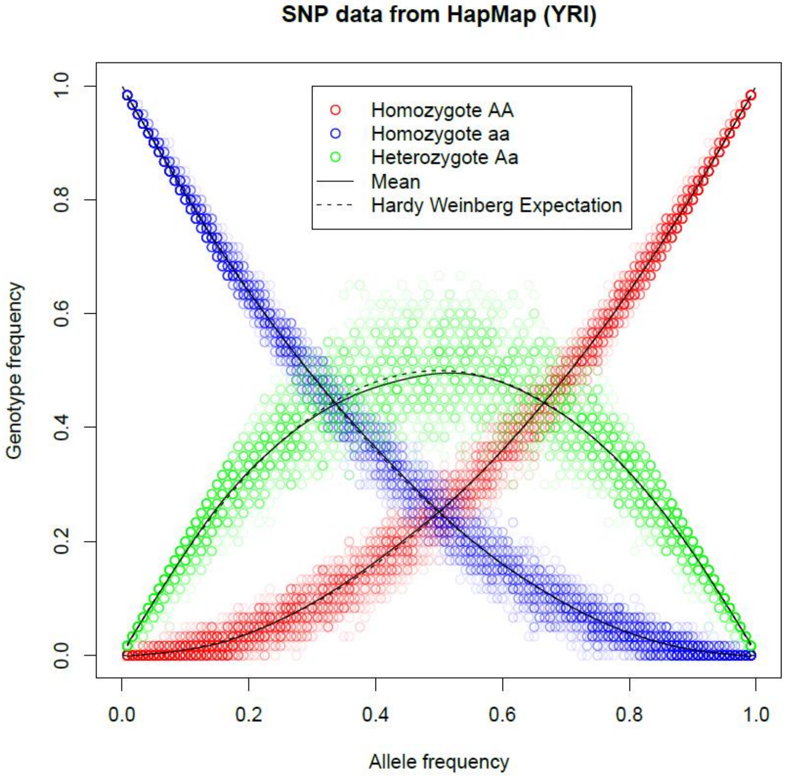
Population Genetics

2023



Ida Moltke
Hans R. Siegismund

Section for Computational and RNA Biology
Department of Biology
University of Copenhagen

Contents

Please read the highlighted content before you come the first time☺.

Welcome to the population genetics course!

This document contains important practical information about the population genetics course ranging from the course schedule to information about how to prepare for the computer use during the course. Please read the parts highlighted in red in “Contents” before the course starts. Also, please note that this file includes links to webpages, articles in journals, books and email addresses like this one:

Ida Moltke
Hans R. Siegismund

Course description

The amount of molecular genetic data (especially nucleotide sequences) has increased tremendously in recent years and is expected to explode as the next generation sequencing methods become standard tools. This has implications for a wide spectrum of biological disciplines spanning conservation genetics, molecular ecology, molecular medicine, genome research and evolutionary biology. The purpose of the course is to provide the students with knowledge about the principles of population genetics and phylogenetics and their applications in the diverse areas mentioned above. In addition, the course will train the students to choose suitable methods to analyze molecular genetic data.

Students with a limited background in bioinformatics will be offered additional exercises in bioinformatics skills necessary to handle the large amount of genomic data that are available today. This includes an introduction to the operating system Linux and the R environment for statistical computing.

By completing the course the student can:

- employ basic population genetic and phylogenetic principles
- discuss, put into perspective, and criticize original research papers in population genetics and phylogenetics
- choose the most suitable molecular methods to analyze a particular hypothesis
- choose the most suitable analytical tools to analyze molecular genetic data
- perform statistical analyses of population genetic and phylogenetic data, present the results, and put them into perspective

