genomic-projects-tutorials

1 Genomic projects tutorials

:warning: This repository is under construction :warning:

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 to analyze BAM files, and other coming soon!

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

The report of the studies is in progress in Report/build/Genomics_proj.pdf

contents

1.1 Contents

tools

1.2 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 2633–2635 `https://doi.org/10.1093/bioinformatics/btm308`

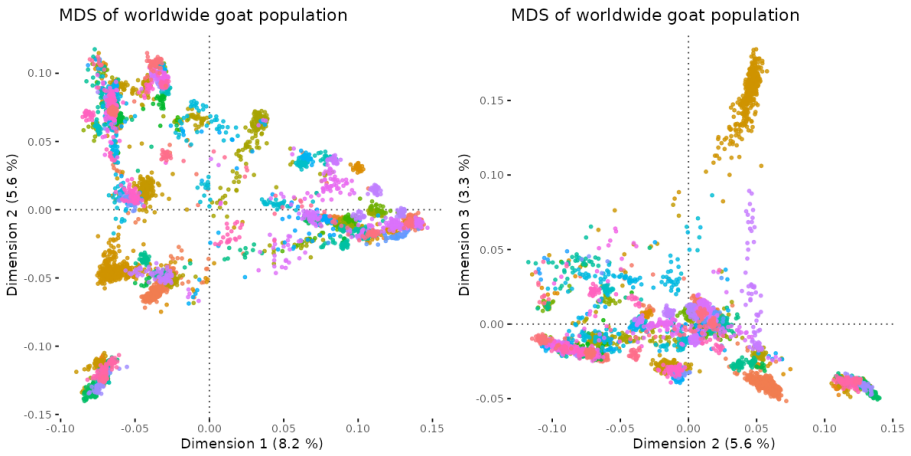

Figure 1: Multidimensional scaling of the genotypes

- `rTASSEL` R package repository documentation. 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, 821–824.
- `rMVP` A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-Wide Association Study https://github.com/xiaolei-lab/rMVP
- `GPtour` Genomic Prediction in R using Keras models https://github.com/miguelperezenciso/GPtour and https://keras.posit.co/articles/getting_started.html
- `GAPIT` Genome Association and Integrated Tools https://github.com/jiabowang/GAPIT

example-case-studies

1.3 Example case studies

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

Multidimensional Scaling (MDS) Plot of a population of 4,653 Individuals from 169 Goat Breeds genotyped with 49,953 SNPs.

The MDS plot visualizes the genetic relationships among 4,653 individuals from 169 goat breeds. Genetic distances were computed using PLINK to generate the distance matrix, and MDS analysis was conducted with the `cmdscale` function based on genotyping data from

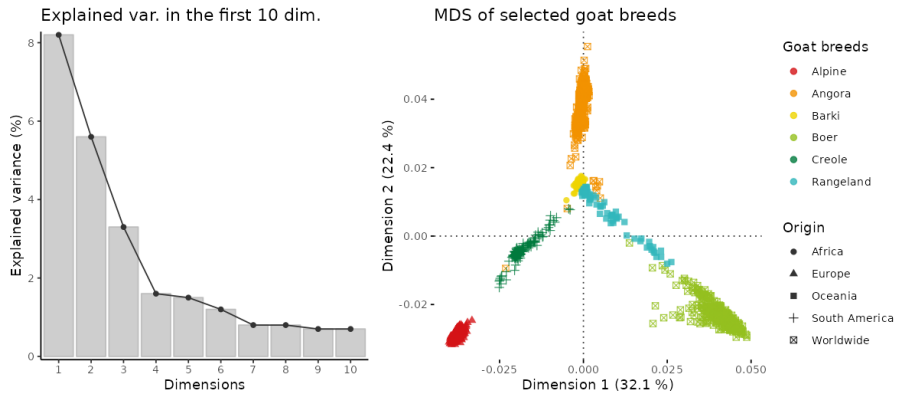

Figure 2: Scree plot of all genotypes and multidimensional scaling of a subset of genotypes

