University of Glasgow | School of Computing Science

The effect of natural selection on relatedness in randomly mating population

Adam Kurkiewicz

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Level 4 Project — March 25, 2016

Abstract

A flexible population simulation framework is designed and implemented in Python. The framework allows to simulate biological populations, whose evolution is specified by the user, through implementing python functions defining desired breeding, mating, mutation, migration and inheritance schemes. The effect of running each simulation is a file containing a complete information on the entire history of the population, including genotypes of all simulated individuals, parent-child relationships, the generation number each sim belonged to when alive, and the location it occupied. Produced population file can be processed algorithmically and statistically to investigate the simulated population. An experiment is designed in order to answer the research question: Does natural selection affect relatedness in a species? While we fail to reach a definite conclusion using the methods employed, the areas for further investigation are proposed.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: _____ Signature: _____

Contents

1 Introduction **1**

1.1 Context of the work 1

1.2 Research question 1

2 Review of the field **3**

2.1 Basic inheritance 3

2.1.1 Early days 3

2.1.2 Mendelian Inheritance 3

2.1.3 Hardy-Weinberg equilibrium. 5

2.2 Work of Fisher 5

2.3 Coalescent method 6

2.4 Forward-Time Modeling 6

2.5 Modelling the recent common ancestry of all living humans 6

3 Choice of methodology **8**

3.1 Simulation technique 8

3.2 Workload management 8

4 Report of practical work **9**

4.1 Natural Selection Framework 9

4.1.1 High-Level overview 9

4.1.2 Performance and porability 10

4.1.3 Genome 10

4.1.4 Migration 10

4.1.5 Mutations and Inheritance 11
4.1.6 Natural Selection 11
4.1.7 Generations 11
4.1.8 Mating and Breeding 11
4.1.9 Code Correctness 12
4.1.10 Computing MRCA 12
4.1.11 Extensibility 12
4.1.12 Documentation 15

5 Report of research work **16**

5.1 Selective Sweep in Randomly Mating Population 16

6 Conclusion **19**

6.1 Inconclusive results 19
6.2 Topics for the future 21

Glossary **22**

Bibliography **24**

Chapter 1

Introduction

1.1 Context of the work

The publication of "On the Origin of Species" (Darwin 1859) is generally accepted to be a founding stone of still vigorously researched field of evolutionary biology. Over the years, as the understanding of the field deepened, it became clear that certain scientific questions within the field can only be answered by obtaining progress in related fields, like establishing laws governing inheritance (genetics) (Corcos and Monaghan 1993), understanding the role of organic molecules in the process of inheritance (molecular biology) (Watson and Crick 1953), understanding population dynamics of predator-prey system using both continuous and discrete methods (mathematical biology), understanding the character of population variability at molecular level (neutral theory of molecular evolution, Markov processes) (Kimura 1968; Dayhoff 1973), studying historical populations (archaeology, anthropology, linguistics) (Theunissen and Perlin-West 1989), proposing and verifying hypotheses about populations based on quantitative evidence (statistics) (Allentoft et al. 2015), modeling and simulating biological systems *in silico* (systems biology), and many others.

While the research presented in this work falls within the realm of computational biology, when referring to other, connected fields we might occasionally need specialized jargon. A glossary of technical terms in use is included at the end of this work. Throughout the text words and phrases in bold (e.g. **autosomal**) are explained in the glossary.

1.2 Research question

The scientific question that we will try to answer in this piece of work is directly connected with evolution by natural selection and perhaps might have even occurred to Darwin himself:

Research Question. *Does natural selection affect relatedness in a species? Are individuals in a species under selective pressure more likely to be closely related?*

We will seek to answer these questions using a computer simulation. First, we have to decide on how to measure relatedness. We will follow Rohde, Olson, and Chang (2004) and focus on computing time (in generations) that is taken for a group of simulated individuals to reach a Most Recent Common Ancestor (MRCA). More precisely, we will speak of a **population graph**, which we will define to be a Directed Acyclic Graph (DAG), with vertices representing all individuals ever alive and directed edges representing the parent-child relationship between individuals. Additionally, each vertex will be labeled with complete information on individual's traits.

For any point in time all individuals at that point will be referred to as "generation at time t", or simply "generation", if the time can be inferred from the context. An MRCA of a generation at time t will be a common ancestor which is the most recent, i.e. precedes the population by the least number of generations. The time (in generations) that is taken to reach such individual will be denoted by T_{MRCA}.

There are other, auxiliary questions we will have to answer on our way. Most crucially we will have to decide on such important details of our simulation as the mating scheme, migration, competition for resource, survivability, number of offspring, etc. Often we will be able to use the choices of parameters already made by others, e.g. breeding scheme (Peng and Amos 2010). We prepare the ground in Chapter 2. Appropriate methodology is decided upon in Chapter 3. We discuss the architecture of the computer program which constitutes the solution, and is on its own a valuable contribution to the field, in Chapter 4. The experiment, which constitutes a novel contribution to the field is discussed in Chapter 5. The conclusions will be drawn in Chapter 6. The source code of all developed computer programs can be accessed publicly on the internet using the github platform (Kurkiewicz 2016).

Chapter 2

Review of the field

2.1 Basic inheritance

Presented work is by no means the first attempt at quantitative study of population genetics. What follows is a review of the most relevant achievements in the field.

2.1.1 Early days

Ever since three European scientist simultaneously and independently rediscovered & experimentally confirmed Gregor Mendel's theory of inheritance (DeVries 1900; Tschermak-Seysenegg 1900; Correns 1900), the interplay between Mendelian Inheritance and Darwinian Evolution has been hotly debated. One claim popular at the time has fallaciously asserted that Mendelian Genetics cannot be a correct model of trait inheritance (Yule 1902). The reasoning at the core of the claim was that **dominant** alleles over time sweep the entire population, thus eliminating variation within the population, rendering evolution by natural selection impossible. This misconception was debunked in a very short letter to the editor of "Science" by a famous English number theorist (Hardy 1908). The letter is a truly master example of a presentation of non-trivial mathematical thought in a manner both rigorous and understandable, even to non-experts. Let us recall Mendel's laws of inheritance before we present Hardy's argument.

2.1.2 Mendelian Inheritance

Let us start by briefly sketching a relevant subset of the laws of Mendelian Inheritance. A more comprehensive reference can be found in an annotated English translation of the original work of Mendel by Corcos and Monaghan 1993.

