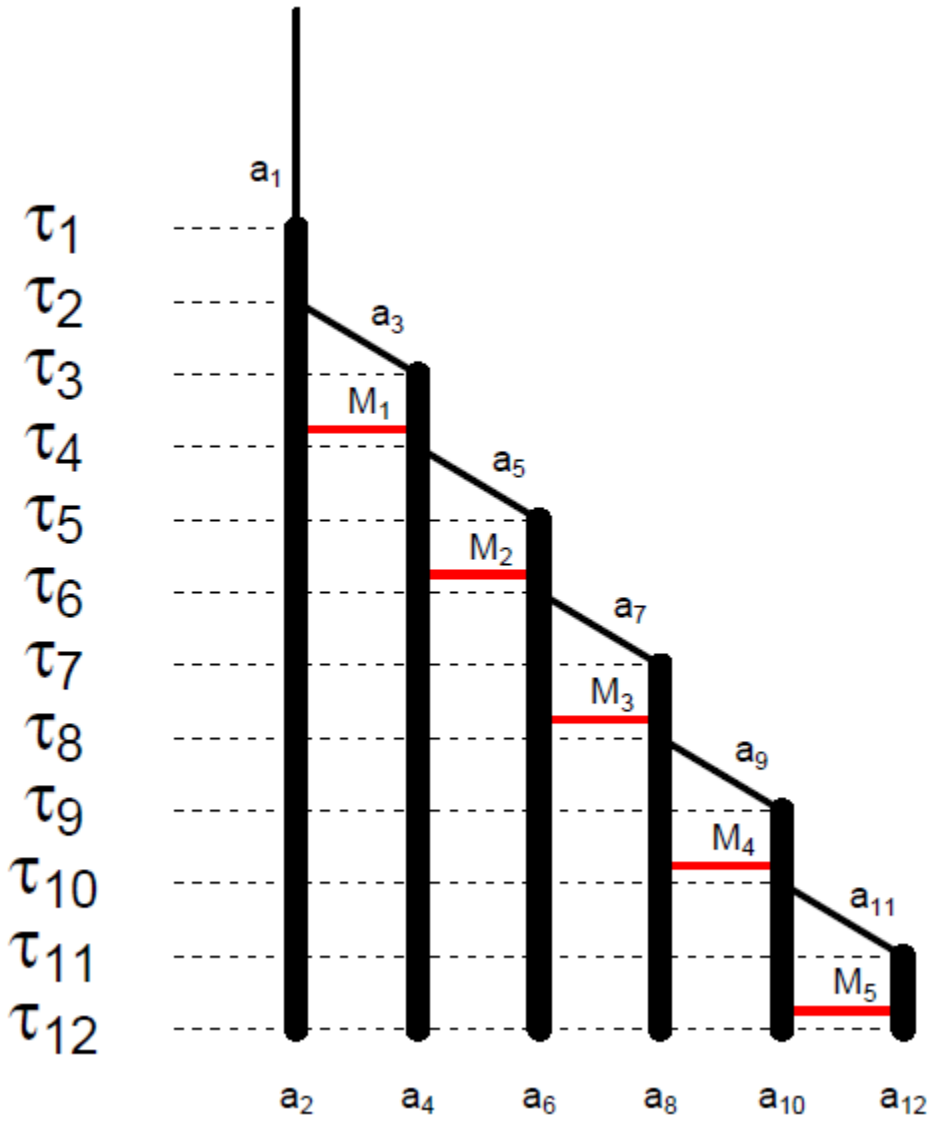

Figure 3-1 Serial founder model with migration when there are 6 extant populations See text for model description.

