Metabolomic Analysis of Mutagenised Lines in Sorghum

Cecilie Svendsen

This thesis is submitted for the degree of
Master of Science in Bioinformatics

Section for Bioinformatics and Computational Research (BiRC)
Department of Molecular Biology and Genetics
Aarhus University

In collaboration with
Center for Quantitative Genetics and Genomics

Internal supervisor: Thomas Bataillon
External supervisor: Guillaume Ramstein

June 2026

Acknowledgements

First, I would like to express my gratitude to my external supervisor, Guillaume Ramstein, for giving me the opportunity to be part of this project and for his continuous guidance throughout the course of this thesis.
I am also very grateful to my interval supervisor, Thomas Bataillon, for his guidance and valuable feedback. The code, literature, insights, and discussions provided by both supervisors have been very helpful throughout this work.
I would also like to acknowledge everyone who has directly or indirectly contributed to the completion of this thesis. In particular, I would like to thank Nikola Micic for providing the datasets and useful information about them.
Finally, I would like to thank my family and friends for their encouragement and support throughout my studies.

Abstract

The changing climate poses significant challenges for agriculture, including increased frequency of extreme weather events that can negatively affect crop yield. Improving crop yield is therefore an important area of research. In recent years, precision breeding and targeted variant selection have been increasingly used to accelerate the development of climate-resilient crops.

This thesis investigates the ability of protein language models to predict genetic variants that may improve plant fitness. To evaluate this, sorghum plants were mutagenised and screened for candidate variants. The aim of this study was to analyze metabolomic and physiological data from mutant and wild-type sorghum plants in order to assess whether the targeted variants resulted in significant differences between genotypes, and whether mutant plants exhibited more beneficial traits.

Statistical analyses and modelling were performed on provided datasets of metabolic features and physiological traits. The results showed few consistent differences between genotypes across most targets. However, differences were observed in some metabolic features and metabolite groups, including sesquiterpenoids and napthalenes. In terms of physiological traits, some differences were identified in certain mutant lines, including increased anthocyanin levels, reduced chlorophyll concentration, and reduced plant height.

Overall, the results suggest that the introduced mutations did not consistently improve plant fitness. However, the observed differences should be interpreted with caution due to uncertainty in the genotype-phenotype relationships, generally weak genotype effects that explained little variation in metabolic features, and large confidence intervals relative to effect sizes for physiological traits, while environmental and systematic effects may also have contributed to the observed differences.

Abbreviations

ABA: Abscisic acid
AIC: Akaike information criterion
DMS: Deep mutational scanning
EIC: Extracted ion chromatogram
EMS: Ethyl methanesulphonate
ESM: Evolutionary scale modeling
FDR: False discovery rate
FIND-IT: Fast Identification of Nucleotide variants by droplet DigITal PCR
F_v/F_m: The maximum quantum yield of photosystem II (PSII)
GERP: Genomic evolutionary rate profiling
g_S: Stomatal conductance
LC: Liquid chromatography
MAPK: Mitogen-activated protein kinase
MS: Mass spectrometry
MSA: Multiple sequence alignment
NADP-MDH: NADP-dependent malate dehydrogenase
NBI: Nitrogen balance index
NDVI: Normalised difference vegetation index
PC: Principal component
PCA: Principal component analysis
PLM: Protein language model
PRI: Photochemical reflectance index
Rep: Replicate
ROS: Reactive oxygen species
SAM: S-adenosylmethionine
SE: Standard error
SIFT: Sorting intolerant from tolerant
TPM: Transcripts per million
VSN: Variance stabilization normalization
WT: Wild-type
ΦPSII: Effective quantum yield of photosystem II

Table of contents