Charles University in Prague

Faculty of Science

Study programme: Bioinformatics

Branch of study: Bioinformatics

Zuzana Vozárová

Genomic Approaches for Studying Speciation

Genomické přístupy ve studiu speciace

Bachelor's thesis

Supervisor: RNDr. Radka Reifová, Ph.D.

Prague 2018

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze dne 6. 5. 2018 podpis autora

I dedicate this thesis to my supervisor RNDr. Radka Reifová, Ph.D. for patience and guidence, to Dávid Jónás, Ph.D., and Hana Pařízková for consultation of informatical chapters and to my parents and Bc. David Kubeša for continues support in my studies.

Abstract

Technological advances in DNA sequencing along with the emergence of new informatics approaches have created new possibilities in many biological fields. In this bachelor thesis, I will focus on the informatics approaches used in speciation genomics, that is research field focused on the problematics of the origin of new species. I will introduce some statistical methods used by these approaches for parameter estimation. The four particular methods I will write about are Maximum likelihood estimation, Bayesian model, Markov chain Monte Carlo and Iterative approach. I will describe several methods used for the detection of interspecific hybrids and recent as well as historical interspecific gene flow. These methods include NewHybrids, the hybrid index, genomic and spatial clines and coalescent-based methods. The thesis demonstrates the usefulness of the connection of applied mathematics and genomics for addressing general biological issues, and speciation particularly.

Keywords: speciation, hybrid zones, gene flow, probabilistic algorithms, bioinformatics

Abstrakt

Pokrok v oblasti technologií sekvenace DNA spolu se vznikem nových informatických přístupů, vedly ke vzniku nových možností na poli biologie. V této bakalářské práci se zaměřuji na informatické přístupy používané v speciační genomice,což je vědecký obor, který se soustředí na problematiku vzniku nových druhů. Představím statistické metody, které tyto přístupy využívají. Čtyři konkrétní metody, o kterých píšu, jsou Odhad maximální věrohodnosti, Bayesovské modely, Markovovské řetězce Monte Carlo a Iterativní přístup. Přiblížím několik metod používaných v detekci mezidruhových hybridů a určovaní jak nedávného tak historického mezidruhového genového toku. Tyto metody zahrnují program NewHybrids, hybridní index, genomické a prostorové klíny a metody založené na koalescenčních modelech. Tato práce vyzdvihuje prospěšnost propojení aplikované matematiky a genomiky při řešení obecných biologických problémů a speciace konkrétně.

Klíčová slova: speciace, hybridní zóny, genový tok, pravděpodobnostní algoritmy, bioinformatika

Contents

1. Introduction

The development of the next-generation sequencing technologies at the beginning of the century offered a possibility to obtain a huge amount of sequence data for both model and non-model species. The simultaneous rise in computational power and various adaptations of algorithms offer us the opportunity to process the massive amount of data. This led to the revolution in many fields of biology, including population and speciation genomics.

The goal of this thesis is to merge biological and informatics view in the context of speciation. In the next chapter, I will briefly introduce the topic of speciation and highlight some basic terms. The phenomenon of hybrid zones is described as the perfect natural laboratory for studying the evolution of new species.

The following chapter briefly introduces four methods widely used in solving biological problems. The primary goal of this chapter is not the deep understanding of how actual algorithms works, but rather a brief overview that also helps biologists to understand the limitations of various approaches using these methods.

The fourth chapter will focus on several approaches used in speciation genomics. I will talk about three issues widely studied in speciation genomics and some particular algorithms used to solve them. I have chosen four algorithms frequently used in the study of hybrid zones, and I will explain how they work from the view of informatics.

2. Speciation

Speciation is the evolutionary process that leads to the formation of new species, typically understood as the origin of reproductive barriers between populations (Coyne and Orr [2004], Seehausen et al. [2014]). In this context, we define species as the groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups, according to the biological species concept, initially introduced by Ernst Mayr (1942).

We distinguish several types of reproductive isolation mechanisms:

- Premating isolation barriers, which occur before mating and include for example behavioral, ecological, mechanical and mating system isolation.
- Postmating, prezygotic isolation barriers, which occur after mating but before the formation of zygote, and include various forms of gametic isolation.
- Postzygotic isolation barriers, which occur after the formation of hybrid zygote. We distinguish extrinsic (environment-dependent) or intrinsic (environment-independent) postzygotic isolation barriers. The first includes worse adaptation to environment, while the later one includes hybrid sterility and inviability.

When reproductive isolation arises in sympatry, i.e. within the same geographical area, we call it sympatric speciation. When reproductive isolation arises in geographic isolation, it is called allopatric speciation. When two distinct or currently diverging species show incomplete reproductive barrier, they exchange genes to some degree and we talk about interspecific gene flow. Such gene flow occurs between species during the sympatric speciation, but can also occur after secondary contact of previously allopatric species in the secondary hybrid zones.

2.1 Speciation drivers

Divergent selection

We distinguish two primary drivers of the arising isolation (Seehausen et al. [2014]). The divergent (or disruptive) selection comes from a combination of ecological and sexual selection. This kind of selection prefers individuals with extreme phenotypes and penalize individuals of an intermediate phenotype. We call this genotype-environment interaction.

This type of selection leads to extrinsic postzygotic and prezygotic isolation barriers and is essential during sympatric speciation as well as for divergence of species after secondary contact.

Intrinsic postzygotic barriers

Intrinsic postzygotic barriers originate as a by-product of allopatric divergence and can be driven by genetic drift or selection. Origin of these barriers is described by Dobzhansky-Muller model (Coyne and Orr [2004] p. 269–272).