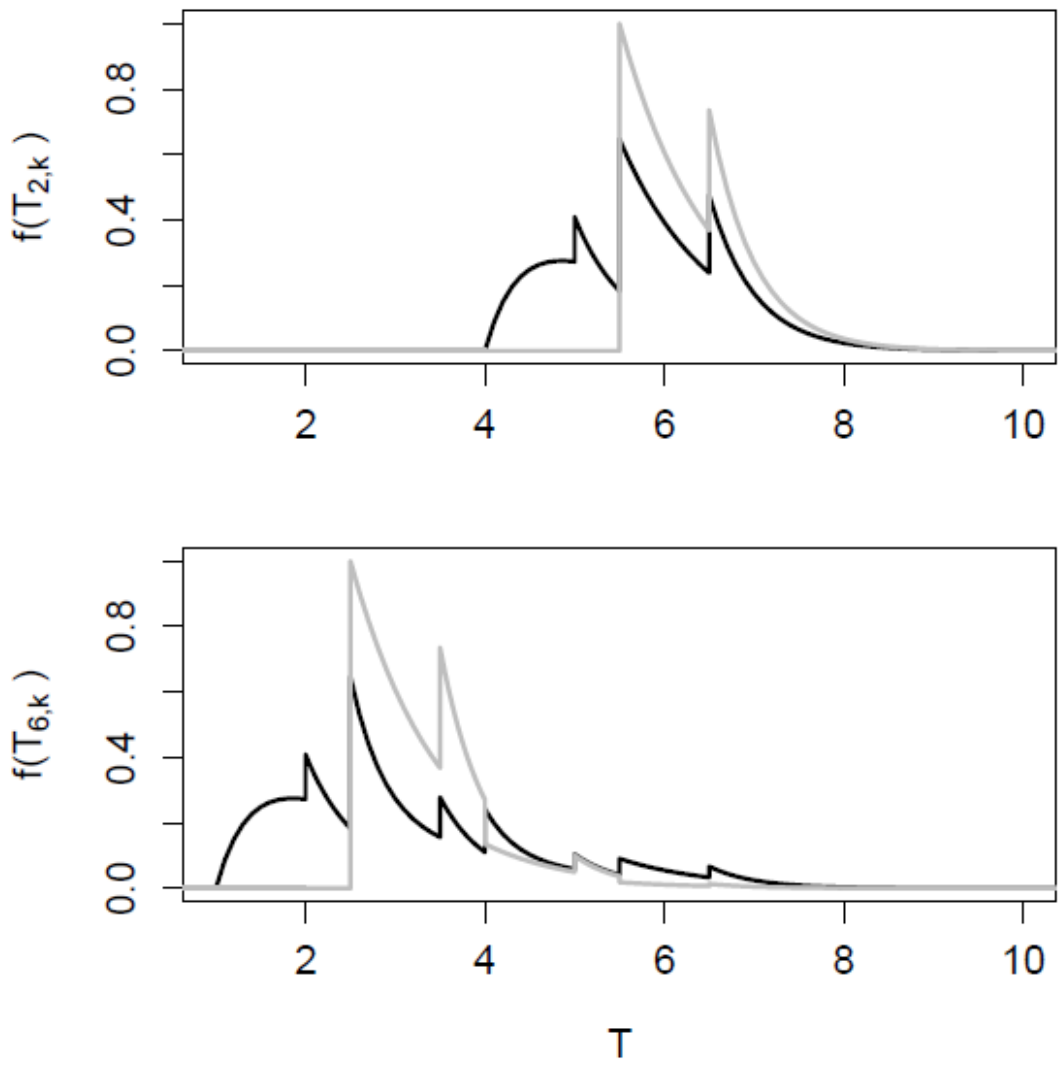

Figure 3-2 Distribution of coalescent times in a model with historical migration (black) is different compared to the model with no migration (grey). X axis: scaled time. Top: one gene is sampled from population 2 and the other from population $k, k = 4, 6, 8$. *Bottom: one gene sampled from population 6 and the other from population* 8. *Model parameters:* $\tau_b = 0.5$, $\tau_M = 1$, $b = 0.5$, $M = 1$, *8 populations.*