Firstly, we will be acting under the assumption that each animal characteristic (trait) is either present, or not present in a given organism because of:

Assumption 1. *possession by this organism of one, two or none copies of a gene encoding this characteristic.*

Assumption 2. *the mode of expression of the characteristic, which can be either* ***dominant*** *or* ***recessive****.*

We have to remark that these assumptions may not be immediately true for certain seemingly continuous, partially environment-dependent characteristics such as height, weight or individual's position in the autism

spectrum. However experiments using various methods have consistently asserted that 80% of human height is inheritable (Silventoinen et al. 2003; Visscher et al. 2006). Although the challenge of identifying all genes affecting height is still open, recent progress using a cross-discipline technique called Genome Wide Association Study (GWAS) identified over 40 genes, which contribute 5 percent points towards total height heritability. To play devil's advocate further, it has been recently established that important traits can be acquired and inherited by non-genetic means in both plants (Ong-Abdullah et al. 2015), non-human animals (Dolinoy, Huang, and Jirtle 2007) and humans (Yehuda et al. 2015), giving rise to the field of Epigenetics. We shall consider both environmental and epigenetic factors to be outside of the scope of this research. A curious reader is referred to an accessible, popular science discussion of the topic of GWAS, height inheritance and epigenetics by Maher (2008). A concerned reader should find assurance in that a great many traits do obey Mendelian's Inheritance straightforwardly. A comprehensive database for human traits, maintained by John Hopkins University at omim.org (OMIM 2016), finds 24617 such genes using the match-all wildcard $*$, including 22044 **autosomal** genes (as of March 13, 2016).

Before we proceed let us recall the following, standard notation: if an **allele** encoding a trait is determined to be **dominant** we will represent it with capital letter, e.g. A. In such case a lack of trait will be considered **recessive** and be denoted as a. Under the assumption 1, we will consider all three possible genotypes: (A,A) – homozygous dominant, (A,a) – heterozygous and (a,a) – homozygous recessive. Of these the first two the genotypes (A,A) and (A,a) will result in an individual possessing the characteristic. An individual with the last genotype (a,a) will not exhibit the trait. Analogously, one can consider a trait inherited recessively, and in such case the above description is true when "trait" is mutually exchanged with "lack of trait", "not exhibit the trait" with "exhibit the trait", etc.

Further, we will assume that

Assumption 3. *in offspring each of the two alleles present comes from the parents, precisely one from the mother and one from the father. Each of the parent's two alleles has a 50% chance of being inherited.*

We can use our assumptions 1, 2, 3 to work out that in a specific case of breeding two heterozygous individuals, an (A,a) mother and an (A,a) father, each individual offspring will have either of 3 possible genotypes with probabilities $\frac{1}{4}$ for (A,A), $\frac{1}{2}$ for (A,a) and $\frac{1}{4}$ for (a,a). Table 2.1 gives all possible outcomes of breeding all possible genotypes involving one characteristic along with corresponding probabilities. If the desired mode of inheritance is dominant, the reader can interpret (A,a) and (A,A) as trait-positive and (a,a) as trait-negative. For the recessive case, (a,a) is trait-positive and both (A,A) and (A,a) are trait-negative.

		mother					
		(a,a)		(A,a)		(A,A)	
father	(a,a)	(a,a)	1	(a,a) (A,a)	$\frac{1}{2}$ $\frac{1}{2}$	(A,a)	1
	(A,a)	(a,a) (A,a)	$\frac{1}{2}$ $\frac{1}{2}$	(a,a) (A,a) (A,A)	$\frac{1}{4}$ $\frac{1}{2}$ $\frac{1}{4}$	(A,a) (A,A)	$\frac{1}{2}$ $\frac{1}{2}$
	(A,A)	(A,a)	1	(A,a) (A,A)	$\frac{1}{2}$ $\frac{1}{2}$	(A,A)	1

Table 2.1: Possible offspring genotypes with corresponding probabilities

Finally, we have to remark that the assumption 3 may not be correct in rare cases of **sex-linked** characteristics

		mother			probability		
		(a, a)	p	(A, a)	2q	(A, A)	r
probability	(a, a)	(a, a)	pr	(a, a)	qr	(A, A)	r^2
	r			(A, a)	qr		
	(A, a)	(a, a)	pq	(a, a)	q^2	(A, a)	qr
father	2q	(A, a)	pq	(A, a)	$2q^2$	(A, A)	qr
				(A, A)	q^2		
	(A, A)	(A, a)	p^2	(A, a)	qp	(A, A)	pr
	p			(A, A)	qp		

Table 2.2: Possible offspring genotypes with corresponding probabilities expressed in terms of probabilities of parental phenotypes.

(e.g. colour-blindess). This research will focus exclusively on **autosomal** genes.

2.1.3 Hardy-Weinberg equilibrium.

Let us desribe a distribution of genotypes in the population as the ratio of dominant homozygotes to heterozygotes to recessive homozygotes, symbolically

$$p : 2q : r$$

$2q$ appears in the ratio to ease the computation.

The ingenious insight of Hardy 1908 is in observing that for certain values of p, q and r, the ratio will stay unchanged from generation to generation; ratios with this property he calls "stable". Hardy gives conditions for stability, and shows that ratios that aren't stable converge to "stable" ratios as the population develops. Hardy's argument is based on assuming an arbitrary ratio in the parent generation and computing the ratio of the children's generation. He accomplishes that by considering "A little mathematics of the multiplication-table type" in the spirit of Table 2.2, and then solving a system of three degree-2 polynomials in three indeterminates: p, q and r. The result, $q^2 = pr$, is a necessary and sufficient condition for the ratio to be stable. We have to remark a few assumptions made by Hardy for his proof of the stability condition. Namely, the population is assumed to be isolated, randomly mating, it is assumed to be large, and the number of men and women is assumed to be equal. Hardy's argument does not take into account natural selection.