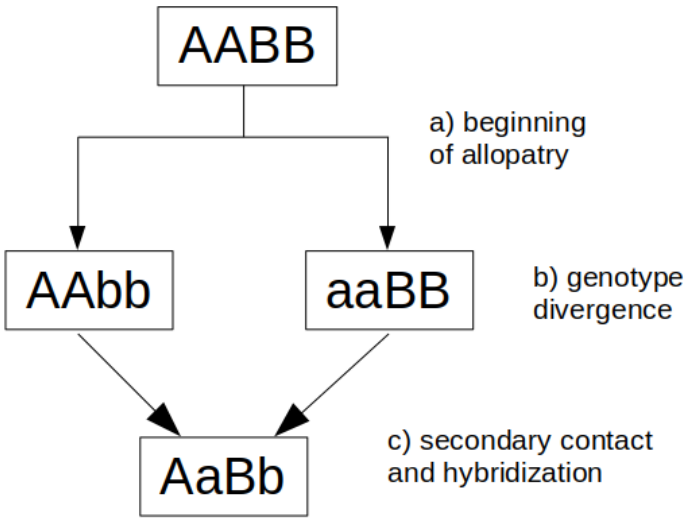

Figure 1: In Dobzhansky-Muller model, we begin with one population. After geographical isolation of two subpopulations (a), selection or genetic drift leads to genetic differentiation of these subpopulations (b). If these populations meet and hybridize in secondary contact (c), the new previously untested combinations of genes may lead to reduced viability or even sterility of hybrids.

This model is based on the premise, that in evolution, every fixed allele of each gene is tested against other genes of the species and the selection favors those, that lead to the best viability in context of the whole genome. However, when an individual with a new previously untested combination of alleles occurs, the result is uncertain (see figure 1). Incompatibilities between mutations that arose in the two geographically isolated populations lead to sterility or inviability of interspecific hybrids. They are mostly caused

by the genomic conflict regarding the genetic or epistatic incompatibilities and lead to the intrinsic reproductive isolation.

This kind of isolation is independent on the environment. The intrinsic postzygotic barriers typically drive the allopatric speciation.

2.2 Hybrid zones and interspecific gene flow

Allopatric speciation in the absence of gene flow was the dominant view of speciation in second half of the 20^{th} century (Sara [2001]). Therefore many scientists concluded, that it was the most typical way how new species were formed (e. g. Coyne and Orr [1998], Turelli et al. [2001], Futuyma and Mayer [1980], and according to Sousa and Hey [2013] also Mayr [1942], Mayr [1963] and Dobzhansky and Dobzhansky [1937]).

However, nowadays it is acknowledged that in many cases allopatric species that come into secondary contact are able to hybridize and produce partially viable and fertile offspring. It is estimated that interspecific hybridization occurs in about 10–30 % of multicellular species (Abbot et al. [2013]). An area, where the hybridization takes place, is called the hybrid zone.

It is typical for populations which are not entirely separated, for example by distance, that gene flow between them is occurring at some stage of divergence (Abbot et al. [2013]). In hybrid zones, two distinct species and an intermediate population of hybrids are always present. In the context of hybrid population, we distinguish two situations, where the F1 (first generation) hybrids do not produce the offspring (due to sterility or other barriers) or when F1 mate with purebred individuals or other hybrids.

It is important to underline, that in this second scenario hybrids are not only one genotype, but represent a wide range of recombinants. Hybrids often show novel phenotypes, that would otherwise not appear (Barton and Hewitt [1985]). In this second scenario, interspecific hybridization can lead to interspecific gene flow between the species when some genes from one species can invade the gene pool of the other species. Interspecific gene flow usually occurs in some loci, but not in other. We call this the concept of semi-permeable barrier that allows some genes or only particular alleles to move freely thought, some only harder and some not at all. This barrier can also be asymmetrical and let the given allele to flow just in one direction.

As the mutations are very rare, from 10^{-8} to 10^{-9} per generation per base pair (Abbot et al. [2013]), so-called 'adaptive introgression' (i.e. when introgression become important source of novel genetic variation maintained by natural selection) could be an essential

source of novel adaptations (Arnold and Kunte [2017]).

Besides that, interspecific gene flow can promote species adaptive radiations as has been shown for example in cichlid fishes (Meier et al. [2015], Darwin's finches (according to Grant and Grant [1994] and Lamichhaney et al. [2015]) and interspecific introgression of genes conditioning Muller mimicry in *Heliconius* butterflies (Dasmahapatra et al. [2012]).). On the other hand, interspecific gene flow between species can lead to break down of reproductive barriers and the reduction or the loss of differentiation (Taylor et al. [2006]).

Speciation can be completed after secondary contact in the presence of gene flow by the reinforcement process when selection against hybridization leads to the evolution of premating barriers. Similarly, competition between species in the secondary contact zone can lead to differentiation of ecological niches and habitat shifts in the two species, which can also strengthen the reproductive isolation between the species (Wu [2001], Via [2009], Sottas et al. [2018]). Because similar processes take place during speciation after secondary contact and sympatric speciation, some researchers prefer to distinguish speciation with gene flow (i.e. sympatric speciation and allopatric speciation when reproductive isolation after secondary contact is not complete) and speciation without gene flow (i.e. allopatric speciation without any gene flow after secondary contact) (Smadja and Butlin [2011]).

As a consequence, we can see species in two different ways. On the one side, species can be understood as populations separated by genetic barriers to the gene exchange, or on the other side as a set of populations maintained in particular stable equilibrium by selection.

3. Mathematical methods for estimating parameters of models

In biological research, we usually begin with the collection of observations of the population, about which we wish to make inferences (Quinn and Keough [2002] p. 14–29). This collection is called a sample, and the number of observation is called sample size. The sample is characterized by measured characteristics, in this context rather called statistics.

When drawing conclusions from this samples, in order to generalize the conclusion to the whole population, we are looking for the explanation of the observed pattern by a model or theory (Ford [2000] according to Quinn and Keough [2002]). The model is represented by a series of statements (or formulae) that explains the observations.

We distinguish two basic types of models (James and McCulloch [1985]) - verbal explanations and mathematical models. Further, we identify empirical and theoretical mathematical models. Empiric models describe relationships resulting from processes we study, on the other hand, theoretical models are used to study the process itself.

The mathematical model in general consists of variables, constants, and equations describing the relations between them. The variables that we can vary as needed to match the observation with the model are called parameters. As the complexity of the model rises, the estimation of precise values of these parameters becomes more difficult.

The methods, I write about in this chapter, offer us the possibility, how to time efficiently get the nearly optimal values of parameters. In other words, we do not necessarily get the right ones, where the model precisely matches the observation, but good enough to explain the system and to study the effects of its different components.

3.1 Maximum likelihood estimation

The maximum likelihood estimation is the general method for calculating statistics that estimates specific parameters of the given model according to given sample of observations to maximize the likelihood of observing those data (Quinn and Keough [2002] p. 23–25).

