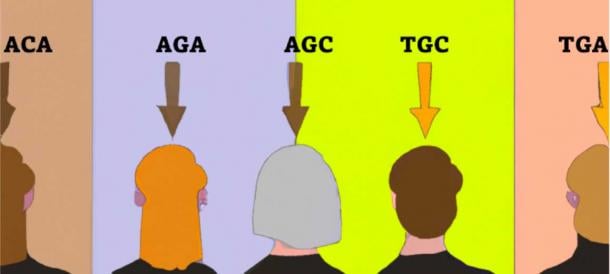
**BINP29 *Population Genetic* Projects**



Analyzing ancient DNA with home DNA tests, scientists can identify specific mutations and determine how your traits have been passed on to you over time. (Enkigen Genetics Limited)

Table of Contents

[What is a bioinformatic project? 3](#_Toc125459341)

[From an idea to a bioinformatic application 3](#_Toc125459342)

[Proposed population genetic projects 4](#_Toc125459343)

[Product design 5](#_Toc125459344)

[Oral presentation 6](#_Toc125459345)

[Written report 6](#_Toc125459346)

[Submission 6](#_Toc125459347)

[Project evaluations 6](#_Toc125459348)

[Writing the paper 6](#_Toc125459349)

[Other guidelines 7](#_Toc125459350)

[FAQ 8](#_Toc125459351)

**Instructions for BINP29 *Population Genetic* Projects**

Welcome to the final part of BINP29 – the development of a Bioinformatics project in population genetics of your choosing!

The goals of BINP29 Projects are to teach you how to develop software by allowing you to sharpen your coding skills, get acquainted with population genetics, and expand your GitHub portfolio. You will also learn how to write an academic paper and present your work.

BINP29 projects include a **development project**, an **oral presentation**, and a **written paper**. Before we start bouncing off ideas for projects, let us first understand what constitutes a bioinformatics project.

# What is a bioinformatic project?

You should develop software (e.g., web application, script, tool) and apply it to some population genetic dataset. You need to think of a problem in population genetics that your code solves and explain why it is an important problem. Consider the following examples:

|  |  |
| --- | --- |
| **Not a biological problem** | **A biological problem** |
| Counting the number of GC% in the 8th position of the DNA sequence is not biologically informative because the 8th position of genomes has no biological function or meaning. | Comparing heterozygosity levels in different populations is informative to study inbreeding patterns. |

If you decide to develop a new tool, you should include 1-3 examples (an example can contain multiple genomes) of how you apply your code to one or more datasets and explain why it is useful.

**Your end goal is to build a useful tool for the science community.**

With that in mind, let us now consider some ideas for bioinformatics projects.

# From an idea to a bioinformatic application

You have different options to choose from in developing your bioinformatics application. Any programming language could be used. You can choose to develop one or more of the following:

• A software that connects to a database

• A web application in pure JavaScript

• A web application that allows querying files

• A web application that connects to a database

Below are suggestions for a Bioinformatics project. You may expand on these suggestions as you wish or develop variants of these ideas, or combine them with other features, for example, a toolbox that is accessible through a web application (note, complex applications would receive a higher grade in the *Software* category).

**All the projects must be pre-approved by the teacher. The work is independent. Two students CAN NOT select the same project.**

# Proposed population genetic projects

1. We provide a dataset of identity by descent (IBD) distances for Europeans, Central Asians, and North Africans. Write an interface for this dataset (see IBD data.zip). The data include IBD distances at the individual (individual-individual) and population (population-population) levels.
   1. Develop a web application that allows choosing populations and drawing circus plots (Linlin Cao).
   2. Develop a web application to choose populations and draw the IBD patterns on a map.
2. The table (Eurasian - Dataset\_tims.xlsx) also has the Y and mtDNA haplogroup information.
   1. Develop an animation that shows their changes over time and space.
   2. Develop a python code that takes a haplogroup, finds the closest ones in the dataset and plots it on a map (Guyuan Tang).
3. We provide a dataset of ancient Eurasians (see [DataS1](https://github.com/sarabehnamian/Origins-of-Ancient-Eurasian-Genomes/tree/main/steps/Step%200)).
   1. Write software that calculates the minor allele frequency data over time for ancient DNA data. Example for an output (ignore the colors): Chart

      Description automatically generated

(Huy Nguyen)

* 1. Develop an interface that allows selecting populations and time and calculates the minor allele frequencies of the chosen populations. The program will merge patterns that are too close. The results will be printed on the screen as admixture plots (Arash Darzian Rostami).