2.2 Work of Fisher

The work of Ronald Fisher had drastic influence on the landcape of both statistics and population genetics. Apart from introducing influential statistical terminology (such as the concept of variance, defined as standard deviation squared, Fisher 1918), Fisher gave modern shape to the notion of null hypothesis testing. He also invented a range of statistical methods, such as one-way analysis of variance (ANOVA). On the biological side, Fisher was behind the realisation that sexual selection is comparatively stronger, and more consistent than natural selection. He formulated the concept of Fisherian runaway and proposed the sexy sons hypothesis. He also proposed the mechanism for inheritance of complex traits (Fisher 1918). His work on coalescent processes gave rise to the coalescent method, which is of interest for the project, as it could be potentially used to answer the research question.

2.3 Coalescent method

The coalescent method, in its most basic form, is used to predict the time (in generations) one has to go back in the population graph for any genetic trait to reach the individual it originated from ("coalesce"). The method is usually employed for a group of individuals with unknown predecessor structure, which makes it suitable for use in situations when data regarding only extant individuals is available. Given an extant population, the coalescent method works backwards in time, with generations in hypothetical ancestry assigned probabilities of containing the coalescent. Although in this basic application the coalescent method is relatively straightforward to apply, once natural selection is included the method gets subtle and very complicated, with most research still open-ended (Wakeley 2016). For a comprehensive review, a reader is referred to a work by Wakeley 2016.

2.4 Forward-Time Modeling

Finally, let us move on to the most recent population modeling method, so called forward-time modeling. While the idea of running a forward-time population model is relatively straightforward, it has long been assumed that its value is only educational (Peng and Kimmel 2005). However, with the advent of vast computational power, and a desire to model complex populations, the forward-time modeling has steadily gained popularity. Genetic Simulation Resources lists 54 publications (GSR 2016), which have utilised just a single forward-time simulation framework simuPOP (Peng and Kimmel 2005).

Forward-Time Modeling is based on the idea of assuming an initial population (generation 0) and rules governing that population's development in time. Each subsequent generation is obtained by applying these rules to the previous generation. The rules are usually probabilistic, and randomness is obtained from pseudo-random number generators, like Mersenne Twister (Matsumoto and Nishimura 1998). This work is extensively based on two particular models, and can be perhaps viewed as an attempt at merging the two approaches. The first model due to Rohde, Olson, and Chang 2004 looks at the relatedness of the simulated population through computing how far back in time one has to go to find the **MRCA** of the final generation in the simulation. Some detail on their approach is given in the next section. The second model due to Peng and Amos 2010 looks at modeling a population undergoing a **selective sweep** of various strength for the purpose of simulating GWAS data. This model in turn is not concerned with computing MRCA. A model very similar to this second model is described in detail in chapter 5.

2.5 Modelling the recent common ancestry of all living humans

A research of Rohde, Olson, and Chang 2004 is a very successful blend of statistical theory and forward-time modeling applied in the context of human population from 20000 BC until the modern day. The purpose of the simulation is to compute the most recent common ancestor of present-day human population. The authors use historical and anthropological data, and an occasional educated guess, to model the spread of humans out of Africa, using a town-country-continent model, with random mating at town level, migration between towns, between countries and between continents. The authors keep the migration rate "conservatively low", as they are trying to avoid underestimating T_{MRCA}, and in accordance with their theoretical investigation the migration rate and T_{MRCA} are inversely proportional.

The authors decide to use a continuous-generation model, with individuals born throughout the lives of their parents (as opposed to at the end of their life, as with discrete generation model), co-inhabiting and perhaps even mating across generations. The authors use their model to predict that T_{MRCA} of present humans is approximately 100 generations, which is not very far from a theoretical number of about 30 under an unrealistic

assumption of random mating.

The model proposed by the authors does not take into account natural selection.

Chapter 3

Choice of methodology

3.1 Simulation technique

After a consideration given to both coalescent and forward-time approach, the forward time approach is chosen for the following reasons:

- As argued by Peng and Amos 2010, forward-time modeling is particularly well-suited for modeling populations under selective pressure.
- Writing computer code is at heart of forward-time modeling, as the technique is fundamentally computational. There are challenges involving code correctness, performance and choice of algorithms. Handling these appropriately can be considered a good challenge for a computer programmer.
- Coalescent method, with its statistical difficulties could be cognitively prohibitive, especially taken into account the amount of time that could be spent on the project and the skill-set of the investigator.

3.2 Workload management

The approach was split into two units of work. The first was the design and implementation of a forward-time modeling framework, suitable for our purpose. The framework is implemented as a python module, and was accomplished to strict software engineering standards, is implemented in a portable way, thoroughly tested, reasonably documented, and easily extensible. The second unit of work was the design and execution of the experiment itself.

We report on both in chapters 4 and 5.

Chapter 4

Report of practical work

4.1 Natural Selection Framework

An important contribution of this research work is an implementation of a forward-time population modeling framework. The framework is then used to investigate the research question that we're interested in by focusing on particular population models of our interest. However, the framework does not force its users into any particular choice of the model. A wide range of real-life populations can be modeled by implementing appropriate mutation, breeding, mating, natural selection, migration and inheritance schemes. The schemes are implemented as python functions, and can be glued together using the framework's API into a full population model. This approach is not dissimilar from simuPOP (Peng and Kimmel 2005), a particularly successful forward-time simulation framework, which has been used extensively to model populations under selective pressure, for example to model population divergence due to local adaptation (Duforet-Frebourg et al. 2016). Let us discuss the design of the framework focusing on similarities and differences with simuPOP.

4.1.1 High-Level overview

The central unit of any simulation in the framework is a "sim". Sims are computational abstractions of biological individuals. Sims are assigned partners in accordance with a mating scheme, can sexually reproduce with other sims of opposite sex, and produce offspring in accordance with a breeding scheme. Both mating and breeding schemes can be chosen from any provided by the framework or be user-defined. Some mating and breeding schemes are capable of taking into account the individual's relative fitness and therefore can be used to get insights into the process of natural selection. Sims live on the surface of a unit sphere. Each sim after birth is assigned the position of its female parent. After birth and before mating and breeding each sim migrates in accordance with a specified migration scheme. Each sim possess a genome, which is an encoding of that sim's genes.

The set of all sims in the simulation forms the set of vertices V in the **population graph** $G_p = (V, E)$. A pair of sims (u, v) belongs to the set of edges E of G_p if the sim u is a parent of the sim v. It is easily noticed that G_p is a DAG, as no sim can be its own ancestor. We call sims capable of interbreeding at time t to be a generation. Like in simuPOP, the framework is designed in such way that generations are discrete, i.e. all sims in the parent generation die before any sims in the children generation are capable of reproduction. This simplifies the construction of the framework.