The method takes into account the probability of data against the model, typically read as probability of data given model, $P_M(Data \mid \Theta)$ (also adopted as the likelihood of given model against the data, $L_M(\Theta \mid Data)$, where Θ represents the set of parameters of the model) and tries to find those values of parameters that maximize the likelihood of

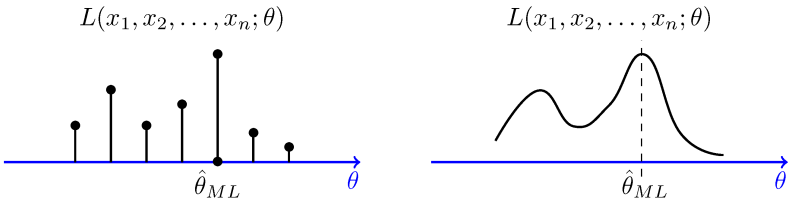

Figure 2: Maximum likelihood estimation. On the left side, the graph shows the sample of calculated values from the real function on the right side. Source: Pishro-Nik [2014]

this observation.

The general formula for likelihood function is:

$$L_M(Data \mid \Theta) = \prod_{i=1}^{n} f(Data_i \mid \Theta) \tag{3.1}$$

where n is the number of observations, $Data_i$ represents the data of observation with index i and $f(Data_i \mid \Theta)$ is the probability distribution of $Data$ for possible values of Θ.

It is more common to use the log-likelihood function for simplification of the computation:

$$L_M(Data \mid \Theta) = \ln \prod_{i=1}^{n} f(Data_i \mid \Theta) = \sum_{i=1}^{n} \ln f(Data_i \mid \Theta) \tag{3.2}$$

In some cases, the estimators of parameters have exact arithmetical solutions (e.g. when estimating means or parameters for linear models). In other cases, when the distribution of observations is non-normal, calculation of estimators use complex iterative algorithms. Most of the time, we do not know the maximum value of likelihood, values of parameters are non-discrete, and it is inefficient to calculate values across an entire likelihood function (see figure 2).

3.2 Bayesian model

The approaches based on frequentist statistics, including Maximum likelihood estimation, are loaded with some limitations. First, there is no way, how to easily incorporate any prior information besides the observed sample data. Additional knowledge about probabilities of values of the parameters can not be easily considered. Second, the interval estimate (an interval, which includes the right solution with a given probability) we have obtained has a frequentist interpretation and contains the fixed population parameter. Bayesian approach removes these limitations by using Bayesian statistics (Quinn and Keough [2002] p. 26–31).

In Bayesian probability consider, that the parameter takes a range of possible values, each according to different probabilities or degrees-of-belief of being true (Barnett [1999] according to Quinn and Keough [2002]).

The general formula of Bayesian statistics is:

$$P_M(\Theta \mid Data) = \frac{P_M(Data \mid \Theta)P_M(\Theta)}{P_M(Data)} \tag{3.3}$$

where $P(\Theta)$ is so-called unconditional prior probability summarizing our prior knowledge about the Θ distribution, $P(Data|\Theta)$ is the likelihood $Data$ given values of parameters Θ and $P(\Theta|Data)$ is the posterior probability of θ given observed $Data$. $P(Data)$ is equal to the expected value (mean) and standardizes the likelihood, so that the area under the posterior probability distribution (simply understand as the sum of all probabilities) equals one.

The prior probability distributions measure, how strong is the belief in possible values of the parameter to be the right one (Dennis [1996]). It can be in the form of prior ignorance or substantial prior knowledge two forms (Barnett [1999] according to Quinn and Keough [2002]). The first one has the form of non-informative prior distribution, also called diffuse, because a wide range of values is possible.

3.3 Markov chain Monte Carlo

Probabilistic algorithms, also called randomized, are based on the premise, that the result and/or the way the result is obtained depend on chance (Törn [2001], Motwani and Raghavan [1995]). In some problems, where the result is stochastic, the use of these algorithms is natural (e.g. simulating the behavior of some existing or planned system over time). In other cases, the deterministic problem is transformed into a stochastic one

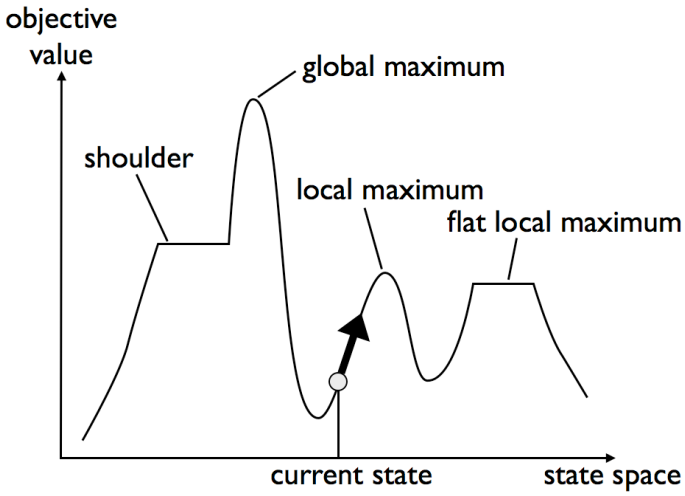

Figure 3: Markov chain Monte Carlo. One dimensional state space is shown for easier understanding. The graph also shows important types of maximum. Source: Rößler [2013]

(e.g. numerical integration, optimization). For these applications, the result is always approximate, and its expected precision is proportional to the time available to use. The techniques that apply probabilistic algorithms to numerical problems are called Monte Carlo methods.

In biological problems, we use Markov Chain Monte Carlo method (see figure 3) to estimate values of parameters of the model, to get the best (among available) objective value.

The model is always represented by an objective function and a state space. The state space consists of all combinations of all values of every explanatory variable of the model, and the objective function is defined in each point in this space. The function represents the measure of how accurate is the representation of observed data by the given combination of parameters. We call this number the objective value.

In real implementations, these algorithms do not search the whole state space; it would be very computationally and time-consuming. Instead, they use various heuristics and approximations of the problem to choose areas in state space, where the probability of finding the optimal objective value is high enough. This assumption leads to finding the best or at least good enough solution in short time, usually given by the fixed number of steps.

As in Maximum likelihood estimation (described above), typically we do not know the optimal solution for the objective value, and the number of points in the state space is

potentially infinite.