Instructors

AA:	Anders Albrechtsen, Bioinformatics, Department of Biology
AB-O:	Anna Brüniche-Olsen, Bioinformatics, Department of Biology
DADG:	David A. D. Garzon, Globe Institute
FR:	Fernando Racimo, Globe Institute
GGE:	Genis Garcia Erill, Bioinformatics, Department of Biology
HRS:	Hans R. Siegismund, Bioinformatics, Department of Biology
IM:	Ida Moltke, Bioinformatics, Department of Biology
PF:	Peter Frandsen, Copenhagen Zoo & Bioinformatics, Department of Biology
PP:	Patricia Pecnerová, Bioinformatics, Department of Biology
RB:	Renzo Fidel Ferdinando Balboa
RH:	Rasmus Heller, Bioinformatics, Department of Biology

Guest lecturer

Bo T. Simonsen (BTS)
Head of Division
Section of Forensic Genetics
University of Copenhagen

Shyam Gopalakrishnan (SG)
Globe Institute
University of Copenhagen

Preliminary schedule

Please note that the hours indicated in blue below are intended for those that need to be introduced to Linux and R, which we use for analyzing data sets in the exercises and in the projects. Also, please note that a few lectures indicated in red start early and end late.

The first six weeks will be used for lectures and exercises while the last two weeks are used for project work.

(L) Lecture
(E) Exercise
(S) Seminar
NOTE: THE TABLE ON THE NEXT PAGE IS A PRELIMINARY SCHEDULE:

Program for Population Genetics 2023

	Monday	Wednesday
Week 6	**6 February**	**8 February**
		8-12 (RB+TAs) Introduction to Linux and R (Linux part) - for those who do NOT know Linux
	13-16 (IM) Introduction to the course (L) Introduction to population genetics (L)	13-17 Introduction to Linux and R (R part) - for those who do NOT know R A brief introduction to parameter estimation - for those who know R
Week 7	**13 February**	**15 February**
		9-12 (FR) Genetic drift and basic coalescence theory (L/E)
	13-14 (HRS) Genetic diversity: from phenotype over chromosomes, genes to DNA (L) 14-16 (GE) Exercise, Genetic diversity in Chimps and humans (E)	13-16 (RH) Coalescence theory: Relating theory to data (L/E)
Week 8	**20 February**	**22 February**
		9-12 (FR) Estimation of effective population size (L/E)
	13-14 (SG) Linkage disequilibrium (L/E) 14-16 (SG) Linkage disequilibrium in gorillas (E)	13-16 (RH) Population history and demography (L/E)
Week 9	**27 February**	**1 March**
		9-11 (AA/Genis) More population structure + selection (L/E) 11-12 Research talk by Patricia Pečnerová
	13-14 (IM) Population subdivision and admixture (L) 14-16 (IM) PCA, Admixture (Chimp data) (E)	13-16 (AA) Selection (L/E)
Week 10	**6 March**	**8 March**
		9-11 (FR) Archaic admixture and introgression (L/E) 11-12 Research talk by Bo Simonsen
	13-14 More theory about selection (L) (HS) 14-15 Research talk by Ida Moltke 15-16 Student projects	13-16 (AA) Inbreeding and relatedness across genomes (L/E)
Week 11	**13 March**	**15 March**
		9-12 (DD) Phylogenetic inference (L/E)
	13-16 Genome-wide association studies (L/E) (IM)	13-14 Phylogenetic inference continued (L/E) 14-15 Research talk by Shyam Gopalakrishnan 15-16 Evaluation of course
Week 12	**20 March**	**22 March**
		Project work
	Project work	Project work
Week 13	**27 March**	**29 March**
		Project work
	Project work	Project work
Week 14	**3 April**	**5 April**
	Easter holiday	**Easter holiday**
Week 15		
	ORAL EXAM week	**ORAL EXAM week**

Teachers

IM	Ida Moltke (kursusansvarlig)
HRS	Hans Siegismund (kursusansvarlig)
RH	Rasmus Heller
AA	Anders Albrechtsen
FR	Fernando Racimo
SG	Shyam Gopalakrishnan
DD	David du Chêne
GE	Genis Erill
PP	Patricia Pečnerová
FFS	Frederik Filip Stæger
RB	Renzo Fidel Ferdinando Balboa

Practical information about computer programs and operating system used

The course consists of a mixture of lectures, theoretical exercises and computer exercises. **Therefore, please bring your own laptop to ALL lectures and exercises.**

Many of the exercises we use in this course are based on up-to-date data sets, which can get quite large. The analytical tools we use are often most practical to implement on a Linux operating system. In order for all students to have access to a Linux operating system we use a virtual server provide by Science IT.