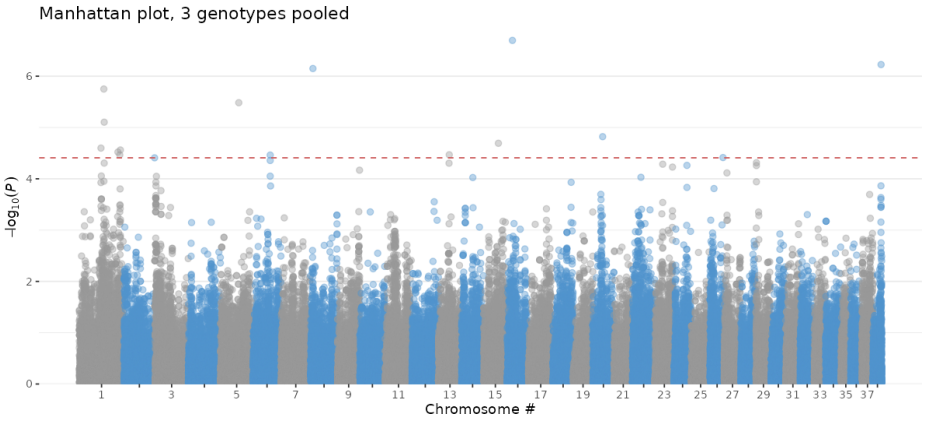

Figure 3: Manhattan plot

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.

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

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 -$\log_{10}$ transformed P-values indicating the strength of association with the trait on the y-axis. The red-dashed lines are representation of the 99.99 percentile threshold of the LOD values.

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

Points are jittered around their respective chromosome.