3.4 Iterative approach

The iterative approach is based on the premise that for each suboptimal representation of the parameters we can estimate new representation, that better suits the observation. In this context, we express each parameter with an equation, which allows us to calculate its value according to the values of other parameters. Whole section is base on Kelley [1995].

The algorithm starts with some random values of parameters. We run multiple sessions, typically the given number of times or until the convergence condition is met. In each session, we iterative recalculate the values of all parameters in the model according to given equations.

The most common way, how the convergence is given, is by the so-called equilibrium achievement. This concept is based on the experience, that after each session the values of parameters more or less differ from their previous values. When after the session the values in parameters does not significantly differ, the equilibrium is met. In other words, we find the values, which satisfie the equation.

Another convergence method is given by an objective function that quantifies the match between the model and the observation.

4. Genomic approaches

The development of the next-generation sequencing methods has created new possibilities for studying the genetic basis of species formation and quantifying levels of gene flow between these species (Sousa and Hey [2013]). This chapter describes several approaches used in the speciation genetic research.

The first section writes about the classification of distinct individuals of a population in hybrid zones according to Mendelian genetic markers. I focus on two different types of classification, in the form of gene frequency classes and genotype frequency classes. Both approaches are represented each by one particular program used for calculation. I underline the main differences between these two approaches.

In the second section, the theory of clinal models is described. This approach studies the distribution of different allelic variations along a gradient in hybrid zones and provides us the information about the quantity of gene flow in distinct regions of a genome. This gradient can be taken as geographical distribution of different allelic forms of particular genes or as the distribution of allelic forms across a spectrum of admixture (e.g. hybrid index).

In the third section, approaches for quantifying historical gene flow are briefly introduced.

4.1 Identification of interspecific hybrids

Mendelian genetic markers give us valuable tools for studying species hybridization by the characterization of individuals as purebred individuals or hybrids (Avise [1994] according to Anderson and Thompson [2002]). In the last decades, various approaches to distinguish between pure parental species and interspecific hybrids have been developed (e. g. Anderson and Thompson [2002], Szymura and Barton [1991], Rieseberg et al. [1998], Pritchard et al. [2000], Buerkle [2005], Rannala and Mountain [1997]). This section describes two widely used approaches that classify the purebred and the hybrid individuals in two slightly different ways.

The first method by is based on the calculation of the hybrid index, the parameter defined as the proportion of alleles inherited from one of two parental species. It can be used on one individual at a time, but a priori needs frequencies of parental alleles. The hybrid index method has been used in several studies of hybrid zones (e. g. Rieseberg

et al. [1998], Rieseberg et al. [1999], Hardig et al. [2000], Rogers et al. [2001] and Watano et al. [2004]). I describe this method as it was implemented by Buerkle [2005].

Anderson and Thompson [2002] introduced the second method called NewHybrids. This method must be used on the whole population at one time to classify all individuals in various classes of hybrids (F1's, F2's, and several backcrosses). This type of classification is vital for documenting gene exchange and introgression between species.

4.1.1 Hybrid index

The hybrid index represents the proportion of alleles of an individual, that were inherited from one of the two parental species. In this context, we call one of the parental species *reference*, its hybrid index equals to 1, and the second species is called *alternative*. The value of hybrid index ranges from 0 to 1 as the ratio of alleles inherited from reference species to all alleles of interest. Using this number, individuals of the hybrid population can be categorized in so-called gene frequency classes, where all individuals with the same value of hybrid index belong to one class (according to Anderson and Thompson [2002]).

When all markers used in the analysis are codominant and fixed between the parental species, the hybrid index can be calculated directly from the genomic information of an individual (i.e. we simply count the number of alleles inherited from the reference species; Barton and Gale [1993] according to Buerkle [2005]). However, not all markers are codominant (e.g. random amplified polymorphic DNA and amplified fragment length polymorphism), and some markers do not exhibit fixed differences between taxa. Therefore more complex approaches that require the aid of software are needed.

The algorithm introduced by Buerkle [2005] is based on Maximum likelihood estimation approach. The likelihood function is given by:

$$MLE(h|g) = \sum_{i=1}^{n} \begin{cases} ln(p_{i,j}^2 + 2p_{i,j}p_{i,k}) & \text{if the locus is positive in a dominant marker,} \\ ln(2p_{i,j}p_{i,k}) & \text{if the locus is heterozygous in a codominant marker, and} \\ ln(p_{i,j}^2) & \text{if the locus is negative in a dominant or homozygous in a codominant marker} \end{cases} \tag{4.1}$$

where h is the value of the hybrid index of a given individual, g is the genotype of the given individual, n is the number of loci and j and k are the alleles of the individual in loci

at position i. The dominant markers were assumed to be for example random amplified polymorphic DNA or amplified fragment length polymorphism. The codominant markers are typically gene alleles.

Variable $p_{i,j}$ is introduced as the probability of a given allele j at locus i. It is calculated by formula:

$$p_{i,j} = hr_{i,j} + (1-h)s_{i,j} \tag{4.2}$$

where $r_{i,j}$ is the frequency of allele j at locus i in the reference species, $s_{i,j}$ is the frequency of allele j at locus i in the alternative species and h is the hybrid index. This means, that $p_{i,j}$ is the probability that this individual with the hybrid index h have allele j at locus i.

As the formula suggests, frequencies of parental alleles given by $r_{i,j}$ and $s_{i,j}$ must be known for each locus and each allele. The algorithm is looking for the best model, represented by hybrid index h, according to a given genotype of an individual.

4.1.2 Genotype frequency classes

The program NewHybrids is intended to classify the sympatric population, which consists of two distinct species and the hybrids risen from the mating between them. The approach assumes result genotype by using the data on multiple unlinked markers. Classification is presented in the form of posterior probability computed by Markov chain Monte Carlo and reflect the level of certainty for each individual belonging to various hybrid classes (purebred individuals, F1's, F2's, and multiple backcrosses).

The enumeration of expected genotype frequency classes and computation of the expected proportions of the genotypes follows Mendel's laws. For the classification, we typically use markers that are fixed for one allele in each of parental species, but different allele in each. For each individual, markers are homozygote for one or other parental species or are heterozygote. For example, F1 (hybrid of the first generation) hybrids are expected to be heterozygous for all markers; BC1 (backcross of F1 and purebred individual) hybrids are supposed to be heterozygous in a 50 % of markers and homozygous (allele from parental species to which is backcrossed) in 50 % of markers.