Please make sure that you have access to the servers before the first Monday in the course.

Population genetic programs

There are a large number of population genetic programs available. For a (somewhat outdated) review about them you can take a look at Excoffier & Heckel (2006) Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics* 7, 745-758. Many of these programs are still in use and you might encounter references to them when you read papers.

We are going to use the following programs for different purposes. Note, the citations describing the programs/packages are from the homepages of programs.

Data management and statistical analyses

Plink: Whole genome data analysis toolset

> "PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.
>
> The focus of PLINK is purely on analysis of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data)"

Admixture: fast ancestry estimation

> "ADMIXTURE is a software tool for maximum likelihood estimation of individual ancestries from multilocus SNP genotype datasets. It uses the same statistical model as STRUCTURE but calculates estimates much more rapidly using a fast numerical optimization algorithm."

R

The R Project for Statistical Computing

> "R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror."

There are a lot of free packages available for use in R. We will use the following:

SNPMatrix

“Implements classes and methods for large-scale SNP association studies”

Do not download it from the homepage mentioned above. We will use an older version available from another server.

HardyWeinberg

“Contains tools for exploring Hardy-Weinberg equilibrium for bi and multi-allelic genetic marker data. All classical tests (chi-square, exact, likelihood-ratio and permutation tests) with bi-allelic variants are included in the package, as well as functions for power computation and for the simulation of marker data under equilibrium and disequilibrium. … Implements several graphics for exploring the equilibrium status of a large set of bi-allelic markers: ternary plots with acceptance regions, log-ratio plots and Q-Q plots.”

Ape: Analysis of Phylogenetics and Evolution

“Ape is a package written in R for the analysis of phylogenetics and evolution. It is reasonably used within the community of evolutionary biologists for data analysis and as a framework for the development of new analytical methods.”

Phangorn: Phylogenetic Analysis in R

“Phylogenetic analysis in R: Estimation of phylogenetic trees and networks using Maximum Likelihood, Maximum Parsimony, distance methods and Hadamard conjugation.”

You can read a short tutorial about the package here: “Estimating phylogenetic trees with phangorn”.

Simulation programs

PopG genetic simulation program

“This is a one-locus, two-allele genetic simulation program for use by students. It simulates multiple populations and allows you to see the effect of natural selection, mutation, migration, and genetic drift. It is freely downloadable. It is written in Java, and will run on Windows, Mac OS X, and Linux systems if they have Java installed on them”

Coalescent.dk

“This page is home to two educational tools for understanding the coalescent process. The tools need to be downloaded and run as java applications.”

“Wright-Fisher animator

The Wright-Fisher application animates the random sampling of genes under the discrete Wright-Fisher model and it is possible to track the number of descendant genes as well as the ancestry of genes in the final generation.

Hudson animator

The Hudson application animates the coalescent process in continuous time. The basic coalescent process as well as the coalescent with

> recombination, coalescent with growth, coalescent with subdivision and coalescent with selection can be studied, tracking the time and type of events as well as their consequences to the coalescent tree”

Fastsimcoal2: fast sequential markov coalescent simulation of genomic data under complex evolutionary models

> “While preserving all the simulation flexibility of simcoal2, fastsimcoal is now implemented under a faster continous-time sequential Markovian coalescent approximation, allowing it to efficiently generate genetic diversity for different types of markers along large genomic regions, for both present or ancient samples. It includes a parameter sampler allowing its integration into Bayesian or likelihood parameter estimation procedure.
>
> fastsimcoal can handle very complex evolutionary scenarios including an arbitrary migration matrix between samples, historical events allowing for population resize, population fusion and fission, admixture events, changes in migration matrix, or changes in population growth rates. The time of sampling can be specified independently for each sample, allowing for serial sampling in the same or in different populations.
>
> Different markers, such as DNA sequences, SNP, STR (microsatellite) or multi-locus allelic data can be generated under a variety of mutation models (e.g. finite- and infinite-site models for DNA sequences, stepwise or generalized stepwise mutation model for STRs data, infinite-allele model for standard multi-allelic data).”

