Genomic projects

Stefano Scansani

March 1, 2024

Abstract

So far this is the LaTeX versin of the README.md.

Contents

List of Figures

List of Tables

1 Genomic projects tutorials

!!! warning "This repository is under construction"

This repository contains a collection of genomic projects that I am working on. GitHub repository of bioinformatic projects recolving around genomics using different tools like Plink through `plinkr` R package, `rTASSEL` and TASSEL 5 (GUI), GEMMA for mixed models analysis in R, SAMtools from the command line to analyze BAM files, gBLUP coming soon.

The repository has been created for self-teaching purposes of biological concept and bioinformatic tools, and make use of other repositories, scripts and data sources, taken or modified as such.

1. SNP profiling of goat breeds.<br>*Data source*: **Colli et al.** (2018) https://doi.org/10.1186/s12711-018-0422-x

<img src="Figures/goat*mds*2.png" height="350px">
<small>**Multidimensional Scaling (MDS) Plot of a population of 4,653 Individuals from 169 Goat Breeds genotyped with 49,953 SNPs.**</small>
<small>The MDS plot visualizes genetic relationships among 4,653 individuals from 169 goat breeds. Genetic distances were computed using PLINK to generate the kinship matrix, and MDS analysis was conducted with the `cmdscale` function based on genotyping data from 49,953 SNPs. Each point represents a goat, and spatial arrangement reflects genetic dissimilarities. This exploratory analysis offers insights into genetic diversity, population structure, and relatedness. </small>

1. a. Manhattan plot of a GWAS on dog population for deafness.*Data source*: **Hayward et al.** (2020) https://doi.org/10.1371/journal.pone.0232900

<img src="Figures/manhattan*dogs.png" height="350px"> <img src="Figures/manhattan*filter_acd.png" height="350px"> <small>Manhattan plots showing the genome wide association (GWA) between dog deafness and their genotype. The plot displays the genomic positions of single nucleotide polymorphisms (SNPs) across the genome on the x-axis, with the corresponding -log<sub>10</sub> transformed *P*-values indicating the strength of association with the trait on the y-axis. </small>

2. b. Plot of the top significant SNPs identified in the above GWAS.

<img src="Figures/top*SNPs*comb.png" height="350px">
and a zoom in the chromosome 3 above LOD 3.5. <img src="Figures/SNP*chr3*top.png" height="350px">

1.1 Tools

- **PLINK 1.90** `https://www.cog-genomics.org/plink2/`
- `plinkr` R package repository documentation. `https://github.com/AJResearchGroup/plinkr`
- **TASSEL 5** `https://www.maizegenetics.net/tassel`. **Bradbury** et al., (2007) TASSEL: software for association mapping of complex traits in diverse samples, Bioinformatics, Volume 23, Issue 19, Pages 26332635 `https://doi.org/10.1093/bioinformatics/btm308`
- `rTASSEL` R package repository documentation. <br> Vignettes: `https://rtassel.maizegenetics.net/index.html`, Repository: `https://github.com/maize-genetics/rTASSEL`. **Monier et al.**, (2022). rTASSEL: An R interface to TASSEL for analyzing genomic diversity. *Journal of Open Source Software*, 7(76), 4530, `https://doi.org/10.21105/joss.04530`

- GEMMA Genome-wide Efficient Mixed Model Association https://github.com/genetics-statistics/GEMMA. **Xiang Zhou and Matthew Stephens** (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44, 821824.

