

## Dissecting the Multi-Phenotype Effects for Cardiometabolic Traits in Highly Dimensional Whole Genome and Omics Data through Usage of Multivariate Analytical Methods

Mila Desi Anasanti

Prof Philippe Froguel | Prof Inga Prokopenko | Dr Marika Kaakinen

This project report is submitted in partial fulfilment of the requirements for the PhD in Clinical Medicine award, Imperial College London. I confirm that it is an original work representing my own academic effort and that all sources have been fully acknowledged.

## Abstract

Over a decade, single-phenotype genome-wide association studies (SP-GWAS) have been used to identify the association between variants and cardiometabolic traits. Initially, our team performed an SP-GWAS meta-analysis of fasting insulin (FI) and fasting glucose (FG) of European and trans-ethnic ancestries within the Meta-Analysis of Glucose and Insulin-related traits Consortium (MAGIC). However, after calculating the variance explained, I found FG only slightly increased from MAGIC's previous analysis from 1.5% to 4.3%.

We proposed multi-phenotype GWAS (MP-GWAS) to boost the statistical power and performed MP-GWAS of fatty acids in NFBC1966 (N=4955) and replication in NFBC1986 (N=2687) to investigate fatty-acid metabolisms. The meta-analysis conducted by our team detected 10 signals associated with FAs ( $P < 5 \times 10^{-8}$ ) at *PCSK9*, *GCKR*, *LPXN*, *FADS1*, *GPR137*, *ZNF259*, *LIPC*, *PDXDC1*, *PBX4*, and *APOE*. For subsequent analysis, I proposed a new direct conditional analysis method within MP-GWAS, which detected multiple distinct signals within these loci.

While MP-GWAS is a powerful method for locus discovery, it could increase missing phenotype data drastically. I further investigated the properties of seven imputation methods within the MP-GWAS framework via an extensive simulation study.

I found that random forest (RF) is the best under various scenarios. However, as there was no available RF software designed for high-dimensional data, I developed the fastest to date RF imputation software, imputeSCOPA.

I applied imputeSCOPA to the NFBC data and performed an MP-GWAS of 31 metabolites on the imputed data and complete-cases (CC). I found that the analysis

using imputed data boosted the power of MP-GWAS's by identifying two novel signals at rs61803025 within *FCGR3B* ( $P_{CC}=5.68 \times 10^{-7}$  vs  $P_{imp}=5.49 \times 10^{-9}$ ) and rs181847072 within *ADAMTS3* ( $P_{CC}= 5.67 \times 10^{-7}$  vs  $P_{imp}= 9.27 \times 10^{-11}$ ).

These results demonstrate the increased power from MP-GWAS as compared to the traditional SP-GWAS. This work further highlights the importance of addressing missing data correctly and introduces a fast RF-based software imputeSCOPA.

Abstract word count: 300

Report word count: 35,562

## Declaration of Originality

I declare that this thesis for submitting a Doctor of Philosophy is based on my own work. Proper recognition is given to the part of work performed in collaboration with other colleagues.

**Mila Desi Anasanti**

## Copyright Declaration

This thesis's copyright rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives license. Unless otherwise indicated, its contents are licensed under an Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, do not use it for commercial purposes, and do not alter, transform, or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work.

**Mila Desi Anasanti**

## Acknowledgements

Foremost, praises to Allaah Almighty for His showers of blessing, for giving me the knowledge and ability to undertake and complete this thesis.

My deep and sincere gratitude to my research supervisors Prof Inga Prokopenko, Dr Marika Kaakinen, and Prof Philippe Froguel, for providing me invaluable support, guidance, and immense knowledge. Inga, thanks for your trust in my capabilities to conduct this research, which was a new field for me. Marika, I am deeply indebted to your sincere support and advice at all times; you have given me invaluable guidance, inspiration, and suggestions in my quest for knowledge. Many thanks for being there at times when I required motivation during my challenging situation.

I humbly extend my thanks to all my colleagues and fellow research scholars at the Prokopenko group, Dr Elena Loizidou, Dr Zhanna Balkiyarova, Dr Harmen Draisma, and Igor Pupko. Their support, encouragement, insightful feedback, and credible ideas have widened my research from various perspectives.

I would like to thank all the scientists involved in the preparation of the Northern Finland Birth Cohort (NFBC) samples. Dr Momoko Horikoshi genotyped NFBC1966 samples, and Dr Paul O'Reilly and Clive Hoggart calculated its PCs. NFBC1986 was genotyped by Dr Afshan Siddiq and Hutokshi Crouch, while Dr Momoko Horikoshi calculated its MDS.