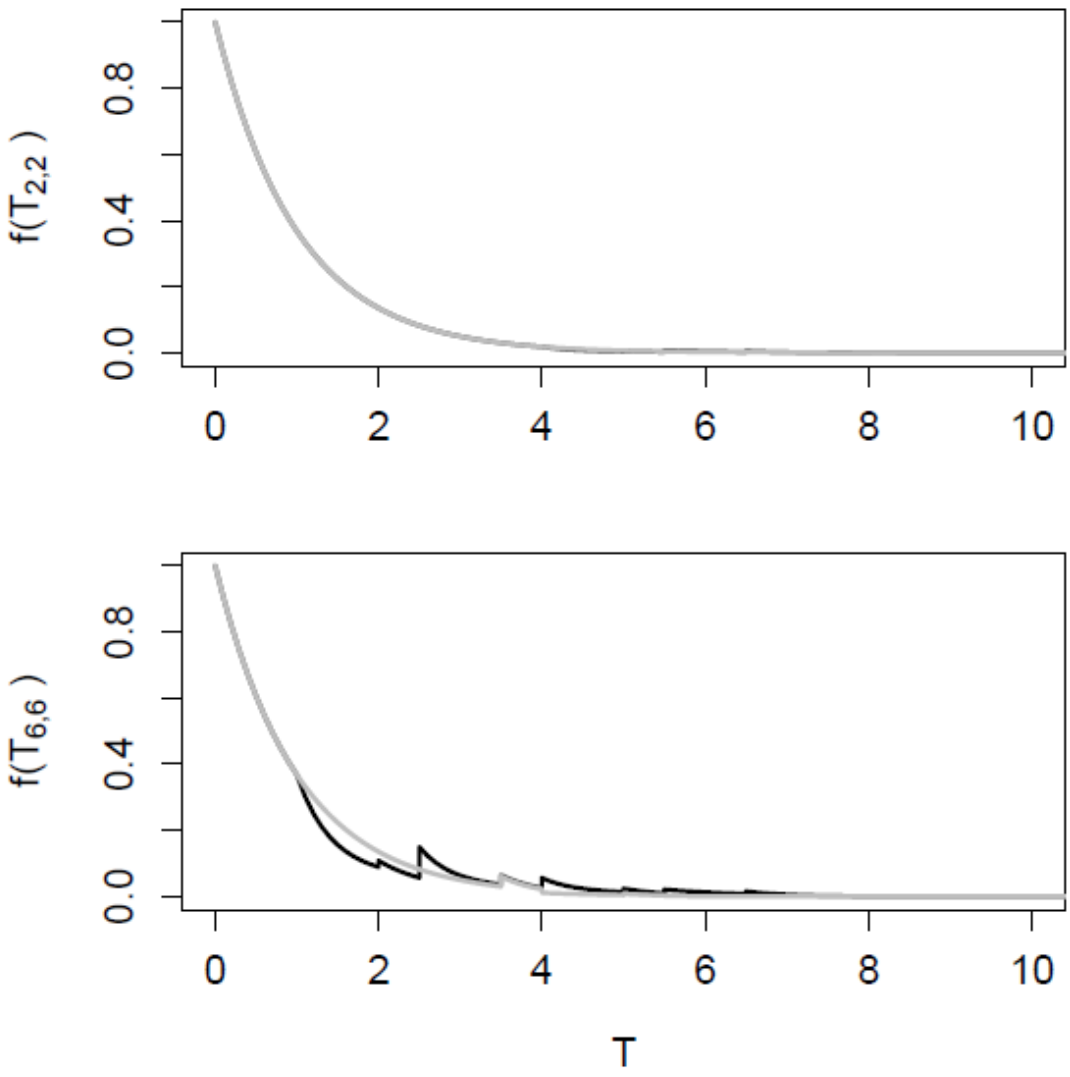

Figure 3-3 Distribution of coalescent times in a model with historical migration (black) and a model with no migration (grey). X axis: scaled time. Genes sampled from population 2 (top) and 6 (bottom).Model parameters: $\tau_b = 0.5$, $\tau_M = 1$, $b = 0.5$, $M = 1$, *8 populations.*