1.2 Resources & Data

- **Marees et al.** (2018) A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res*. 27:e1608. https://doi.org/10.1002/mpr.1608

- **Marees et al.** (2018) tutorial https://github.com/MareesAT/GWA_tutorial

- **Gábor Mészáros** (2021) Genomic Boot Camp Book https://genomicsbootcamp.github.io/book/

- **Gábor Mészáros** video tutorials https://www.youtube.com/c/GenomicsBootCamp

- **Colli et al.** (2018) Genome-wide SNP profiling of worldwide goat populations reveals strong partitioning of diversity and highlights post-domestication migration routes. *Genet Sel Evol* 50, 58. https://doi.org/10.1186/s12711-018-0422-x

- DATA: **Colli et al.** (2020). Signatures of selection and environmental adaptation across the goat genome post-domestication [Dataset]. *Dryad*. https://doi.org/10.5061/dryad.v8g21pt

- **Decker et al.** (2014) Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *PLOS Genetics* 10(3): e1004254.https://doi.org/10.1371/journal.pgen.1004254[1],

- DATA: **Decker et al.** (2015) Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle [Dataset]. Dryad. https://doi.org/10.5061/dryad.th092

1.3 Setup of the working environment

Install R: The Comprehensive R Archive Network (CRAN)[2]
IDE:VSCode[3]/RStudio[4]<sup>*</sup>
Install Python: Miniconda 3[5]<sup>*</sup>
OS: Linux<sup>*</sup>/WSL
<small><sup>*</sup>Suggested</small>

[1] https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004254

[2] https://cran.r-project.org/

[3] https://code.visualstudio.com/

[4] https://posit.co/download/

[5] https://docs.anaconda.com/free/miniconda/index.html

1.3.1 Get PLINK working in Linux

1. Download PLINK 1.90 Linux 64-bit[6]
2. Install PLINK `cd Downloads/ sudo unzip plink_linux_x86_64_20200616.zip -d plink_install`
3. PLINK in `usr/local/bin`

   ```
   cd plink_install sudo cp plink /usr/local/bin sudo chmod 755
   /usr/local/bin/plink
   ```

4. Add PLINK to PATH

   with bash/zsh/...

   ```
   sudo nano ~/.bashrc
   ```

   adn include the line:

   ```
   export PATH=/usr/local/bin:$PATH
   ```

   Save and exit. Refresh the terminal and you should be able to call `plink` from the terminal at any user position in the system.

   ```
   source ~/.bashrc plink -help
   ```

1.3.2 Get TASSEL (GUI) on Linux

1. Go on the website `https://www.maizegenetics.net/tassel` and download the last UNIX verison.
2. Download the TASSEL_{xxx}_unix.sh and make it executable `chmod +x ~/Downloads/TASSEL_{xxx}_unix.sh`
3. Run the TASSEL installer `~/Downloads/TASSEL_{xxx}_unix.sh`

1.3.3 Get `rTASSEL` working in Linux

1. `rJava` installation

   ```
   sudo apt install default-jdk sudo R CMD javareconf R install.packages("rJava")
   ```

2. Installation in R

```
    if (!require("devtools")) install.packages("devtools") devtools::install_github(
repo = "maize-genetics/rTASSEL", ref = "master", build_vignettes =
TRUE, dependencies = TRUE )
```

3. Run `rTASSEL`

   - Allocate job's memory<sup>1</sup> and start the logger (here at the root of the project):

<sup>1</sup>"-Xmx50g" and "-Xms50g", "*50g*" represents 50 Gigabytes of memory.

!! Choose an appropriate value that fits your machine !!

```
    options(java.parameters = c("-Xmx50g", "-Xms50g")) rTASSEL::startLogger(fullPath
= NULL, fileName = NULL)
```

[6] `https://s3.amazonaws.com/plink1-assets/plink_linux_x86_64_20231211.zip`

- Run & infos

```
library(rTASSEL) ??rTASSEL
```

Useful resource for `rTASSEL` are the vignettes and tutorials at `https://rtassel.maizegenetics.net/index.html`

1.3.4 Get GEMMA

GEMMA can be installed from source at the GitHub repo, but is also available through Bioconda `http://www.ddocent.com/bioconda/`. To install is suggested to have miniconda installed and working, and then added the channel for Bioconda, you should already have defaults and conda-forge.

```
conda config -add channels defaults conda config -add channels
conda-forge conda config -add channels biocond conda install gemma
```