1. Develop the Geographic Population Structure (GPS) algorithm in python. GPS takes DNA and predicts its geographical coordinates anywhere in the world. [The algorithm](https://github.com/homologus/GPS/blob/master/GPS.r) is written in R. The link also has data (under Step3-GPS). You can encode it in python, expand on it, and produce graphs (Felicia Schulz).
2. Write software that shows the minor allele frequency for a chosen SNP in worldwide populations for a selected period (Saghar Toresson):

Example of output:

Diagram

Description automatically generated

<https://popgen.uchicago.edu/ggv/?data=%221000genomes%22&chr=14&pos=31573859>

You can use the modern DNA dataset (PLINK files for Projects.zip).

1. Develop a tool that traces the allelic changes of an SNP over time and space on a map.
2. Develop a tool that does time forward population simulation (e.g., <https://www.frontiersin.org/articles/10.3389/fgene.2020.601439/full>). Note, that this is NOT a genome simulator, but a simulator of SNP set.
3. Develop a simulator of genomes in populaitons (e.g., <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-223>). Note, that this is NOT a genome simulator, but a simulator of SNP set. Here, it is useful to look at PLINK’s simulator <https://zzz.bwh.harvard.edu/plink/simulate.shtml>
4. Write & code that reads a user’s plink file (.ped format) and a plink file with similar data but for multiple individuals (same .ped format). Report the number of homozygotes and heterozygotes matches of the user with each individual (Joel Moreira).
5. Read ClinVAR markers (markers associated with diseases). Develop an application that allows selecting a ClinVAR marker from the list and plotting its frequency over time in regional populations (S. Europe, S. Asia, etc.) (Dominique Philip Fastus)

You are welcome to come up with ideas outside of this list, as long as they are pre-approved by the teacher, to ensure that their scope fits the timeline and expected complexity of your project. You can use the databases of modern (PLINK files for Projects.zip) and [ancient populations](https://github.com/sarabehnamian/Origins-of-Ancient-Eurasian-Genomes/tree/main/steps/Step%200).

# Product design

* Your code should work and be well-documented.
* Ensure a smooth and presentable design of your product.
* When you are done developing your software, upload it to GitHub.
* Implement your knowledge of version control, documentation, and proper instructions.
* Provide Readme files with instructions on how to execute your code on a dataset or example file that you will also include allowing anyone to REPLICATE your work.

# Oral presentation

**You will be asked to present your work to the class**. You will have 3-4 minutes to present + 2 minutes for questions. Your presentation and claims would be evaluated by a referee (the teacher).

You don’t need to actually run your tool (unless you must, in the cases of web applications). Don’t show your code. Instead, show screens of your tool working and convince us that you built a valuable and useful tool. The order of presentations will be by surnames. If you do not present, you will get a 0 in that component.

# Written report

A 2-4 pages report should be written using the Bioinformatics templates for writing papers (MS Word Template Bioinformatics.dotx). The template is found in the **Projects directory** on the course website. You are advised to find similar application notes or papers to what you will be writing to get an idea of the academic style that you will use (of course, not all the papers are perfect!). A few application notes are provided (Application notes.zip) alongside two documents that have good writing tips (Writing tips - How to write a great science paper.pdf, Writing tips - Ten simple rules for structuring papers.pdf).

# Submission

* Submit your paper in Word format (firstname\_lastname.docx). The paper should contain a link to GitHub.
* Submit your presentation in PowerPoint format (firstname\_lastname.pptx)
* Submit your project files zipped in one file (firstname\_lastname.zip)
* The deadline for the projects and papers: the 21st  of March at 12:00.
* Late submissions would be penalized.

# Project evaluations

The following will be evaluated:

|  |  |
| --- | --- |
| Product design (20%) | Database/web design (if any)  Clarity of instructions and GitHub presentation.  Code documentation  Organization of the GitHub folder.  Existence of input/output files and folders |
| Software (20%) | Software replicability  Accuracy  Level of complexity |
| Oral presentation (30%) | Clarity and understanding of the project. |
| Written report (30%) | Publication quality |

# Writing the paper

|  |
| --- |
| **Follow** [**bioinformatics guidelines**](https://academic.oup.com/bioinformatics/pages/instructions_for_authors) **for authors on how to format your paper.**  **Title Section**  Only include one author, you. You are also the corresponding author. Your  affiliation is *Department of Biology, Box 118, 221 00, Lund University, Sweden*.  **Abstract**  Follow Bioinformatics guidelines for [writing an abstract](https://academic.oup.com/bioinformatics/pages/instructions_for_authors).  **Introduction**  Describe briefly what you have done. Motivate why this is good or interesting. Compare your work with other studies (How is tool work better? In what way does it improve the existing tools?).  **Methods**  Describe what programming languages you used including version number. What data did you use? Where was it obtained or how was it generated? Give references. Include version number in parentheses.  **Features**  Describe what the application does. What input is needed? What output does the user get? Pros and cons of having a web application instead of a standalone program (if you did a web application).  **Discussion**  Briefly describe what you have done and give your application's advantages. What do you foresee when it comes to the user base? What are the limitations and known bugs of the tool? What could be the future directions of developing this technology?  **Acknowledgment**  We are grateful to the editor and two anonymous reviewers for their valuable comments on the manuscript and for Lund university for their financial support.  **References**  Include 1-5 references.  Follow the guidelines of Bioinformatics for writing [references](https://academic.oup.com/bioinformatics/pages/instructions_for_authors). |

Are you still confused as to what to put in each section? Check out this [webpage](https://psychology.ucsd.edu/undergraduate-program/undergraduate-resources/academic-writing-resources/writing-research-papers/research-paper-structure.html#References).