The number of generations of inbreeding is known or estimated from observations and should be rather small. Individuals arising from many generations of backcrossing are difficult to distinguish from pure individuals even with many diagnostic markers (Boecklen and Howard [1997]).

In contrast with the hybrid index reviewed in the previous section, Newhybrids does not require any prior information about a model. The program itself can estimate, which

markers are fixed in which parental population. If we know, that some sampled individuals come from parental classes, the information can be added for improvement of the result.

The method used in the NewHybrids program is called Gibbs sampling and is based on the Hastings [1970] algorithm, which combines Markov chain Monte Carlo and Iterative approach (see the previous chapter). The iteration is driven by given number of sessions.

4.2 Quantifying levels of interspecific gene flow

The distribution of different allelic variations in hybrid zones can be influenced by selection against introgressed alleles, and this information can be used to identify genomic regions with limited gene flow (Rieseberg et al. [1999]).

The study of distinct alleles in the context of gene flow is usually done by spatial and genetic clines. The cline theory (Barton and Hewitt [1985]) provides the new level of understanding the dynamics of reproductive barriers and increases our knowledge about the causes of the observed patterns of gene frequencies. This approach uses natural admixture between divergent lineages (i.e. hybridization) to investigate the genetic architecture of reproductive isolation and adaptive introgression (e. g. Gompert and Buerkle [2009]).

4.2.1 Spatial clines

The spatial cline represents the distribution of a given allele among the geographic area of the hybrid zone. We can distinguish various types of spatial clines according to the value of the characteristic scale of selection l, that is given by $l = \frac{\sigma}{\sqrt{s}}$ (Barton and Hewitt [1985]), where σ is the dispersal rate or the standard deviance of the distance between parent and offspring (see figure 4) and s is the parameter of selection or the inverse of the time since contact for neutral clines.

Another parameter is the width, w, defined as the inverse of the maximum gradient (see figure 4). This parameter is estimated for each distinct locus and represents the measure of gene flow. Loci with a low level of gene flow are mostly those that are sources of gene incompatibility.

In dispersal-dependent cline, w is mostly of the same order as l (Bazykin [1969]). In dispersal-independent cline, this parameter must be much greater than l.

The following section introduces the method for estimating spatial clines (Slatkin [1973]). The author made several assumptions in order to simplify the problem, most importantly that each distinct generations consist of three separate phases: a) mating and

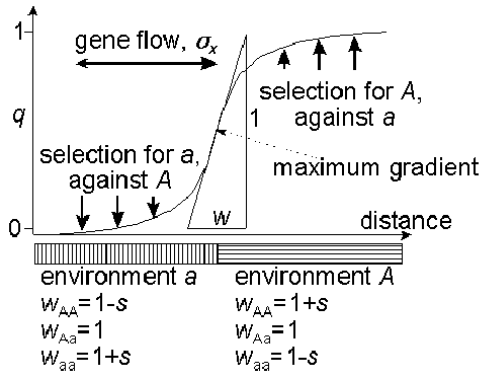

Figure 4: Example of a spatial cline as the representation of the frequency of allele A according to the geographic gradient. The value of parameter w is represented by the triangle at the highest derivation of the function at its adjacent side. Variables w_{AA}, w_{Aa} and w_{aa} represents effect of selection in two separate locations along the gradient. σ_x shows the dispersal rate. Source: Mallet [2006]

the production of offspring, b) natural selection, and c) the movement of individuals to other locations.

The final model is based on the formula of population genetics by Crow and Kimura [1970]:

$$p'(x,t) = \frac{p(x,t) + sy(x)(p(x,t))^2}{1 + sy(x)(p(x,t) - q(x,t))} \tag{4.3}$$

where $p(x,t)$ is the frequency of a given allele at the location x at given time t, $q(x,t)$ is the mortality index, $p'(x,t)$ is frequency of given allele after mating but before dispersal, s is the selection parameter and *y(x)* is the effect of selection at a given location x.

Another assumptions by Slatkin [1973] lead to neglecting the time t as parameter (i. e. p(x,t) $\rightarrow$p(x) and p'(x,t) $\rightarrow$p'(x)).

To this thought we add the dispersal phase as the migration from all different locations summed up:

$$p(x') = \sum_x M(x,x')p'(x) \tag{4.4}$$

where $p'(x')$ is obtained from equation 4.3. The new variable $M(x,x')$ is the probability of an individual to migrate from the location x to the location x'. The matrix M is a representation of the "migration matrix" introduced by Kimura and Weiss [1944].

The very algorithm starts with (randomly) chosen initial variables for $p(x)$ at each

location x. The final value is calculated by iterating the formula 4.4 until the equilibrium is approached (see Iterative approach). The spatial cline for the given locus is then calculated from the final values of $p(x)$ for every location x.

4.2.2 Genomic clines

Similarly, as the spatial clines study gradient of distinct alleles according to the geographical position of individuals, the genomic clines compare the quantity of introgression of distinct alleles from one parental species to another. Therefore genomic clines are more appropriate to use in mosaic hybrid zones (where two parental species are not simply meeting somewhere in the middle, but are both distributed among some area) and in other cases, where a simple spatial pattern is not available (e. g. Szymura and Barton [1986], Rieseberg et al. [1999], Tang et al. [2007] and Macholán et al. [2011]).

Models of genomic clines are intended to generate a null-distribution for patterns of neutral introgression across genome-wide admixture using various methods. Widely used are the simple parametric model, the permutation method, and the hierarchical Bayesian framework (e. g. Gompert and Buerkle [2009], Gompert and Buerkle [2011]).

The null model is used to identify outlier (here non-neutral) loci, which are linked to candidate genes of reproductive isolation or hybrid vigor. Typically we talk about the markers with reduced introgression (a deficit of heterozygotes) or markers with increased introgression (an excess of heterozygotes).