And use GEMMA with

```
gemma -h
```

---

2 Project: SNP Profiling of worldwide goat populations

2.1 Data

Data source: Colli et al. (2018) https://doi.org/10.1186/s12711-018-0422-x

3 Methods

3.0.1 Data import

4 Results

4.0.1 Multidimensional scaling

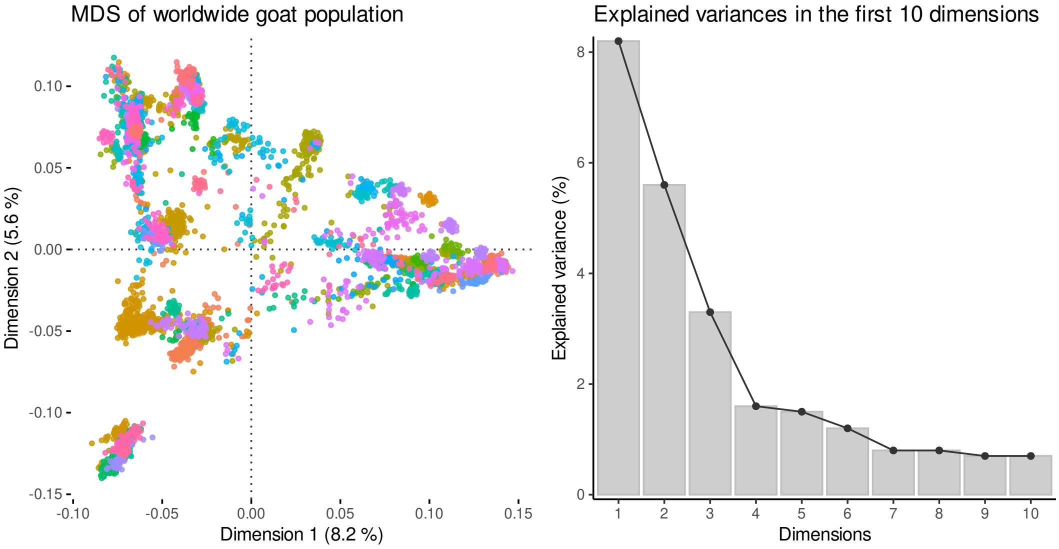

Figure 1: Multidimensional Scaling of the goat dataset

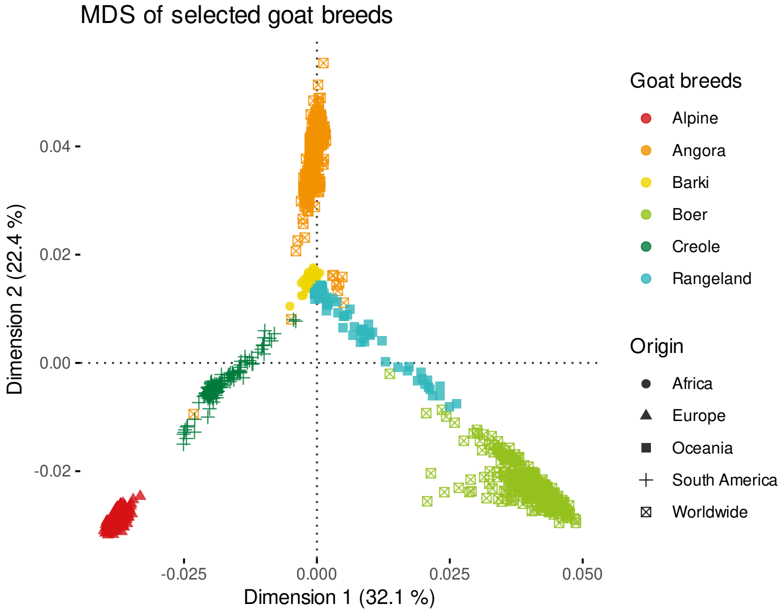

Figure 2: Multidimensional Scaling of the goat dataset, subset of 6 goat breeds.