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 24, 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. A file created in such way is processed algorithmically and statistically 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 furhter 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

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 of 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 this question 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 capable of inter-breeding 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). The choice of the parameters is discussed in chapter 3, along with the appropriate methodology, to which we prepare ground in chapter 2. 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 results are presented in chapter 4 and the conclusions will be drawn in chapter 5.

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 [find the original works of DeVries, Correns and Tscherma, Roberts, H. F. Plant Hybridization before Mendel. Princeton: Princeton University Press, 1929.], 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 [cite this guy whose name starts with U]. 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**.*

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

Table 2.1: Possible offspring genotypes with corresponding probabilities

We have to remark that these assumptions may not be obviously 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.

Finally, we have to remark that the assumption 3 may not be correct in rare cases of **sex-linked** characteristics (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 using a table much like 2.1, 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 Fischer

The work of Ronald Fischer had drastic influence on the landcape of both statistics and quantitative biology. Apart from introducing influential statistical terminology (such as the concept of variance, defined as standard deviation squared) [todo quote original Fischer's paper], Fischer 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, Fischer was behind the realisation that sexual selection is comparatively stronger, and more consistent than natural selection. He formulated the concept of Fischerian runaway and proposed the sexy sons hypothesis. 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) 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 individuals in simulated populations assigned appropriate probability functions.

2.4 Forward-Time Modelling

for this let us echo concerns raised by Peng and Amos 2010

2.5 Population graph

Write about Rohde and Olsen's work. In particular, write about migration around various islands. Discuss the use of MRCA and IA.

Chapter 3

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.

3.1   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 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.

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

Table 3.1: statistics for Selective Sweep in Randomly Mating Population experiment

	Mann-Whitney	Kolomogorov-Smirnov
p-values	0.258987	0.999987

Table 3.2: p-values of null-hypothesis for Selective Sweep in Randomly Mating Population experiment

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 3.1 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 3.2 we report the p-values.

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

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

Mine is more accurate because of the unit sphere.

4.1.5 Mutations and Inheritance

4.1.6 Natural Selection

4.1.7 Generations

discrete generations as in simuPOP. This is unlike Rohde and Olhsen.

4.1.8 Mating and Breeding

Depending on capture difficulty in random mating -¿ random fluctuations in population

Chapter 5

Conclusion

This section should summarize the work and its significance, comparing practical results with theoretical expectations where appropriate. Suggestions for possible extensions, or research projects arising from the work, may also be included here.

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, 13

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, 13

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, 7

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. 7

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. 7

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

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

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 the sex. 4, 13

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.

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.

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`.

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`.

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`.

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`.

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.

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`.

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).