Data collection, management, and analyses have been supported by the grants from the Academy of Finland (Grant no. 1114194, Grant no. 12926901), University Hospital

Oulu and Unit of Primary Care, Biocenter, University of Oulu, and the European Commission (EURO-BLCS, Framework 5 award QLG1-CT-2000-01643 and EU FP7-EurHEALTHAgeing - European Research on Developmental, Birth and Genetic Determinants of Ageing. Grant no. 24000522).

Finally, I would like to thank all the participants who participated in the NFBC1966 and NFBC1986 studies at the University of Oulu.

# Table of Contents

|                                                                                                  |           |
|--------------------------------------------------------------------------------------------------|-----------|
| <b>Abstract .....</b>                                                                            | <b>2</b>  |
| <b>Declaration of Originality.....</b>                                                           | <b>4</b>  |
| <b>Copyright Declaration.....</b>                                                                | <b>5</b>  |
| <b>Acknowledgements .....</b>                                                                    | <b>6</b>  |
| <b>List of Tables .....</b>                                                                      | <b>12</b> |
| <b>List of Figures .....</b>                                                                     | <b>13</b> |
| <b>List of Supplementary Tables .....</b>                                                        | <b>16</b> |
| <b>Abbreviations.....</b>                                                                        | <b>17</b> |
| <b>The overall aim, hypotheses, research questions, scope, and structure of the thesis .....</b> | <b>19</b> |
| <b>The overall aim of the thesis .....</b>                                                       | <b>19</b> |
| <b>Study-specific hypotheses .....</b>                                                           | <b>21</b> |
| <b>Research questions for the thesis .....</b>                                                   | <b>22</b> |
| <b>Scope of the thesis.....</b>                                                                  | <b>23</b> |
| <b>Structure of the thesis.....</b>                                                              | <b>25</b> |
| <b>1.    Introduction .....</b>                                                                  | <b>27</b> |
| <b>1.1.    Complex traits .....</b>                                                              | <b>27</b> |
| 1.1.1.    Metabolic health .....                                                                 | 29        |
| <b>1.2.    Genome-wide association studies (GWAS).....</b>                                       | <b>37</b> |
| 1.2.1.    Missing heritability.....                                                              | 47        |
| 1.2.2.    Meta-analysis and Trans-ethnic meta-analysis of multiple GWAS .....                    | 47        |
| 1.2.1.    Multi-phenotype genome-wide association studies .....                                  | 50        |

|             |                                                                             |            |
|-------------|-----------------------------------------------------------------------------|------------|
| 1.2.1.1.    | Multi-phenotype effects .....                                               | 50         |
| <b>1.3.</b> | <b>Missing data .....</b>                                                   | <b>52</b>  |
| 1.3.1.      | Missing data patterns and mechanisms .....                                  | 54         |
| <b>2.</b>   | <b>Review of the literature.....</b>                                        | <b>58</b>  |
| 2.1.        | Polygenic risk score (PRS) calculation.....                                 | 62         |
| 2.2.        | Correlated phenotype analysis .....                                         | 66         |
| 2.3.        | Conditional analysis on multi-phenotype GWAS .....                          | 70         |
| 2.4.        | Several approaches to tackle missing phenotype data.....                    | 73         |
| 2.4.1.      | Imputation methods based on statistical approaches.....                     | 74         |
| 2.4.2.      | Imputation methods based on Machine Learning (ML) approaches.....           | 82         |
| <b>3.</b>   | <b>Materials and Methods .....</b>                                          | <b>93</b>  |
| 3.1.        | Study populations.....                                                      | 93         |
| 3.1.1.      | Northern Finland birth cohort (NFBC) 1966 & 1986 .....                      | 93         |
| 3.2.        | Statistical analyses.....                                                   | 100        |
| <b>4.</b>   | <b>GWAS of glycemic traits in two healthy birth cohorts.....</b>            | <b>101</b> |
| 4.1.        | Meta-analyses of Glucose and Insulin-related traits Consortium (MAGIC)..... | 101        |
| 4.2.        | Study populations.....                                                      | 103        |
| 3.2.1.      | Genetic data .....                                                          | 105        |
| 3.2.2.      | Phenotypic data .....                                                       | 106        |
| 4.3.        | Methods.....                                                                | 106        |
| 3.3.1.      | Variance explained by polygenic risk score (PRS) calculation.....           | 106        |
| 4.4.        | Results.....                                                                | 110        |
| 4.5.        | Discussion and conclusions.....                                             | 115        |
| <b>5.</b>   | <b>MP-GWAS of Fatty acid (FA) metabolism .....</b>                          | <b>118</b> |
| 5.1.        | Introduction .....                                                          | 118        |
| 5.2.        | Methods.....                                                                | 119        |
| 4.2.1.      | Phenotype imputation .....                                                  | 119        |
| 4.2.2.      | Conditional analysis .....                                                  | 120        |
| 5.3.        | Results.....                                                                | 123        |