Several empirical studies of hybrid zones used these methods to repeatedly provide evidence for variable introgression among genomes (e. g. Macholán et al. [2007], Carling and Brumfield [2009], Rieseberg et al. [1999], Nolte et al. [2009], Gompert et al. [2010] and Teeter et al. [2010])

Genomic clines can be defined as mathematical functions describing the probability of locus-specific ancestry along a gradient represented by genome-wide admixture or by hybrid index (Gompert and T. L. Parchman [2012]). According to this approach, Gompert and Buerkle [2011] introduced the Bayesian model for estimating genomic clines, which describes this function by two critical genomic cline parameters, α and β. These variables quantify various patterns of introgression and form the primary model for detecting outlier loci (figure 5).

In the context of these parameters, we set one of the parental populations as the population 1. Parameter α represents an increase ($a > 0$) or decrease ($a < 0$) of the probability, that the locus originate from parental population 1. This parameter also

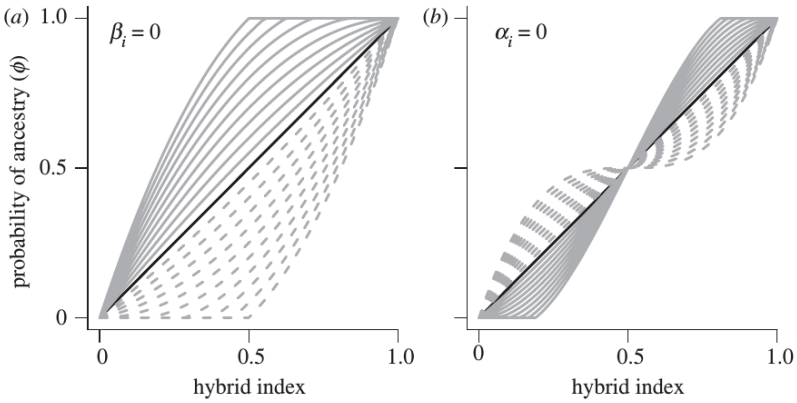

Figure 5: Two graphs show hypothetical genomic clines. Axis y represents the probability of ancestry of given allele from parental population 1, axis x represents the gradient along hybrid index. In (a) parameter β_i is fixed to 0 and parameter α_i varies from -1 to 1. Negative values are symbolized by dashed grey lines, positive values by solid grey lines and the solid black line represent the value of 0. Likewise in (b) parameter α_i is fixed and parameter β_i varies from -1 to 1. Source: Gompert and Buerkle [2011]

defines the center of the cline. β quantifies the rate of change in the probability function along the gradient. Positive values of β specify a steeper cline, negative values of β specify a wider (i.e. less steep) cline.

4.3 Historical gene flow

While the approaches mentioned in the previous sections are intended to measure recent interspecific gene flow, this section briefly introduces one of the approaches used to quantify historical gene flow between species.

4.3.1 Coalescent models

The coalescent theory describes the distribution of gene trees under a given demographic model. This model can be used to compute the probability of a given gene tree (Sousa and Hey [2013]). This approach is based on modeling population divergence according to different models.

Some basic models are (also see figure 6):

a) isolation without migration (also called allopatric divergence scenario),

b) isolation with migration,

c) isolation after migration and

d) secondary contact.

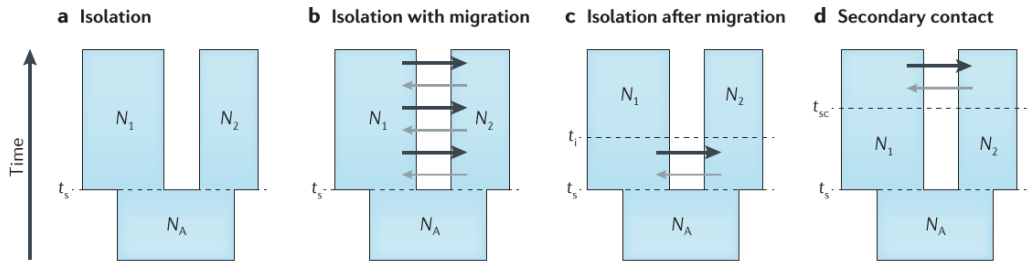

Figure 6: The basic population divergence models: a) isolation without migration, b) isolation with migration, c) isolation after migration and d) secondary contact. The different shades of grey represent the potentially different migration rates. Source: Sousa and Hey [2013]

The allopatric divergence scenario can be distinguished from the models with migration by patterns of genetic variation in samples from two closely related populations or species (Nielsen and Wakeley [2001] and Hey and Nielsen [2004]).

In this approach, methods are not focusing on the best gene tree; the highest likelihood is obtained by integrating over all possible genealogies (Felsenstein [1988]). This way we can estimate for example effective population sizes, migration rates, admixture contributions and time of species divergence (Nielsen and Beaumont [2009], Marjoram and Tavaré [2006]).

Multiple programs use coalescent model approach. A few widely used methods are briefly mentioned in the following text.

DIYABC (for 'do it yourself', Cornuet et al. [2014]) is based on approximate Bayesian computation (ABC) used for the analysis of single nucleotide polymorphism data at a large number of loci by Bayesian model choice.

IMa (Hey and Nielsen [2007]) implements the Isolation with Migration model. Markov chain Monte Carlo simulations are used for the likelihood-based analyses.

ms (Hudson [2002]) is the program for generating samples under neutral models, based on a Monte Carlo simulation and according to a Wright-Fisher neutral model. Its extension **msHot** (Hellenthal and Stephens [2007]) allows both crossover and gene conversion hotspots for simulating genetic variation data for a sample of chromosomes from a population. Another extension **msms** (Ewing and Hermisson [2010]) includes selective sweeps.

5. Conclusion

The aim of this thesis was to summarize genomic approaches used in speciation research and introduce mathematical tools used by these methods. Besides the general principles of these approaches, I also mentioned concrete implementations for deeper illustration.

I wrote about two different algorithms for hybrid classification. In the context of the hybrid index, we can talk about gene frequency classes, the groups of individuals with the same or similar enough value of the hybrid index. This classification is rarely used, because the potential number of classes is infinite, according to the precision of calculation and number of markers. Despite this fact, with at least three generations of inheritance, there is always a higher number of genotype frequency classes than gene frequency classes. For example, F2 and F3 are two distinct genotype classes, but both have the same hybrid index of 0.5.