It has links to more resources.

# Other guidelines

There are various papers with useful writing tips in the Project folder. You are advised to read them.

Writing a paper always takes longer than you realize so do not leave it to the last minute.

# FAQ

* Which programming language\technology should I choose?
  + It is up to you! You should pick something that you are familiar with or interested in learning how to use. Please recall that not everyone is familiar with all the languages. Therefore, the sooner you will inform us of your choice, the better it is so that we can assign someone familiar with this language\technology to assist you.
* I need help using this function…
  + Did you google it? One of the goals of this course is to help you learn to find the answers on your own through online searching. Therefore, approach the teacher only after you have exhausted these options since our goal is to prepare you for real life.
* What are my outputs? What should I send you?
  + Read the **Submission** subject above.
* Where can I find relevant data?
  + We made several datasets available to you.
  + There are several ways that papers report data, which also depend on the kind of data.
  + Some deposit their data in NCBI (mainly sequence data and genomes).
  + A variety of biological data can be found in <https://datadryad.org/stash>.
  + Some post their data on GitHub or their personal websites.
* Do I have to publish my code on GitHub?
  + Yes!
* Do I have to use the Bioinformatics template and follow all their other guidelines?
  + Yes!
* What should I put in GitHub?
  + Your GitHub depository should include the following:
    - A readme file that explains about your project, methodology, files included, how to run your code, and unknown bugs. You should explain how to install dependencies and answer common questions.
    - Your code should be well documented.
    - If there are multiple versions, show it and explain the difference between the versions.
    - Provide several (typically, 1-3) input and output files that demonstrate that your code works in separate folders.
    - Don’t put all the files in one folder, be organized.
    - Assume that the user who wants to use your code is ignorant.
  + You can also have a folder with previous versions. This demonstrates that you did version control.
* Should I do “live demonstrations” during my presentation or just show slides?
  + It’s up to you how to allocate the time for your talk. Screens capture of your tool is acceptable.
* Can I use an existing toolbox/package for my project?
  + No. You have to build something of your own. You can use existing tools if you need their functionality, but you would be evaluated based on what you built. In other words, **your tool should be publishable**, and only new tools that promote the field are considered publishable. Learning how to use existing tools does not satisfy this requirement.
* Should I use a citation manager?
  + **It is strongly recommended**. You are far more likely to make mistakes if you do not use a citation manager. If you want to stay in Academia, you have to learn how to use one.
* Which citation manager should I use?
  + There are many citation managers like EndNotes (license required), Mendeley (free), or other ones. They are compared here: <https://en.wikipedia.org/wiki/Comparison_of_reference_management_software>. Most of the students like Mendeley or RefWorks.
* I don’t know how to write academic papers.
  + That’s OK. If you never wrote an academic paper, there is no reason that you should know how to write one. Writing is something that we learn by doing and through examples. You were provided with examples of papers. Many more examples are available online. In addition, you were provided documents with helpful writing tips; use them. Finally, when you are done, proofread your paper. Word can catch many typos. Grammarly is also useful, even if you use the free version. Finally, exchange your draft with your friend and proof each other papers. This is how we write papers.
* Can I cite Wikipedia? What types of references are appropriate?
  + No. Read this [website](https://psychology.ucsd.edu/undergraduate-program/undergraduate-resources/academic-writing-resources/writing-research-papers/appropriate-references.html) and this [one](https://psychology.ucsd.edu/undergraduate-program/undergraduate-resources/academic-writing-resources/writing-research-papers/citing-references.html) concerning citation format and rules.
* Can I copy code from other tools?
  + No. That is considered plagiarism. You can get ideas from other codes. If you reuse parts of other codes, you have to acknowledge that in your paper, and it cannot be the main function of your code (otherwise, what is the value of your code?).
* Should I include example data in my paper?
  + Yes! Show the output of your tool when applied to real biological data in a graph or table. Remember that it is your job to convince others to use your software. So you must provide evidence that your tool works.
* Should I provide figure and table legends?
  + Yes! This is where you explain what the figure/table show and say. Just writing the title is not enough.
* How can I get a good grade on this assignment?
  + Read this document. Prepare a checklist and go over it before you submit it. Show your paper to your peers and proofread each other papers. If something is unclear, ask the teacher.

**Good luck!**