Finally sims are serialised into json format and output to hard drive for further analysis. The analysis stage consists of computing MRCA of each generation of sims in the population graph.

simuPOP provides most of the same functionality, however there are significant differences between the two programs. Exact comparison is difficult, because simuPOP is mature software in constant development for over a decade, whereas the proposed framework is an effect of 7 months of part-time development.

Most notably, simuPOP does not allow for graph-theoretic analysis of simulated populations, and does not support outputting population graphs into permanent storage for later analysis. Aside from the educational aspect of creating my own framework this was the main reason why the effort was undertaken.

4.1.2 Performance and porability

SimuPOP achieves low memory and CPU footprint by using python's *extension modules*, which allow more direct access to underlying hardware than pure python. However, there are significant drawbacks of using *extension modules*.

Firstly, development is slowed down by higher cognitive burden put on the programmer, as *extension modules* are typically written in C++, which is arguably less user-friendly than pure python. The following snippet of simuPOP's source code serves as good example:

```
PyObject * genoObj = PyTuple_New(alleles.size());
// set value
for (size_t j = 0; j < alleles.size(); ++j)
    PyTuple_SET_ITEM(genoObj, j, PyInt_FromLong(alleles[j]));
return genoObj;
```

which could be equivalently expressed in pure python

```
genoObj = tuple(alleles)
return genoObj
```

Secondly, the use of extension modules limits execution environment of simuPOP to CPython, whereas the framework proposed has been successfully tested both under CPython 2 and 3 (the entire framework), pypi and jython (with exception of the graph analysis module). Using pure python therefore results in much higher portability.

Let us end this note on performance by saying that the performance of the framework is satisfactory for our purpose. Timing one of the simulation scripts (with MRCA analysis) reveals that the simulation of an initial population of 1000 sims for 70 generations accomplishes in under 2 minutes on consumer grade hardware.

4.1.3 Genome

SimuPOP allows for extensive modeling of genomes, such as modeling multiple loci situated on multiple **chromosome**s, even in **polyploid** organisms (simuPOP 2016).

The proposed framework takes much simpler approach consistent with our requirement of investigating a single feature, in its present form allowing for only one **locus** capable of containing only 2 alleles.

4.1.4 Migration

The proposed framework takes a different approach to modeling migration than either simuPOP or Rohde, Olson, and Chang 2004. Both simuPOP and the program of Rohde, Olson and Chang allow for maintaining partial population substructure through simulating many small randomly breeding sub-populations with limited migration

between these sub-populations. The proposed framework is different in that it allows for continuous diffusion of sims on the surface of the world they live in, which is a unit sphere. Appropriate spherical trigonometry was derived using a blend of traditional paper-based computation and Sage (Sage 2014). The README.md file accompanying the source gives sufficient description to reproduce these results.

The sims can migrate on the surface of the sphere using two different schemes, the wanderer scheme, which positions any sim in a random (pole-adjusted) point on the sphere, and homebody, which moves a sim by at most a great-arc distance k on the sphere, in a random direction. Additional types of migrators can be implemented by the user.

4.1.5 Mutations and Inheritance

The framework allows for a simulation of **autosomal** recessive and **autosomal** dominant alleles. The **sex-linked** genes are not supported, unlike in simuPOP.

A mutator – a function defining mutation event, which can change genotype of any number of sims in any generation – can be provided by the user as a python function with signature as expected by the framework. In the experiment the only non-empty mutator is used at the beginning of each simulation to introduce a single individual with an advantageous feature.

4.1.6 Natural Selection

The framework implements the natural selection scheme proposed by Peng and Amos 2010, and discussed in detail in chapter 5. The user can easily define any other natural selection scheme.

4.1.7 Generations

The framework operates using discrete generations, much like simuPOP. Introduction of continuous generations, like in (Rohde, Olson, and Chang 2004) was considered to be conceptually too difficult to implement in the framework given the limited amount of time. It needs to be admitted that modeling populations using discrete generations can lead to discrepancies between the effects of natural selection observed *in silico* against those observed *in vivo* (Waples, Antao, and Luikart 2014). Discrete generations are therefore a limitation of this study.

4.1.8 Mating and Breeding

Two mating and breeding schemes are implemented in the framework and are discussed below. Both mating schemes are based on proximity.

In the first mating scheme the closest two sims of opposite sex on the unit sphere are chosen to form the couple, and are removed from further mating. The process is repeated until no uncoupled sims remain. The corresponding breeding scheme is used to reproduce coupled sims by drawing the number of children for each couple from appropriate probability distribution. This mating and breeding scheme was not used in the simulation for two reasons:

- The size of resulting population is difficult to control. If the expected value of the corresponding probability distribution of the number of children was close to 2, small populations using these schemes were

dying out due to random imbalances between the number of men and women and gradually decreasing sizes of subsequent generations. If the expected value was greater than 2 the populations were exploding exponentially.

- It was difficult to come up with a realistic natural selection scheme, which could be used along these mating and breeding schemes.

It is worth pointing out that described mating scheme is monogamous.

The second mating and breeding scheme is based on the relative advantage of sims possessing an advantageous characteristic against sims without the characteristic. The first parent of any sim in each simulated generation is chosen in accordance with their genetic advantage. The closest sim of opposite sex to the first parent is chosen as the second parent. The process is continued until the next generation achieves the size of the previous generation. We should remark that described mating scheme is also monogamous. This mating and breeding scheme is due to Peng and Amos (2010) and does not suffer from the issues of the previous mating scheme. More detail of this scheme is described in Chapter 5.

Neither the first nor the second mating and breeding scheme incorporate preferential mating (sexual selection), as they are both purely proximity based. This is acknowledged as a limitation of the study. Ideas regarding expansion of the former scheme with preferential mating are addressed in Chapter 6 as an area for further investigation.

4.1.9 Code Correctness

Great care was taken to ensure correctness. As of the project submission no discovered bugs or outstanding issues remain, and extensive unit testing has been in place throughout the project. The framework consists of about 1600 lines of code and contains 25 test cases, which provide about 75% test coverage. The most challenging module, "processGraph.py" is most extensively tested, with 10 test cases and 97% test coverage.