Conversion programs

Population genetic data are stored in a wealth of different formats, which makes it hard to combine different programs. Fortunately, Heidi Lischer has written a program PGDSpider, a program for converting data between population genetics programs (with many different formats...). PGDSpider is written in Java and therefore platform independent.

Background reading

We recommend as textbook:

Nielsen, R. & M. Slatkin 2013. *An Introduction to Population Genetics: Theory and Applications.* Sinauer Associates.

Errata: http://cteg.berkeley.edu/errata.html (Link is broken)

"An Introduction to Population Genetics is intended as a text for a one-semester biology course in population genetics at the undergraduate or graduate levels. The goal of the book is to introduce both classical population genetics theory developed in terms of allele and haplotype frequencies and modern population genetics theory developed in terms of coalescent theory. Numerous applications of theory to problems that arise in the study of human and other populations are presented. Appendices provide the mathematical background necessary to understand the basic theory."

We will list the recommended reading for the different lectures and exercises from that book on Absalon. However, if you are interested in additional reading material there are **other textbooks**, which cover most of what is needed in our course, like

Hahn, M. 2018. *Molecular Population Genetics,* Sinauer Associates.

"Molecular Population Genetics is a general text covering one of the most active and exciting areas in biology. Combining advances in molecular biology and genomics with mathematical and empirical findings from population genetics, work in molecular population genetics has uncovered the extraordinary history of natural selection and demographic shifts in many organisms, including humans."

Coop, G. 2020. *Population and Quantitative Genetics*.

"This book was developed from my set of notes for the Population Biology graduate group core class (PBGG) and Undergraduate Population and Quantitative Genetics class (EVE102) at UC Davis. Thanks to the many students who've read these notes and suggested improvements. Thanks to Simon Aeschbacher, Vince Buffalo, and Erin Calfee who read and extensively edited earlier drafts of these notes. To illustrate these notes I've used old scientific and natural history illustrations, in part because they are out of copyright but mainly because they bring me joy. Many of the old images come from Biodiversity Heritage Library a consortium of natural history institutions that are digitizing their collections and make them freely available online. If you enjoy the images consider donating to the BHL. Many of the data and simulation graphics in the book were prepared in R, the code for each is linked to from the caption of each figure. In many cases data were extracted from old figures using the WebPlotDigitizer tool, as such I advise re-extracting the data if you wish to use it for research purposes."

(The notes can be downloaded from https://github.com/cooplab/popgen-notes)

Hartl, D.L. & A.G. Clark 2007, *Principles of Population Genetics,* Fourth edition, Sinauer Associates (Good, but a little bit old.)

Hedrick, P. W. 2011. *Genetics of Populations*, 4. ed., Jones and Bartlett Publishers, Sudbury, MA. (Also a very good book)

Another book that can be **freely downloaded** has been written by Joe Felsenstein (2019) is *Theoretical Evolutionary Genetics*. It covers our course but is relatively theoretical.

A **comprehensive handbook** about population genomics is

Balding, D.J., Ida Moltke, J. Marioni (Eds.) 2019. *Handbook of Statistical Genomics*. 4th edition, Wiley.

> *"A timely update of a highly popular handbook on statistical genomics*
> This new, two-volume edition of a classic text provides a thorough introduction to statistical genomics, a vital resource for advanced graduate students, early-career researchers and new entrants to the field. It introduces new and updated information on developments that have occurred since the 3rd edition. Widely regarded as the reference work in the field, it features new chapters focusing on statistical aspects of data generated by new sequencing technologies, including sequence-based functional assays. It expands on previous coverage of the many processes between genotype and phenotype, including gene expression and epigenetics, as well as metabolomics. It also examines population genetics and evolutionary models and inference, with new chapters on the multi-species coalescent, admixture and ancient DNA, as well as genetic association studies including causal analyses and variant interpretation.
>
> ⋮
>
> The *Handbook of Statistical Genomics* is an excellent introductory text for advanced graduate students and early-career researchers involved in statistical genetics."