This is because the hybrid index takes the alleles only by their frequency, but omits the information whether the locus is homo- or heterozygote. According to this view, we can tell, that the NewHybrids program gives us more detailed information. On the other hand, the NewHybrid program can have some problems, when working with markers from sexual chromosomes. It can be expected, that about half of the population, depending on the sample, will be haploid in all these markers. Therefore some manipulations with data must be done in order to get relevant results.

As I worked with this program, I used two basic changes of a dataset. The first change was to omit the individuals with haploid markers (females in Abraxas system, males in Drosophila system). The second change is to omit the markers on haplotype chromosome. Both of these manipulations lead to loss of information. Therefore the further work with this open source program is suggested, to incorporate understanding of haplotype markers by NewHybrids.

I introduced two clinal approaches. When we use the spatial cline, we suppose, that along one spatial dimension the frequency of one allele (in biallelic approximation) arise at the expense of the other in the same marker. In the middle of the spatial gradient, the frequency of both is about $\frac{1}{2}$. According to Hardy–Weinberg equilibrium, the more balanced frequencies of alleles are in the location along the spatial gradient, the higher is the probability of an individual to be heterozygote in this marker.

We assume, that the clines of distinct neutral alleles typically correlate to each other, which concludes that in the view of neutral alleles, the locations with a high frequency

of heterozygotes for distinct alleles are near each other. Therefore the probability of an individual to be heterozygote in multiple alleles rises in the middle of the spatial cline.

The individuals in the edges of the spatial cline are purebred individuals, because the frequency of some allele is 1. The probability of heterozygote alleles rise along the first half of the gradient and then decline in the second half. Therefore the spatial gradient may correlate with the occurrence of individuals from various hybrid classes (i. e. with hybrid index). When we assume, that the edges of the cline are continuing source of purebred individuals from parental populations, the spatial gradient would correlate with the genomic gradient.

The outcome of this hypothesis leads to cross usage of genomic and spatial principles. For example, we can sequence individuals across some area, build genomic cline along the hybrid index. Then we map back the two-dimensional spatial gradient of distinct alleles according to the presence of individuals of given hybrid index. This would not show us local differences between various local preferences of other minor alleles, but this approach can offer us the insight, where to look for them. Then we can make deeper analysis along these new small spatial gradients.

All the tools I wrote about in this thesis have had a significant impact in the field of the speciation genomic. Despite this considerable improvement, there are still some areas where the limitations of current implementations meet the cases not included in initial assumptions. The deeper understanding and broader dialogue between biology, applied mathematics and programming should be the right way how to solve this issues.

Bibliography

R. Abbot, D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, et al. Hybridization and speciation. *Journal of Evolutionary Biology*, 26:229–246, 2013.

E. C. Anderson and E. A. Thompson. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics Society of America*, 160:1217–1229, 2002.

M. L. Arnold and K. Kunte. Adaptive genetic exchange: A tangled history of admixture and evolutionary innovation. *Trends in Ecology and Evolution*, 32(8):601–611, 2017.

J. C. Avise. *Molecular Markers, Natural History and Evolution*. Chapman and Hall, New York, 1994.

V. Barnett. *Comparative Statistical Inference*. Wiley, New York, 1999.

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

N.H. Barton and K. S. Gale. *Genetic analysis of hybrid zones. In: Hybrid Zones and the Evolutionary Process*. Oxford University Press, New York, 1993.

A. D. Bazykin. Hypothetical mechanism of speciation. *Evolution*, 23:685–87, 1969.

W. J. Boecklen and D. J. Howard. Genetic analysis of hybrid zones: numbers of markers and power of resolution. *Ecology*, 78:2611–261, 1997.

C. A. Buerkle. Maximum-likelihood estimation of a hybrid index based on molecular markers. *Molecular Ecology Notes*, 5:684–687, 2005.

M. D. Carling and R. T. Brumfield. Speciation in passerina buntings: introgression patterns of sex-linked loci identify a candidate gene region for reproductive isolation. *Molecular Ecology*, 18:834–847, 2009.

J. M. Cornuet, P. Pudlo, J. Veyssier, A. Dehne-Garcia, M. Gautier, R. Leblois, J. M. Marin, and A. Estoup. Diyabc v2.0: a software to make approximate bayesian computation inferences about population history using single nucleotide polymorphism, dna sequence and microsatellite data. *Bioinformatics*, 30(8):1187–1189, 2014.

J. Coyne and H. Orr. *Speciation*. Sinauer Associates, 2004. ISBN 0-87893-089-2.

J. A. Coyne and H. A. Orr. The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society*, 353:287, 1998.

J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory.* Harper and Row, New York, 1970.

K. K. Dasmahapatra, J. R. Walters, A. D. Briscoe, J. W. Davey, A. Whibley, et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 7405(487):94–98, 2012.

B. Dennis. Discussion: should ecologists become bayesians? *Ecological Applications*, 6: 1095–1103, 1996.

T. G. Dobzhansky and T. Dobzhansky. *Genetics and the Origin of Species.* Columbia University Press, 1937.

G. Ewing and J. Hermisson. Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16): 2064–2065, 2010.

J Felsenstein. Phylogenies from molecular sequences: inference and reliability. *Annual Reviews Genetics*, 22:521–565, 1988.

E.D. Ford. *Scientific Method for Ecological Research.* Cambridge University Press, Cambridge, 2000.

D. J. Futuyma and G. C. Mayer. Non-allopatric speciation in animals. *Systematic Biology*, 29:254–271, 1980.

Z. Gompert and C. A. Buerkle. A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology*, 18:1207–1224, 2009.

Z. Gompert and C. A. Buerkle. Admixture in european populus hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. *Molecular Ecology*, 20:2111–2127, 2011.

Z. Gompert and C. A. Buerkle T. L. Parchman. Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society B*, 367:439–450, 2012.

Z. Gompert, L. K. Lucas, J. A. Fordyce, M. L. Forister, and C. C. Nice. Secondary contact between lycaeides idas and l. melissa in the rocky mountains: extensive introgression and a patchy hybrid zone. *Molecular Ecology*, 19:3171–3192, 2010.

P.R. Grant and B.R. Grant. Phenotypic and genetic effects of hybridization in darwin's finches. *Evolution*, 48:297–316, 1994.