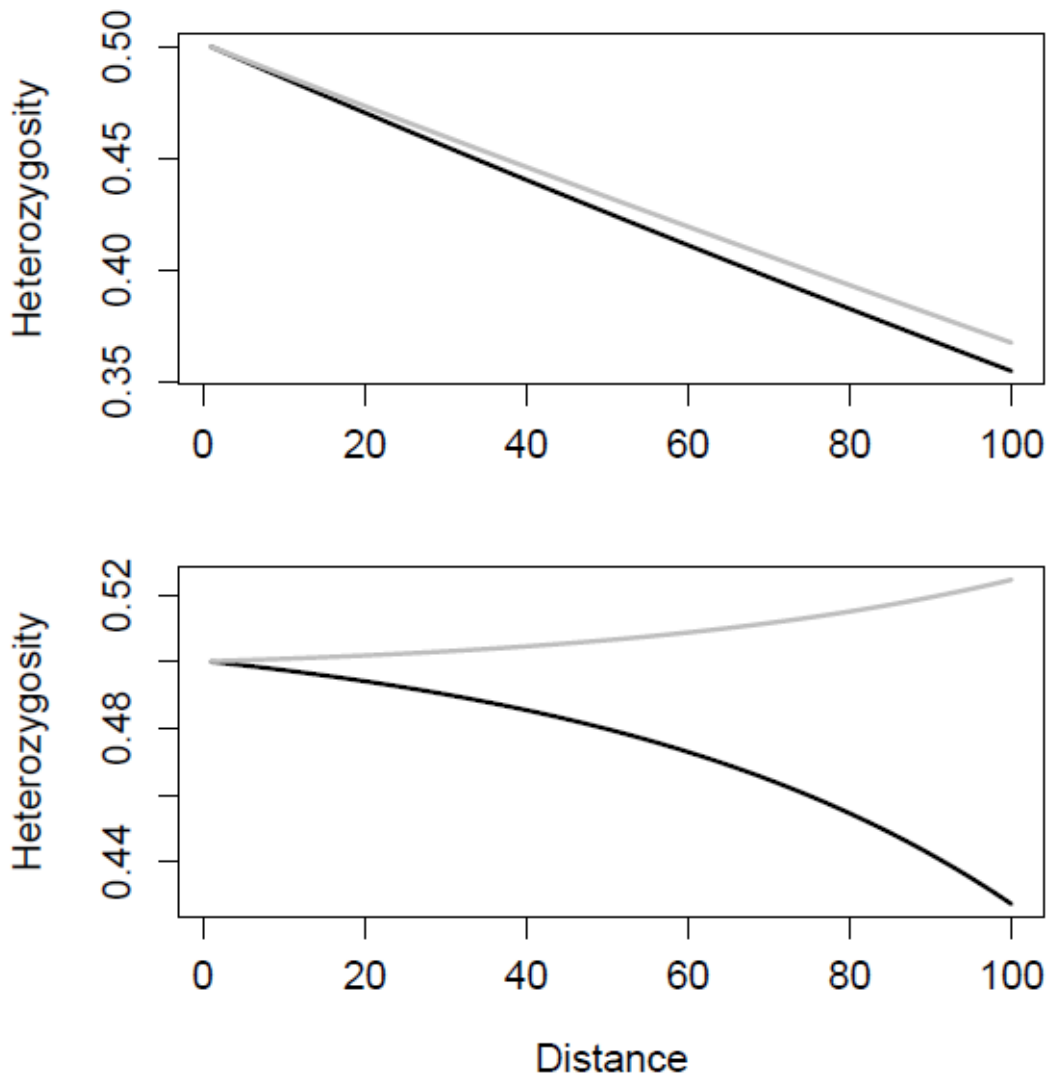

Figure 3-4 In a migration model, heterozgosity can decrease or increase in distant populations depending on parameters. X axis: distance from the first observable population, corresponds to population number in (DeGiorgio et al 2011). Model parameters $\tau_b = 0.025$, $M = 100$ *(grey lines),* $M = 0$ *(black lines),* $t_b = 0.0001$, $b = 0.025$, *(top)* $\tau_M = 0.00095$, *(bottom)* $\tau_M = 0.01$.