An interesting way of testing the framework was devised to solve the issue of inherent lack of determinism in test case results. As simulated processes are probabilistic in nature, a traditional way of testing the program through an instance of input for which output is known could not be realised. Instead, whenever possible, theoretical consideration of the expected distribution of output was undertaken, and a range of values was chosen as valid. The range was chosen in such way as to make it incredibly unlikely for the test case to fail. For binomial distribution for example, 5 standard deviations around mean were taken to form the range.

4.1.10 Computing MRCA

The most challenging part of the framework was the graph processing module due to a difficult task of computing MRCA. Two algorithms **??** were implemented, and only the former algorithm turned out to satisfy our performance needs. The algorithms are designed very differently, with **??** designed by the author of this study and **??** is a simplified version of an algorithm due to Rohde, Olson, and Chang (2004).

4.1.11 Extensibility

A major change to the library occurred between revisions eadc6756 and 1af29cde. The change was due to a realisation that a particular algorithm chosen to compute MRCA was performing poorly, and the implementation of

Algorithm 1 procedure MRCA1

```
procedure MRCA1 (populationGraph)
    for sim in populationGraph.firstGeneration do
        COUNTDESCENDANTS(sim)
    end for
    return DISCOVERMRCA(populationGraph.lastGeneration)
end procedure
```

Algorithm 2 procedure discoverMRCA

```
procedure DISCOVERMRCA (sims)
    q ← initialise empty queue
    minimumCAcount ← sims.size
    for sim in sims do
        add sim to the rear of q
    end for
    while q is not empty do
        sim ← retrieve and remove an element from the front of q
        if sim.descendantsCount equals minimumCAcount then return sim
        end if
        for parent of sim do
            add parent to the rear of q
        end for
    end while
    return null
end procedure
```

Algorithm 3 procedure countDescendants

```
procedure COUNTDESCENDANTS (sim)
    if sim.descendantsCount not initialised then
        sims.descendantsCount ← 0
    else return sim.descendantsCount
    end if
    for child in sim.children do
        increment sim.descendantsCount by COUNTDESCENDANTS (child)
        if sim.children is empty then
            sim.descendantsCount ← 1
        end if
    end for
    return sim.descendantsCount
end procedure
```

Algorithm 4 procedure MRCA2

```
procedure MRCA2 (populationGraph)
    i ← 0
    q ← initialize empty queue
    for sim in population.lastGeneration do
        sim.tracers ← initialise array of size populationGraph.lastGeneration.size
        for j s.t. 0 ≤ j < populationGraph.lastGeneration.size do
            if j equals i then
                sim.tracers[j] ← true
            else
                sim.tracers[j] ← false
            end if
        end for
        append sim to the rear of q
    end for
    while q not empty do
        sim ← remove and retrieve an element from the front of the q
        for parent of sim do
            if parent.tracers is initialized then
                parent.tracers ← parent.tracers — sim.tracers          ▷ element-wise or
            else
                parent.tracers ← COPY(sim.tracers)                     ▷ a copy of entire array
            end if
            if ALL(parent.tracers) then return parent                  ▷ all elements of the array are true
            end if
            append parent to the rear of q
        end for
    end while
end procedure
```

an alternative algorithm was undertaken. This required several changes across two modules, the module responsible for graph generation and the module responsible for graph processing. The change was accomplished in a few days of work, and the difficulties rather than in code integration lied in it's high cognitive load. Presumably the change would be more difficult if the safety net of extensive test coverage was not present.

4.1.12 Documentation

A reasonable part of the framework is documented and is made available through pydoc (and pydoc3). An instruction on how to access the manual is given in detail in the project's README.md, which accompanies the source code. The source code can be accessed on github (Kurkiewicz 2016)

Chapter 5

Report of research work

This should bring out the insights and new ideas produced by the review and may continue to suggest further practical work, which would test these ideas. Most importantly, the report should clearly state the overall value of the work which has been done - this should not be lost in a welter of detail.

5.1 Selective Sweep in Randomly Mating Population

An experiment was carried out in order to quantify the effect of positive selection on relatedness in a small, randomly mating population. The relatedness was measured using T_{MRCA} (as discussed in chapter 1) of a carefully selected generation in this population.

Throughout the experiment randomness was obtained from python's default random module, which implements Mersenne Twister, a good choice of pseudo-randomness for scientific simulations (Matsumoto and Nishimura 1998). Let us explain the experiment's design.

The initial simulation size was set to be 100 sims, of random sex, and uniformly distributed on the unit sphere. One of the sims, chosen at random, was assigned a **heterozygous** genotype. All other sims were assigned **homozygous** recessive genotypes. The chosen mode of gene expression was **dominant**. The presence of a trait in an invidual would give that individual a comparative breeding advantage of 5% over an individual without the trait. Each subsequent generation of the simulation was obtained as follows. For each sim in the next generation, first parent was chosen at random from the previous generation. The probability of any sim s being chosen as the first parent can be expressed by a formula

$$p(s, S) = \frac{f(s)}{\sum_{s' \in S} f(s')}$$

where S is the set of all sims in the previous generation (including s), and $f(s)$, the sim's absolute fitness, is defined by

$$f(s) = \begin{cases} 1.05, \text{ if sim } s \text{ has the trait} \\ 1, \text{ if sim } s \text{ does not have the trait} \end{cases}$$

The second parent was chosen to be the closest sim of opposite sex (as measured by the great circle distance on the unit sphere). The number of sims in the next generation was adjusted to be equal to the number of sims in the previous generation. After birth and before mating each sim migrated to a randomly chosen place on the

	sample	control
average	6.415	6.37
median	6	6
mode	7	7
min	5	5
max	7	7

Table 5.1: statistics for population of size 100 and relative advantage 1.05

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.258987	0.999987

Table 5.2: p-values of null-hypothesis for population of size 100 and relative advantage 1.05

unit sphere. This model closely follows the one described by Peng and Amos 2010. It differs from the model of Rohde, Olson, and Chang 2004 by usage of discrete generations (the previous generation dies before the next generation can breed) as opposed to continuous real-life like generations.

The simulation was run until the trait had spread to the entire population or the trait died out. In the former case, the simulation was included in the sample, in the latter case the simulation was discarded. It is worth pointing out that the length of each simulation measured by generations was variable, i.e. it did not always take the same number of generations for a simulation to reach full trait saturation.

