MASTER THESIS IN BIOINFORMATICS

Simulating generations of the Danish population to evaluate BLUPF90 variance component estimation

by
Giovanni Vincenzo Cavallo
202402143

Supervisor:
Jakob Grove

Co-supervisor:
Juan Cordero

AARHUS UNIVERSITET

Bioinformatics Research Center (BiRC)
Faculty of Natural Sciences
Aarhus University
Denmark

15th June, 2026

Abstract

Several epidemiological studies have reported an association between maternal cardiometabolic conditions and increased offspring autism risk, an association that can arise through multiple mechanisms simultaneously, including direct pregnancy effects, shared familial environment, and shared genetic liability.
Disentangling these mechanisms in large population registers requires statistical models capable of explicitly partitioning maternal and offspring genetic components, validated under controlled conditions before application to real data.

This thesis develops a generalizable simulation pipeline, built on the AlphaSimR R package, for generating multi-generation pedigrees resembling human population registers across stratified generations.
Although motivated by the Danish population and the iPSYCH cohort, the workflow is designed to be adaptable to different populations, demographic structures, and trait pairs with minimal modification.
Realistic founder haplotypes are generated using a coalescent Markov chain simulation approach and expanded to a founder pool of 850,000 individuals through random crossing.
From the resulting population of approximately 6 million individuals, an iPSYCH-like genotyped cohort of 150,000 individuals is drawn from the two most recent generations, including 25,000 autism cases selected through a relatedness-aware greedy sampling algorithm.

Joint maternal and offspring phenotypes are simulated under a liability threshold model, partitioning the offspring liability into direct additive genetic effects, maternal genetic effects, shared maternal environment, and the effect of a maternal condition during pregnancy on offspring risk.
This framework has been built with CMC and autism in mind, but can be generalized and applied to any joint mother-child trait pair.

The BLUPF90 software suite is evaluated on the simulated data using four model configurations, from a simple pedigree-based model to a full bivariate single-step GBLUP model.
The results show that BLUPF90 can correctly estimate maternal genetic effects and direct offspring heritability when the model is correctly specified, while also highlighting the challenges of estimating permanent environment effects and genetic correlations in human-like pedigree structures.

Acknowledgements

I would like to start by thanking my supervisor for his kindness, for giving me this opportunity, and for the trust he placed in me throughout this project.
I could not have asked for a better person to guide my thesis.

A sincere thank you goes also to my co-supervisor for the continuous support and for always being there while still leaving room for my personal growth and vision.
That balance made a real difference.

I want to thank everyone at the Bartholin Building for the warm welcome, and GenomeDK for providing the computational infrastructure that made this work possible.

To the teachers of the Master's programme in Bioinformatics: thank you for always being available and for creating an environment where curiosity and growth were never taken for granted.

To my Master's colleagues: thank you for always being there, for reminding me I was never alone through the highs and lows of this journey, and for being simply the best colleagues I could have hoped for.

To my family, my partner, and my friends back in Italy: thank you for the support, the patience, and for being a constant even from a distance.
To my dormmates: thank you for making Aarhus feel like home and for the smiles on the harder days.

Finally, a thank you to everyone who crossed this path at some point and left something behind, even briefly.

Contents