This is a relatively expensive set of books ($370), but if you are logged in as a student at Copenhagen University, you can have it for free here☺:
https://onlinelibrary.wiley.com/doi/book/10.1002/9781119487845

For those especially interested in the **coalescent**, two books can be recommended

Hein, J., M. H. Schierup, and C. Wiuf. 2005. *Gene Genealogies, Variation, and Evolution*. Oxford University Press, New York, NY.

Wakeley, J. 2009. *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, CO.

Classical books on population genetics:

Crow, J. F., and M. Kimura. 1970. *An Introduction to Population Genetics Theory*. Harper & Row, New York, NY. (Rather theoretical)

Ewens, W. J. 2004. *Mathematical Population Genetics. I. Theoretical Introduction*, 2nd ed. Springer-Verlag, Berlin. (Rather theoretical. We are still waiting for volume II…)

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection*, The Clarendon Press. (A classic. You can read it here: https://archive.org/details/geneticaltheoryo031631mbp)

Haldane, J. B. S. 1932. *The Causes of Evolution*. Longmans, Green, & Co., Ltd., London. (Another classic. You can read it here: https://archive.org/details/causesofevolutio00hald_0)

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge. (From the father of the neutral theory.) You can read it here https://www.cambridge.org/core/books/neutral-theory-of-molecular-evolution/0FF60E9F47915B17FFA2620C49400632

Wright, S. 1968-1978. *Evolution and the Genetics of Populations*, 4 vols. University of Chicago Press, Chicago, IL. (A four volume treatise from one of the three founders of population genetics. Sewall Wright lived from 1889 to 1988. He started publishing scientific papers in 1914 and published his last paper in 1988 at the age of 99.)

Review papers

Here is a list of review papers about various phylogenetic and population genetic topics. When you are logged in with an account at the university, you have access to the papers by clicking on the links.

Phylogenetics

Mario dos Reis, Philip C. J. Donoghue & Ziheng Yang 2016. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*17, 71–80

Paschalia Kapli, Ziheng Yang & Maximilian J. Telford 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics 21*, 428–444

Ziheng Yang, Bruce Rannala 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* 13, 303–314.

Computer simulations

Sean Hoban, Giorgio Bertorelle, Oscar E. Gaggiotti 2012 Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics* 13, 110–122

Mutation rates

Aylwyn Scally and Richard Durbin 2012. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics* 13, 704–753

Michael Lynch, Matthew S. Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W. Kelley Thomas & Patricia L. Foster 2016. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* 17: 745–714.

Vladimir B. Seplyarskiy & Shamil Sunyaev 2021. The origin of human mutation in light of genomic data. *Nature Reviews Genetics* 22: 672–686.

Genetic diversity

Hans Ellegren & Nicolas Galtier 2016. Determinants of genetic diversity. *Nature Reviews Genetics* 17, 422–433.

Madlen Stange, Rowan D. H. Barrett & Andrew P. Hendry 2021. The importance of genomic variation for biodiversity, ecosystems and people. Nature Reviews Genetics *22*: 89–105

Inbreeding, mutation load and inbreeding depression

Giorgio Bertorelle, Francesca Raffini, Mirte Bosse, Chiara Bortoluzzi, Alessio Iannucci, Emiliano Trucchi, Hernán E. Morales & Cock van Oosterhou 2022. Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics* 23: 492–503.

Francisco C. Ceballos, Peter K. Joshi, David W. Clark, Michèle Ramsay & James F. Wilson. 2018. Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews Genetics* 19: 220–235.

Deborah Charlesworth, John H. Willis. 2009 The genetics of inbreeding depression. *Nature Reviews Genetics* 10, 783–796

Philip W. Hedrick, Aurora Garcia-Dorado 2016. Understanding Inbreeding Depression, Purging, and Genetic Rescue. Trends in Ecology & Evolution 12, 940 –952.