The investigated sample consisted of 200 simulation runs of the experiment described above. For each simulation in this sample, a control simulation was generated with the relative advantage of possessing the trait removed (all sims equally likely to become the first parent) and the length in generations equal to that of the corresponding observation from the sample. This had given a control sample of 200 sims.

Both the investigated and the control sample were processed in the following way: for each simulation in either sample a generation exactly in the middle of the simulation (rounded down if the number of generations happened to be odd) was chosen, and T_{MRCA} was computed for this generation.

This had resulted in two samples of 200 non-negative integers each. The table 5.1 gives some common statistics for these numbers. Looking at the table we can infer that both the sample and the control are actually collections of integers in the interval $[5, 7]$. We expect a lot of repetitions.

The null hypothesis was chosen so that both the investigated sample and the control sample T_{MRCA}s come from the same distribution. This hypothesis was tested using Mann-Whitney U-test (ties adjusted and continuity corrected) and Kolomogorov-Smirnov (without continuity correction) two sample test. In Table 5.2 we report the p-values.

Analogous experiments were executed for all combinations of relative advantage in the rage [1.05, 1.10] and population sizes in the range [120, 140, 160, 180, 200]. Resulting 10 statistical tests are reported below. The reader is trusted to apply the Bonferroni correction of 11, when interpreting the p-values.

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.475	0.999

Table 5.3: p-values of null-hypothesis for population of size 120 and relative advantage 1.10

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.0962	0.981

Table 5.4: p-values of null-hypothesis for population of size 140 and relative advantage 1.10

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.0663	0.708

Table 5.5: p-values of null-hypothesis for population of size 160 and relative advantage 1.10

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.0406	0.708

Table 5.6: p-values of null-hypothesis for population of size 180 and relative advantage 1.10

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.0859	0.674

Table 5.7: p-values of null-hypothesis for population of size 200 and relative advantage 1.10

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.05	0.999

Table 5.8: p-values of null-hypothesis for population of size 120 and relative advantage 1.05

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.104	0.893

Table 5.9: p-values of null-hypothesis for population of size 140 and relative advantage 1.05

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.244	0.999

Table 5.10: p-values of null-hypothesis for population of size 160 and relative advantage 1.05

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.0609	0.708

Table 5.11: p-values of null-hypothesis for population of size 160 and relative advantage 1.05

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.0769	0.999

Table 5.12: p-values of null-hypothesis for population of size 200 and relative advantage 1.05

Chapter 6

Conclusion

6.1 Inconclusive results

The experiment carried out did not provide evidence allowing to reject the null hypothesis with the cut-off value of 0.05 (0.0045 after Bonferroni correction), which states that selective sweep has no influence on T_{MRCA}. There can be numerous reasons for this outcome. Firstly there might be no influence indeed, or the influence might be so slim, that a very big number of measurements would be required to discover it. Looking at the data it seems intuitive that there is no correlation in T_{MRCA} between cases and controls, which would be greater than 0.5 a generation, but a more formal evaluation of the power of the statistical tests used would be necessary to substantiate this intuition. Further, it could be that the difference actually exists, and is substantial, but our tests did not have enough power to detect it. This again would require a formal analysis of statistical power. An interesting approach would be to set the relative advantage to an unrealistically high value and repeat the experiment in such setting. Would we see any influence of T_{MRCA} in such scenario?

Alternatively, we might try to look for compelling heuristics in support of the negative answer to the research question. A closer look at any of the case simulations reveals a possible explanation: Natural selection seems to be both slow and inconsistent! The Figure 6.1 shows a population of size 120 undergoing selective sweep. The abscissa runs through all generations in the simulation. The blue irregular line shows how many advantageous alleles there are in the population at each generation (at most $2 * n = 240$, where $n = 120$ is the number of individuals. There are 2 alleles per individual). The green line shows T_{MRCA} at each generation. It takes over 300 generations for the characteristic to spread to the entire population. It is hard to expect that such a slow process would have a big impact on something as quick as T_{MRCA}. Colloquially speaking T_{MRCA} might be so quick that it doesn't even notice the selective process happening in the background.

Further, one might not rule out undiscovered bugs having an effect on the result of this investigation. However, the quality and high coverage of unit-tests diminishes the risk. Additionally, it is observed that setting the relative advantage constant to very high values (20 instead of 1.05) results in populations, where the advantageous feature spreads very quickly. Additionally, a careful inspection of the source code by another programmer apart from the investigator could decrease the risk even further.

Eventually, some of our assumptions, such as discrete generations or lack of sexual selection (preferential mating) may not correspond to reality. These issues should be considered in the future work.

Figure 6.1: The x axis runs through all generations in the simulation. The green line shows T_{MRCA} at each generation, the blue line shows how Smany advantageous alleles there are in the population at each generation (at most $2 * n = 240$, where $n = 120$ is the number of individuals. There are 2 alleles per individual).