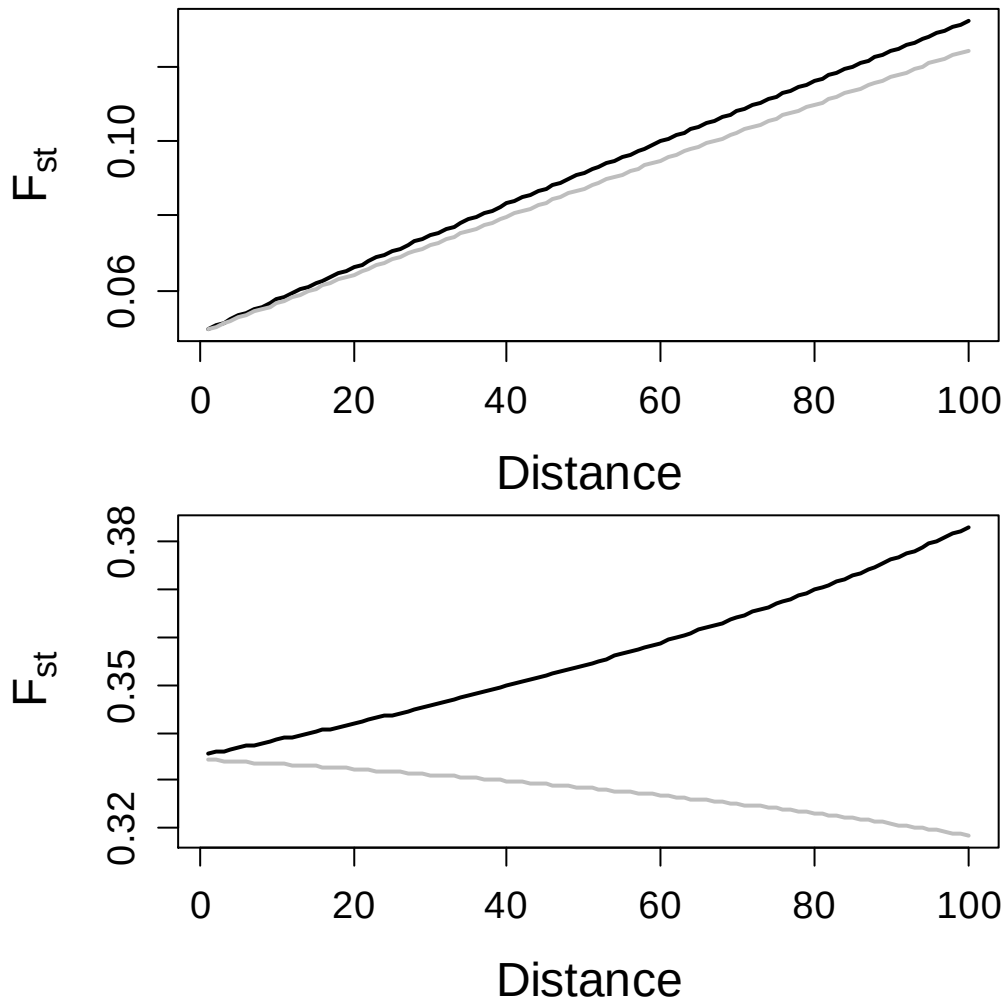

Figure 3-5 Pairwise F_{st} *in models with (grey) and without (black) migration when j = 2.* F_{st} *is a function of expected coalescent times, therefore it can decrease in distant populations in the model with migration. X axis: distance from the first observable population, corresponds to population number in (DeGiorgio et al 2011). Parameters* $\tau_b = 0.025$, $M = 100$ *(grey lines),* $M = 0$ *(black lines),* $t_b = 0.0001$, $b = 0.025$, *(top)* $\tau_M = 0.00095$, *(bottom)* $\tau_M = 0.01$.

Terms on the right hand side of equations 3.24 and 3.25

To obtain the expected value, terms on the right side are:

$$
\begin{aligned}
A_E &= 1 - e^{-\tau_{j+2}}(1+\tau_{j+1}) \\
B_{E1}(t_1,t_2) &= \sum_{r=1}^{2}(A_{0r}(1/\lambda_r + t_1 - e^{-\lambda_r \tau_b}(1/\lambda_r + t_2))) \\
B_{E2}(t_1,t_2) &= \sum_{r=1}^{2}(A_{1r}(1/\lambda_r + t_1 - e^{-\lambda_r \tau_b}(1/\lambda_r + t_2))) \\
C_{E1}(t_1,t_2) &= e_{1,1}^{Q\tau_M}(1+t_1 - e^{\tau_M}(t_2+1)) + e_{1,3}^{Q\tau_M}(b+t_1 - e^{\tau_M/b}(t_2+b)) \\
C_{E2}(t_1,t_2) &= e_{2,1}^{Q\tau_M}(1+t_1 - e^{\tau_M}(t_2+1)) + e_{2,3}^{Q\tau_M}(b+t_1 - e^{\tau_M/b}(t_2+b)) \\
D_E(t_1,t_2) &= e_{1,3}^{Q\tau_M}(1+t_1 - e^{\tau_M}(t_2+1)) + e_{1,1}^{Q\tau_M}(b+t_1 - e^{\tau_M/b}(t_2+b)) \\
E_E(t_1,t_2) &= 1 + t_1 - e^{-(t_2-t_1)}(1+t_2) \\
F_E(t_1) &= b + t_1
\end{aligned}
$$