T. M. Hardig, S. J. Brunsfeld, R. S. Fritz, M. Morgan, and C. M. Orians. Morphological and molecular evidence for hybridization and introgression in a willow (salix) hybrid zone. *Molecular Ecology*, 9:9–24, 2000.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

G. Hellenthal and M. Stephens. mshot: modifying hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, 23(4):520–521, 2007.

J. Hey and R. Nielsen. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of drosophila pseudoobscura and d. persimilis. *Genetics*, 167:747–760, 2004.

J. Hey and R. Nielsen. Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8):2785–2790, 2007.

R. R. Hudson. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

F.C. James and C.E. McCulloch. Data analysis and the design of experiments in ornithology. *Current Ornithology*, 2:1–63, 1985.

C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. Society for Industrial and Applied Mathematics, Philadelphia, 1995.

M. Kimura and G. H. Weiss. The stepping stone model of population structure and the decrease in genetic correlation with distance. *Genetics*, 49:561–576, 1944.

S. Lamichhaney, J. Berglund, M. S. Almén, K. Maqbool, M. Grabherr, A. Martinez-Barrio, et al. Evolution of darwin's finches and their beaks revealed by genome sequencing. *Nature*, 518(5):371–375, 2015.

M. Macholán, P. Munclinger, M. Sugerková, P. Dufková, B. Bímová, E. Bozíková, J. Zima, and J. Piálek. Genetic analysis of autosomal and x-linked markers across a mouse hybrid zone. *Evolution*, 61:746–771, 2007.

M. Macholán, S.J. Baird, P. Dufková, P. Munclinger, B. V. Bímová, and J. Piálek. Assessing multilocus introgression patterns: a case study on the mouse x chromosome in central europe. *Evolution*, 65:1428–1446, 2011.

J. Mallet. BIOL2007 - evolutionary genetics, 2006. URL `http://www.ucl.ac.uk/~ucbhdjm/courses/b242/Hz/Hz.html`.

P. Marjoram and S. Tavaré. Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, 7:759–770, 2006.

E. Mayr. *Systematics and the Origin of Species: from the Viewpoint of a Zoologist*. Harvard University Press, 1942.

E. Mayr. *Animal Species and Evolution*. Harvard University Press, 1963.

J. I. Meier, D. A. Marques, S. Mwaiko, C. E. Wagner, L. Excoffier, and O. Seehausen. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, 8(5):14363, 2015.

R. Motwani and P. Raghavan. *Randomized algorithms*. Cambridge University Press, New York, 1995. ISBN 0-521-47465-5.

R. Nielsen and M. A. Beaumont. Statistical inferences in phylogeography. *Molecular Ecology*, 18:1034–1047, 2009.

R. Nielsen and J. Wakeley. Distinguishing migration from isolation: a markov chain monte carlo approach. *Genetics*, 158:885–896, 2001.

A. W. Nolte, Z. Gompert, and C. A. Buerkle. Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. *Molecular Ecology*, 18:2615–2627, 2009.

H. Pishro-Nik. Introduction to probability, statistics and random process, Kappa Research LLC, 2014. URL `https://www.probabilitycourse.com/chapter8/8_2_3_max_likelihood_estimation.php`.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

G. Quinn and M. Keough. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, New York, 2002. ISBN 978-0-511-07812-5.

B. Rannala and J. L. Mountain. Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences USA*, 94:9197–9201, 1997.

L. H. Rieseberg, S. J. Baird, and A. M. Desrochers. Patterns of mating in wild sunflower hybrid zones. *Evolution*, 52:713–726, 1998.

L. H. Rieseberg, J. Whitton, and K. Gardner. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, 152:713–727, 1999.

S. M. Rogers, D. Campbell, S. J. Baird, R. G. Danzmann, and L. Bernatchez. Combining the analyses of introgressive hybridisation and linkage mapping to investigate the genetic architecture of population divergence in the lake whitefish (coregonus clupeaformis, mitchill). *Genetica*, 111:25–41, 2001.

J. Rößler. Genetic algorithms, 2013. URL `http://www.handmade-insights.com/blog/2013/genetic-algorithms/`.

V. Sara. Sympatric speciation in animals: the ugly duckling grows up. *TRENDS in Ecology & Evolution*, 16(7):381–390, 2001.

O. Seehausen, R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman, P. A. Hohenlohe, C. L. Peichel, G. P. Saetre, et al. Genomics and the origin of species. *Nature Reviews — Genetics*, 15:176–192, 2014.

M. Slatkin. Gene flow and selection in a cline. *Genetics*, 75:733–756, 1973.

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

C. Sottas, J. Reif, L. Kuczyński, and R. Reifová. Interspecific competition promotes habitat and morphological divergence in a secondary contact zone between two hybridizing songbirds. *Journal of Evolutionary Biology*, 2018.

V. Sousa and J. Hey. Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews — Genetics*, 14:404–414, 2013.

J. M. Szymura and N. H. Barton. Genetic analysis of a hybrid zone between the fire-bellied toads, bombina bombina and b. variegata, near cracow in southern poland. *Evolution*, 40:1141–1159, 1986.

J. M. Szymura and N. H. Barton. The genetic structure of the hybrid zone between the fire-bellied toads bombina bombina and b. variegata: comparisons between transects and between loci. *Evolution*, 45:237–261, 1991.

H. Tang, S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Cintron, E. G. Burchard, and N. J. Risch. Recent genetic selection in the ancestral admixture of puerto ricans. *American Journal of Human Genetics*, 81:626–633, 2007.

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

K. C. Teeter, L. M. Thibodeau, Z. Gompert, C. A. Buerkle, M. W. Nachman, and P. K. Tucker. The variable genomic architecture of isolation between hybridizing species of house mouse. *Evolution*, 64:472–485, 2010.

A. A. Törn. Probabilistic algorithms: Spring 2001 course, 2001. URL `http://users.abo.fi/atorn/ProbAlg/Abstract.html`.

M. Turelli, N. H. Barton, and J. A. Coyne. Theory and speciation. *Trends in Ecology and Evolution*, 16:330–343, 2001.

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

Y. Watano, A. Kanai, and N. Tani. Genetic structure of hybrid zones between pinus pumila and p. parviflora var. pentaphylla (pinaceae) revealed by molecular hybrid index analysis. *American Journal of Botany*, 91:65–72, 2004.

C.I. Wu. The genic view of the process of speciation. *Journal Evolution Biology*, 14: 851–865, 2001.