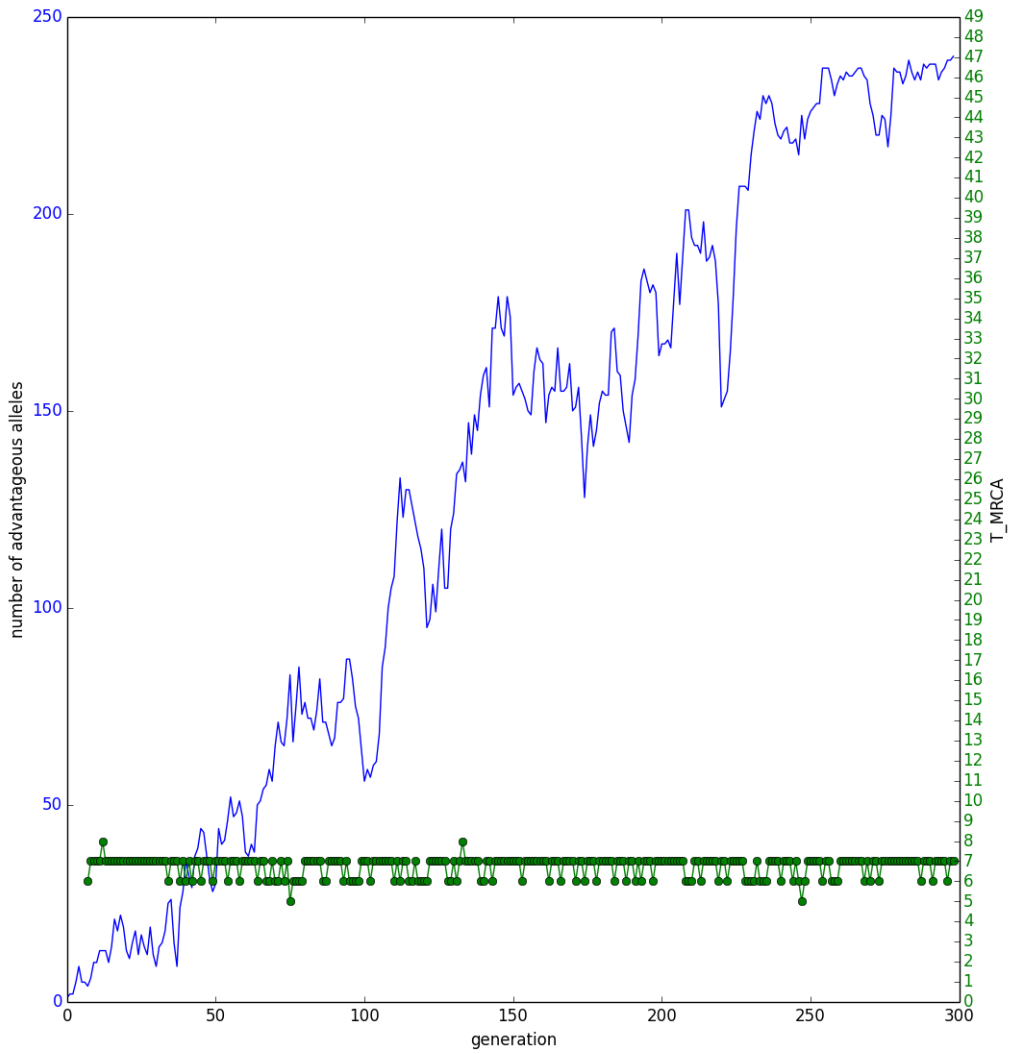

Figure 6.2: T_{MRCA} (green line) and number of advantageous alleles (blue line) for each generation

6.2 Topics for the future

An interesting observations was made during non-rigorous inspection of obtained data. Firstly, only about 4% of advantageous features managed to spread to the entire population in the case study without dying out. It would seem wasteful for nature not to utilise so much reproductive potential. Biologists have looked at sexual selection as both stronger and more consistent form of selective pressure than natural selection (Miller and Svensson 2014). Could it be that switching sexual selection on could boost the spread of the feature considerably? It could make an interesting follow-up study, and another use-case for the implemented framework: to investigate relative strength of natural selection against sexual selection.

Glossary

allele A version of a gene. Each gene can have one or multiple alleles. Existence of many gene versions is the reason for variety within population. Some alleles can corrupt normal body function and result in disease. 4, 22

autosomal An autosomal gene is located on one of the numbered chromosomes (as opposed to the sex chromosomes). Organisms typically have a pair of autosomal chromosomes, one from the mother and one from the father. 1, 4, 5, 11, 22

chromosome chromosome is a collections of genes. Organisms usually have several chromosomes, usually classified into numbered chromosmes, which host autosomal genes, and sex chromosomes, which host sex-linked genes. 10

dominant A dominant inheritance scheme is a scheme in which individuals possessing one or two copies of a gene encoding the trait exhibit the trait. An individual can be called homozygous dominant for a trait if it is homozygous and possesses the trait. 3, 4, 16

heterozygous A heterozygous individual possess two different alleles of a gene, one of which encodes a given trait, and one of which encodes a lack of the trait. 16

homozygous A homozygous individual possesses two identical alleles of a gene, either both encoding a given trait or the lack of it. A homozygous individual without a given trait is called homozygous recessive for that trait. A homozygous individual with a given trait is called homozygous dominant. 16

locus A placeholder for a gene. Locus is situated on a chromosome and can hold a version of the gene – see **allele**. If the locus is situated on one of the numbered chromosomes, the gene is called **autosomal**. If the gene is on one of the sex chromosomes, the gene is called **sex-linked**. 10

MRCA An MRCA of a generation at time t is a common ancestor which is the most recent, i.e. precedes the population by the least number of generations.. 6

polyploid Most organisms have only two versions of each autosomal gene. Some organisms have 4, 6, 8 or more copies of each gene. Such organisms are called polyploid organisms. 10

population graph A population graph $G_p = (V, E)$ is a Directed Acyclic Graph, where V is the set of all simulated individuals, dead or alive, and E is a collection of all pairs of sims (u, v) such that u is a parent of v. 1, 9

recessive The recessive inheritance scheme is a scheme in which only individuals possessing two copies of a gene encoding the trait exhibit the trait. An individual can be called homozygous dominant for a trait if it is homozygous and it does not possess the trait. 3, 4

selective sweep When natural selection exerts positive or negative selective pressure on individuals possessing a given trait the alleles encoding this trait will respectively dominate the population or disappear from the population. The process is referred to as selective sweep.. 6

sex-linked A sex-linked gene is located on one of the sex chromosomes (as opposed to one of the numbered chromosomes). Organisms usually have either one or two copies of sex-linked genes, depending on sex. 4, 11, 22

Bibliography

Allentoft, Morten E. et al. (June 2015). "Population genomics of Bronze Age Eurasia". In: *Nature* 522.7555, pp. 167–172. ISSN: 0028-0836. DOI: 10.1038/nature14507. URL: http://www.nature.com/nature/journal/v522/n7555/full/nature14507.html (visited on 02/19/2016).

Corcos, Alain F. and Floyd V. Monaghan (1993). *Gregor Mendel's Experiments on plant hybrids: a guided study*. Rutgers University Press. ISBN: 0813519217.

Correns, Carl (1900). "Gregor Mendels Regel über das Verhalten der Nachkommenschaft der Bastarde." In: *Berichte der Deutschen Botanischen Gesellschaft* 18, pp. 158–168.

Darwin, Charles (Oct. 1859). *On the origin of species. Or the preservation of favoured races in the struggle for life*. London, Albemarle Street: John Murray.

Dayhoff, Margaret (1973). *Atlas of protein sequence and structure. Supplement*. Washington, D.C: National Biomedical Research Foundation. ISBN: 0912466073.