Brenna M. Henn, Laura R. Botigué, Carlos D. Bustamante, Andrew G. Clark & Simon Gravel. 2015. Estimating the mutation load in human genomes. *Nature Reviews Genetics* 16, 333–343.

Relatedness

Doug Speed and David J. Balding. 2015. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* 16, 33–43.

Population structure

Kent E. Holsinger and Bruce S. Weir 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics* 10, 639–650.

John Novembre & Anna Di Rienzo 2009 Spatial patterns of variation due to natural selection in humans.. *Nature Reviews Genetics* 10, 745–755

Effective population size

Brian Charlesworth 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10, 195–205

J Wang, E Santiago and A Caballero 2016. Prediction and estimation of effective population size. *Heredity* 117, 193–206

Coalescent theory

Siavash Mirarab, Luay Nakhleh, and Tandy Warnow 2021. Multispecies Coalescent: Theory and Applications in Phylogenetics. Annual Review of Ecology, Evolution, and Systematics 52: 247-268.

Magnus Nordborg 2007. Coalescent Theory. *Handbook of Statistical Genetics*, Third Edition . Edited by D . J. Balding, M . Bishop and C. Cannings. John Wiley & Sons, Ltd.

Noah A. Rosenberg and Magnus Nordborg 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphism. *Nature Reviews Genetics* 3, 380–390

Natural selection

Laurence D. Hurst 2009. Genetics and the understanding of selection. *Nature Reviews Genetics* 10, 83–93.

Rasmus Nielsen 2005. Molecular Signatures of Natural Selection. *Annu. Rev. Genet*. 39: 197–218

Nielsen et al. 2007 Recent and ongoing selection in the human genome. *Nature Reviews Genetics* 8: 857–868.

H. Allen Orr 2009. Fitness and its role in evolutionary genetics *Nature Reviews Genetics* 10, 531–539

Linkage disequilibrium

Slatkin, M. 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* 9, 477–485

Genome wide association mapping

Jason Flannick & Jose C. Florez 2016. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nature Reviews Genetics* 17, 535–549.

Teo, Y.-Y., Kerrin S. Small and Dominic P. McCarthy et al. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 9: 357-369.

Rosenberg et al 2010. Genome-wide association studies in diverse populations. *Nature Reviews Genetics* 11, 356–366.

Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré & David Meyre 2019. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* 20: 467–484.

Kwiatkowski 2010. Methodological challenges of genome-wide association analysis in Africa. *Nature Reviews Genetics* 10, 149–160.

Quantitative genetics

Peter M. Visscher, William G. Hill & Naomi R. Wray 2008 Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics* 9, 255–266.

Manolio, TA *et al.* 2009. Finding the missing heritability of complex diseases. *Nature* 461, 747–753.

Human evolution

Joshua G. Schraiber & Joshua M. Akey 2015 *Nature Reviews Genetics* Methods and models for unravelling human evolutionary history 16, 727–740.

Rachel M. Sherman & Steven L. Salzberg. 2021. Pan-genomics in the human genome era. *Nature Reviews Genetics* 21: 243–254.

Exam

The course ends with project work in the last two weeks where a group of students (in the order of 5) analyzes a real data set with the tools that have been learned in the previous weeks. The groups should consist of students from different educational backgrounds. Each group hands in a paper written in a style from a scientific journal. (Title, Authors, Abstract, Introduction, Results, Discussion, References). It should not exceed 10 pages.

The paper is handed in on Absalon, where it is checked for plagiarism. You can read more about plagiarism and how to avoid it at "Stop plagiarism".

The paper is presented individually at an oral defense that last for 20 minutes. The presentation should last at most 10 minutes. This is followed by a discussion of the presentation, the paper, and relevant population genetic topics. Please note that you are expected to have read all parts of the textbook that have been used during the course. This is basically the whole book (Nielsen, R. & M. Slatkin 2013) except Chapter 10. You will be asked questions at the exam that are not directly related to the specific project you have done but have been covered during the course. The exam is with external censorship.