To obtain the expression for the probability of observing l pairwise differences, terms on the right side are:

$$A_E = B(1,l,\theta,0) - e^{-\tau_{j+2}} B(1,l,\theta,\tau_{j+1})$$

$$B_{E1}(t_1,t_2) = \sum_{r=1}^{2} (A_{0r}(B(l/\lambda_r,l,\theta,t_1) - e^{-\lambda_r \tau_b} B(l/\lambda_r,l,\theta,t_2)))$$

$$B_{E2}(t_1,t_2) = \sum_{r=1}^{2} (A_{1r}(B(l/\lambda_r,l,\theta,t_1) - e^{-\lambda_r \tau_b} B(l/\lambda_r,l,\theta,t_2)))$$

$$C_{E1}(t_1,t_2) = e_{1,1}^{Q\tau_M}(B(1,l,\theta,t_1) - e^{\tau_M} B(1,l,\theta,t_2)) + e_{1,3}^{Q\tau_M}(B(b,l,\theta,t_1) - e^{\tau_M/b} B(b,l,\theta,t_2))$$

$$C_{E2}(t_1,t_2) = e_{2,1}^{Q\tau_M}(B(1,l,\theta,t_1) - e^{\tau_M} B(1,l,\theta,t_2)) + e_{2,3}^{Q\tau_M}(B(b,l,\theta,t_1) - e^{\tau_M/b} B(b,l,\theta,t_2))$$

$$D_E(t_1,t_2) = e_{1,3}^{Q\tau_M}(B(1,l,\theta,t_1) - e^{\tau_M} B(1,l,\theta,t_2)) + e_{1,1}^{Q\tau_M}(B(b,l,\theta,t_1) - e^{\tau_M/b} B(b,l,\theta,t_2))$$

$$E_E(t_1,t_2) = B(1,l,\theta,t_1) - e^{-(t_2-t_1)} B(1,l,\theta,t_2)$$

$$F_E(t_1) = B(b,l,\theta,t_1)$$

where function B is given by equation (3.8).

Conclusion

In my dissertation I considered how different evolutionary processes and population histories can produce various DNA polymorphism patterns, and how we can use these patterns to learn about populations' histories.

In the first chapter I clarify numerous recent claims about the evolution of "Genomic Islands of Divergence". I show that the main features of GIDs, such as its shape, are approximated well by analytical results found in the literature dealing with barriers to gene flow. I also show that different "hitchhiking" mechanisms are not needed to describe how GIDs appear and are maintained over time. I dispute claims about the transience of GIDs by showing that GIDs themselves do not change over time, and I clarify the effects of population size, migration, recombination, the strength of selection and initial conditions on GID size. Lastly, I show that weakly selected alleles can rapidly diverge if they are within a GID (close enough to strongly selected gene). Overall, this chapter is an important contribution to the study of speciation since the GID metaphor is widely used in the speciation literature and there is substantial confusion regarding the vocabulary accompanying it. Relating GIDs to standard and well-established vocabulary will facilitate future communication between biologists.

The main result of the second chapter is the derivation of the expression for the distribution of pairwise differences in a hybridization model with migration. I describe how the distribution of pairwise differences depends on model parameters and show that

it can be, in part, described using already known results of the recently studied "Isolation with Initial Period of Migration" model. The most important contribution of this chapter is the ability to use this result to construct the likelihood function which can be used to infer model parameters from whole genome sequences. Inferring parameters using analytical equations rather than extensive simulations is faster and provides "exact" results. However, this approach for parameter estimation is limited to the comparison of two sequences. We also find situations in which a model with no migration, but in which populations change sizes over time cannot be distinguished from a model with migration. This is an unsettling result and future work is needed to fully understand under which parameter combinations two models result in the same distributions of pairwise differences. One possible solution to this problem might be to present all different equally likely models and let the researcher decide which one is more plausible based on other evidence (archeological findings, or historical records for example).

In the third chapter I found that historical migration during human colonization of the world does not qualitatively affect the patterns of pairwise F_{st} and heterozygosity decay. This is because the effects of bottlenecks are stronger than the effects of migration. This result is in agreement with previous research, but was based on exact analytical results rather than simulations. This allows for quick detection of parameter combinations under which the historical migration model produces qualitatively different results from a model without migration. This is an interesting result since it shows that historical migration can cause seemingly counterintuitive results such as a decrease of pairwise F_{st} with distance. One concern with the historical migration model is that it is too simple and

unrealistic to describe migration during human colonization of the world. Different migration schemes can be considered using the same approach I used to study historical migration.

Vita

Ivan Juric was born in Croatia. He graduated with a B.A. in biology from the University of Zagreb in 2007. In the fall of 2007, Ivan began work toward a Ph.D. in the Department of Ecology and Evolutionary Biology at the University of Tennessee.