An analysis of variations in drought stress-related genes of *Arabidopsis thaliana* - a comparison of populations from Spain and Sweden

Halide Aydin[1], Sezin Dogan[1], Aliz Fodor[1], Ingo Giebel[1], Janine Graser[1], and Paula Winnitzki[1]

QBio305: Population and Quantitative Genetics
[1]Heinrich-Heine-Universität Düsseldorf
[2]Institute for Plant Sciences, Universität zu Köln
[2]Prof. Dr. Juliette de Meaux
[2]Dr. Tahir Ali
[2]Dr. Markus Stetter

February 5, 2024

Contents

List of tables

List of figures

1 Abstract

TODO...

2 Introduction

TODO...

A. thaliana is a nonselective model organism that is known to accommodate different climates easily. In this study we want to find out if samples of *A. thaliana* in Spain, that have been under drought conditions, show different genetic diversity than of the samples in Sweden.

It is important to research plants' drought resistance since it is a pressuring issue when it comes to agriculture. Water limitation is an ongoing challenge in agriculture and due to climate change it is significant to see how the plants adjust to the harsh climates and to investigate the effects of changing climate on these plants.

The findings of this study can shed light to which mechanisms and genes help plants against harsh conditions and also later help us in the research of other model organisms and plants.The findings can also contribute to understanding and improving plants against harsh conditions in agriculture via genetic engineering. We wanted to see if the drought would affect the genetic diversity of the plant.

Tahir:

I have reviewed your research project idea and find it to be a very relevant topic for exploration. This investigation can help you to understand how *A. thaliana* adapts to drought stress and the resulting impact on its genetic diversity. While comparing the genetic diversity of *A. thaliana* samples from Spain and Sweden, where plants face varying drought conditions, is a good starting point, consider refining your research question. Focus on drought stress-related genes that may show variation and adaptation in *A. thaliana* populations across different regions. Utilize measures like Tajima's D or Fst to assess genetic diversity and adaptation, observing variations between populations from the two regions. Alternatively, you may choose to conduct a Genome-Wide Association Study (GWAS) to pinpoint SNPs and genes associated with drought tolerance or resistance.

3 Materials and methods

TODO...

3.1 Data source and collection

- Whole FASTA reference genomic data of A. thaliana
- Whole genomic sequencing data of A. thaliana from regions in Spain and Sweden

3.2 Data preprocessing

The raw FASTQ sequencing data was step wise preprocessed as follows:

1. Quality control of the raw FASTQ data.
2. Quality trimming of the raw FASTQ data.
3. Quality control of the trimmed FASTQ data.
4. Sequence aligning of the genomic sequencing data (FASTQ files) with the reference genomic data of A. thaliana (FASTA file). The resulting SAM files were converted to sorted BAM files.
5. The sorted BAM files were quality filtered, PCR duplicates were removed, and only properly aligned paired-end reads were kept: mapping quality -q 30, secondary alignments and reads failing the quality checks were removed.
6. the sorted and filtered BAM files were converted to a VCF file. Minimum calling threshold for variant alleles: -p 0.01. Therefore, only variants with an allele frequency of at least one percent were called.
7. Further quality filtering: All monomorphic variant sites and all indels were removed.

TODO: Figure illustrating the preprocessing pipeline.

The geospatial data of the accessions were determined using the 1001 KML data (TODO: link to the XML file). For the further downstream analysis, the following geospatial data of the accessions was acquired using the NASA Earth Observatory database:

- Land surface temperature
- Total rainfall
- Snow cover

3.3 Selection of genomic regions

3.4 Bioinformatics and statistical analysis

3.5 Software and tools

The following software and tools were used for data processing and analysis:

- FastQC v0.12.1 - quality control of the raw FASTQ data
- fastp v0.23.4 - quality trimming of the raw FASTQ data
- bwa-mem2 v2.2.1 - indexing the FASTA reference file + mapping of the trimmed FASTQ files to the reference FASTA file
- samtools v1.18 - conversion of the SAM files into BAM format + quality filtering + removing PCR duplicates + indexing and sorting the BAM files
- QualiMap v.2.3 - evaluation of the mapping quality
- bcftools v1.18 - variant calling and manipulating files in the Variant Call Format (VCF) and its binary counterpart BCF
- vcftools v0.1.17 - additional filtering of the VCF files
- PLINK v1.90b6.21 - whole-genome association and population-based linkage analyses
- ADMIXTURE v1.3.0 - ancestry estimation in a model-based manner from large autosomal SNP genotype datasets
- Stacks v2.65 - modular pipeline to perform several different types of analyses
- R programming language v4.3.2 - statistical analysis using a self developed script (see appendix ...)
- RStudio v2023.09.1+494 - integrated development environment (IDE) for R

For the statistical analysis and plotting with R, the following packages were used:

- adegenet v2.1.10 - for multivariate analysis of genetic markers
- BiocManager v1.30.22 - to access the Bioconductor project package repository
- RColorBrewer v1.1-3 - color palettes for creating graphics
- factoextra v1.0.7 - to extract and visualize the results of multivariate data analyses
- FactoMineR v2.9 - for multivariate exploratory data analysis and data mining
- ggrepl v0.9.4 - to automatically position non-overlapping text labels
- gplots v3.1.3 - for plotting
- here v1.0.1 - to determine the project root (required for accessing files using relative path names)
- LEA v3.14.0 - for landscape and ecological association studies
- plotly v4.10.3 - to create interactive plots
- popReconstruct v1.0-6 - for reconstructing populations of the recent past

- StAMPP - v1.6.3 - for statistical analysis of mixed ploidy populations
- tidyverse v.2.0.0 - data transformation and presentation
- vcfR v1.15.0 - manipulation and visualization of variant call format (VCF) data

4 Results

TODO...

5 Discussion

TODO...

A Appendix A

The first appendix...

B Appendix B

The second appendix...

C Appendix C

The third appendix...