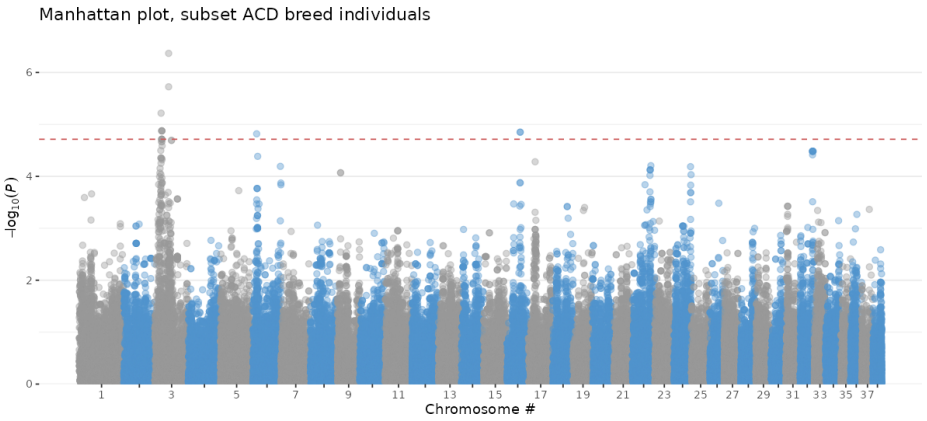

Figure 4: Manhattan plot of a single canine breed

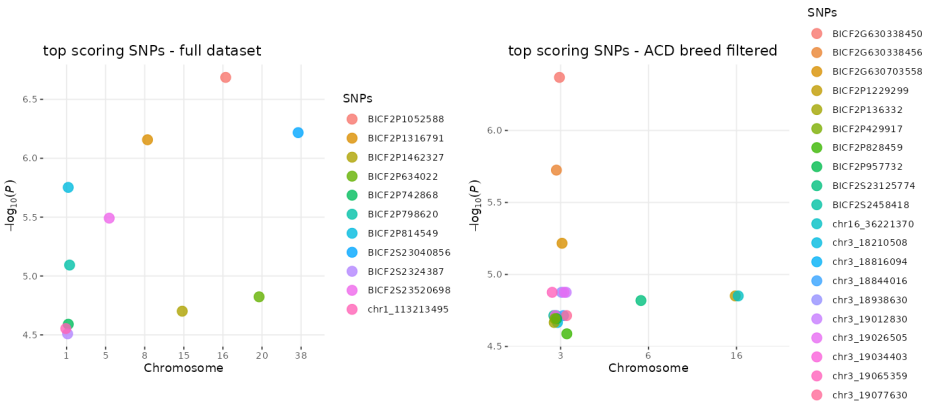

Figure 5: Top scoring SNPs

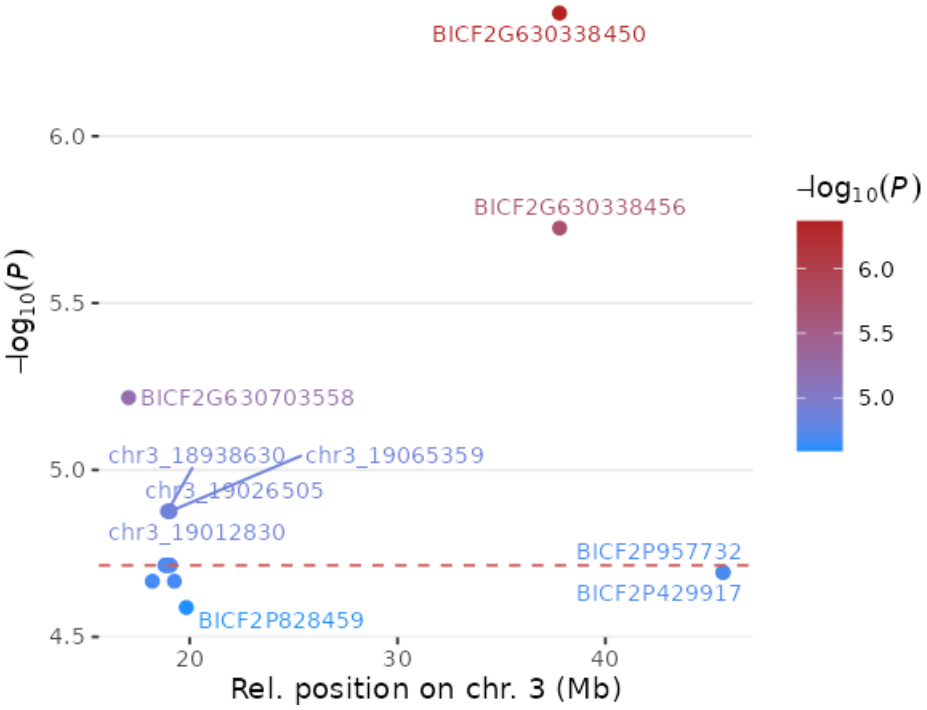

Figure 6: Top scoring SNPs of a ABC breed in the 3rd chromosome

and a zoom in the chromosome 3 above the 99.99 percentile (LOD score = 4.71).

resources-data

1.4 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

. *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://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004254https://doi.org/10.1371/journal.pgen.1004254,
- DATA: **Decker et al.** (2015) Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle

$$Dataset$$

. Dryad. https://doi.org/10.5061/dryad.th092

setup-of-the-working-environment

1.5 Setup of the working environment

Install R: https://cran.r-project.org/The Comprehensive R Archive Network (CRAN)
IDE:https://code.visualstudio.com/VSCode*/https://posit.co/download/RStudio*
Install Python: https://docs.anaconda.com/free/miniconda/index.htmlMiniconda 3*
OS: Linux*/WSL
*Suggested
get-plink-working-in-linux

1.5.1 Get PLINK working in LinuxGet PLINK working in Linux

1. Download https://s3.amazonaws.com/plink1-assets/plink$_{linux}{}_{x}86_64_20231211.zipPLINK1.90Linux64-$
2. PLINK in usr/local/bin

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

3. Add PLINK to PATH
   with bash/zsh/...

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

   and 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
```

get-plinkr-r

PLINK directly in r.

refer to the installation guide at https://github.com/AJResearchGroup/plinkr/blob/master/doc/install.md

```
library(remotes)
install_github("richelbilderbeek/plinkr")
remotes::install_github("chrchang/plink-ng/2.0/pgenlibr")
library(plinkr)
install_plinks()
```

get-tassel-gui-on-linux

1.5.2 Get TASSEL (GUI) on LinuxGet 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
   ```

get-rtassel-r

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[1] and start the logger (here at the root of the project):

   [1]"-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)
```

   - Run & infos

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

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

get-gemma

1.5.3 Get `GEMMA`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
```

get-gapit-r

1.5.4 Get GAPIT (R)Get GAPIT (R)

R package, here we are going to install it through GitHub. For the manual visit https://zzlab.net/GAPIT/gapit_help_document.pdf

```
R> install.packages("devtools")
R> devtools::install_github("jiabowang/GAPIT", force=TRUE)
R> library(GAPIT)
```