DeVries, Hugo (1900). "Sur la fécondation hybride de l'albumen". In: *Biologisches Zentralblatt* 20, pp. 129–130.

Dolinoy, Dana C., Dale Huang, and Randy L. Jirtle (Aug. 2007). "Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development". en. In: *Proceedings of the National Academy of Sciences* 104.32, pp. 13056–13061. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0703739104. URL: http://www.pnas.org/content/104/32/13056 (visited on 03/13/2016).

Duforet-Frebourg, Nicolas et al. (2016). "Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data". In: *Molecular Biology and Evolution* 33.4, pp. 1082–1093. DOI: 10.1093/molbev/msv334. eprint: http://mbe.oxfordjournals.org/content/33/4/1082.full.pdf+html. URL: http://mbe.oxfordjournals.org/content/33/4/1082.abstract.

Fisher, Ronald (1918). In: *Transactions of the Royal Society of Edinburgh* 52, pp. 399–433. URL: http://www.indiana.edu/~curtweb/L567/readings/Fisher1918.pdf.

GSR (Mar. 24, 2016). *Genetic Simulation Resources*. URL: http://popmodels.cancercontrol.cancer.gov/gsr/packages/simupop/#applications (visited on 03/24/2016).

Hardy, G. H. (Apr. 1908). "Mendelian Proportions in a Mixed Population". In: *Science* 28.706, pp. 49–50.

Kimura, M. (1968). "Evolutionary Rate at the Molecular Level". In: *Nature* 217.5129, pp. 624–626. DOI: 10.1038/217624a0.

Kurkiewicz, Adam (Mar. 25, 2016). *naturalSelection source code*. URL: github.com/picrin/naturalSelectoin (visited on 03/25/2016).

Maher, Brendan (Nov. 2008). "Personal genomes: The case of the missing heritability". In: *Nature* 456, pp. 18–21. DOI: 10.1038/456018a.

Matsumoto, Makoto and Takuji Nishimura (Jan. 1998). "Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator". In: *ACM Trans. Model. Comput. Simul.* 8.1, pp. 3–30. ISSN: 1049-3301. DOI: 10.1145/272991.272995. URL: http://doi.acm.org/10.1145/272991.272995.

Miller, Christine W. and Erik I. Svensson (2014). "Sexual Selection in Complex Environments". In: *Annual Review of Entomology* 59.1. PMID: 24160419, pp. 427–445. DOI: 10.1146/annurev-ento-011613-162044. eprint: http://dx.doi.org/10.1146/annurev-ento-011613-162044. URL: http://dx.doi.org/10.1146/annurev-ento-011613-162044.

OMIM (2016). *Online Mendelian Inheritance in Man*. URL: http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm (visited on 03/13/2016).

Ong-Abdullah, Meilina et al. (Sept. 2015). "Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm". en. In: *Nature* 525.7570, pp. 533–537. ISSN: 0028-0836. DOI: 10.1038/nature15365. URL: http://www.nature.com/nature/journal/v525/n7570/abs/nature15365.html (visited on 03/13/2016).

Peng, Bo and Christopher I. Amos (2010). "Forward-time simulation of realistic samples for genome-wide association studies". In: *BMC Bioinformatics* 11, p. 442. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-442. URL: http://dx.doi.org/10.1186/1471-2105-11-442.

Peng, Bo and Marek Kimmel (Sept. 2005). "simuPOP: a forward-time population genetics simulation environment". en. In: *Bioinformatics* 21.18, pp. 3686–3687. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bti584. URL: http://bioinformatics.oxfordjournals.org/content/21/18/3686.

Rohde, Douglas L. T., Steve Olson, and Joseph T. Chang (Sept. 2004). "Modelling the recent common ancestry of all living humans". en. In: *Nature* 431.7008, pp. 562–566. ISSN: 0028-0836. DOI: 10.1038/nature02842. URL: http://www.nature.com/nature/journal/v431/n7008/abs/nature02842.html.

Sage (2014). *Sage Mathematics Software (Version 6.1.1)*. http://www.sagemath.org.

Silventoinen, Karri et al. (Oct. 2003). "Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries". In: *Twin Research and Human Genetics* 6 (05), pp. 399–408. ISSN: 1839-2628. DOI: 10.1375/twin.6.5.399. URL: http://journals.cambridge.org/article_S1369052300004001.

simuPOP (2016). *simuPOP source code*. URL: http://sourceforge.net/p/simupop/code/HEAD/tree/trunk/src/individual.cpp (visited on 03/13/2016).

Theunissen, L. T. and Enid Perlin-West (1989). *Eugène Dubois and the Ape-Man from Java*. Kluwer Academic Publishers. ISBN: 1556080816.

Tschermak-Seysenegg, Erich Von (1900). "Ueber künstliche Kreuzung bei Pisum sativum." In: *Berichte der Deutschen Botanischen Gesellschaft* 18, pp. 232–239.

Visscher, Peter M et al. (Mar. 2006). "Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings". In: *PLoS Genet* 2.3, e41. DOI: 10.1371/journal.pgen.0020041. URL: http://dx.plos.org/10.1371/journal.pgen.0020041.

Wakeley, John (2016). *Natural Selection and Coalescent Theory*.

Waples, Robin S., Tiago Antao, and Gordon Luikart (2014). "Effects of Overlapping Generations on Linkage Disequilibrium Estimates of Effective Population Size". In: *Genetics*. ISSN: 0016-6731. DOI: 10.1534/genetics.114.164822. eprint: http://www.genetics.org/content/early/2014/04/08/genetics.114.164822.full.pdf. URL: http://www.genetics.org/content/early/2014/04/08/genetics.114.164822.

Watson, J. D. and F. H. C. Crick (1953). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* 171.4356, pp. 737–738.

Yehuda, Rachel et al. (Aug. 2015). "Holocaust Exposure Induced Intergenerational Effects on FKBP5 Methylation". English. In: *Biological Psychiatry*. ISSN: 0006-3223. DOI: 10.1016/j.biopsych.2015.08.005. URL: http://www.biologicalpsychiatryjournal.com/article/S0006322315006526/abstract (visited on 03/13/2016).

Yule, Udny (1902). "Mendel's laws and their probable relation to intra-racial heredity". In: *New Phytol* 1.193–207, pp. 222–238.