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...

Tahir:

Having finalized your research question and study regions, you can now focus on the Materials and Methods section, which involves identifying the types of datasets needed and exploring suitable statistical methods and tools for data analysis. I will help you download the required genome-wide SNP data for accessions from Spain and Sweden as a VCF file from https://1001genomes.org/data/GMI-MPI/releases/v3.1/.

If you need climate data, it can be downloaded from https://www.world-clim.org/data/worldclim21.html. Don't forget to identify any additional datasets needed and their potential sources before starting the analysis. We encourage you to contact us with any additional questions, and we are here to guide you throughout the process of designing and executing your research project.

3.1   Data

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

3.2   Preprocessing of the raw data and mapping to a reference genome

- FastQC v0.12.1 - quality control of the raw FASTQ data
- fastp version 0.23.4 - quality trimming of the raw FASTQ data
- bwa-mem2 version 2.2.1 - indexing the FASTA reference file + mapping of the trimmed FASTQ files to the reference FASTA file
- samtools version 1.18 - 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

3.3   Evaluation of the variants calling and filtering

- bcftools version 1.18 - filtering for minimum base quality (-q 30) + minimum calling threshold for variant alleles (-p 0.01) + removing all monomorphic variant sites + some additional filtering

3.4   Statistical analysis tools and methods

3.4.1   Tools

- R version 4.3.2 - statistical programming language
- RStudio version 2023.09.1 build 494 - integrated development environment (IDE) for R

3.4.2 Methods

TODO...

4 Results

TODO...

5 Discussion

TODO...

A   Appendix A

The first appendix...

B   Appendix B

The second appendix...

C   Appendix C

The third appendix...