|        |                                                                                                         |     |
|--------|---------------------------------------------------------------------------------------------------------|-----|
| 5.4.   | Discussion and conclusions.....                                                                         | 128 |
| 6.     | Implementation of phenotype imputation methods to the MP-GWAS framework                                 |     |
|        | 132                                                                                                     |     |
| 6.1.   | Introduction .....                                                                                      | 132 |
| 6.2.   | Methods.....                                                                                            | 133 |
| 6.2.1. | Simulation of genetic data .....                                                                        | 134 |
| 6.2.2. | Simulation of phenotype data .....                                                                      | 135 |
| 6.2.3. | Imputation methods .....                                                                                | 135 |
| 6.2.4. | Association analysis within the MP-GWAS framework.....                                                  | 137 |
| 6.3.   | Results.....                                                                                            | 138 |
| 6.4.   | Discussion and conclusions.....                                                                         | 143 |
| 7.     | Software development for phenotype imputation in the MP-GWAS framework.                                 | 149 |
| 7.1.   | Introduction .....                                                                                      | 149 |
| 7.2.   | Methods.....                                                                                            | 150 |
| 7.2.1. | Implementation of the random forest imputation in C++ .....                                             | 150 |
| 7.2.1. | Data input and output.....                                                                              | 152 |
| 7.3.   | Comparison with existing RF implementations .....                                                       | 154 |
| 7.3.1. | Running time and memory .....                                                                           | 154 |
| 7.4.   | Discussion and conclusions.....                                                                         | 156 |
| 8.     | Phenotype imputation in the MP-GWAS framework: real data application on blood<br>metabolite levels..... | 162 |
| 8.1.   | Introduction .....                                                                                      | 162 |
| 8.2.   | Methods.....                                                                                            | 163 |
| 8.2.1. | Phenotype imputation .....                                                                              | 163 |
| 8.2.2. | Quality control (QC) and multi-phenotype analysis (MPA) .....                                           | 164 |
| 8.3.   | Results.....                                                                                            | 166 |
| 8.4.   | Discussions and conclusions .....                                                                       | 170 |
| 8.4.1  | Established signals in MP-GWAS.....                                                                     | 170 |
| 8.4.2  | Additional signals detected in the imputeSCOPA imputed data .....                                       | 171 |

|            |                                 |            |
|------------|---------------------------------|------------|
| <b>9.</b>  | <b>Discussion .....</b>         | <b>173</b> |
| 9.1.       | General discussion .....        | 173        |
| 9.2.       | Conclusion .....                | 177        |
| 9.3.       | Strengths and limitations ..... | 178        |
| 9.3.1.     | Strengths .....                 | 178        |
| 9.3.2.     | Limitations.....                | 179        |
| 9.4.       | Future directions.....          | 180        |
| <b>10.</b> | <b>References .....</b>         | <b>186</b> |
| <b>11.</b> | <b>Supplementary.....</b>       | <b>204</b> |

## List of Tables

|                                                                                                                                                                                                                                                    |     |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Table 1.1 Different diagnostic criteria of the metabolic syndrome by five organisations (source: (Ogedengbe & Ezeani, 2014)).                                                                                                                      | 30  |
| Table 2.1 The key literature used on the method development and implementation of SP-GWAS and MP-GWAS on cardiometabolic traits.                                                                                                                   | 59  |
| <i>Table 2.2. The key literature used the analysis and implementation of SP-GWAS and MP-GWAS on cardiometabolic traits.</i>                                                                                                                        | 61  |
| Table 3.1 Numbers of invited, participated, and gave participants consent that their data were provided from clinical data collection and questionnaire (area of catchment: NFBC1966 / NFBC1986).                                                  | 95  |
| Table 3.2 Genotype Quality Control criteria applied in the NFBC1966 and NFBC1986 studies.                                                                                                                                                          | 97  |
| Table 3.3 The proportion of missingness for each category of phenotypes in NFBC1966 and NFBC1986. The means and medians were calculated from the phenotypes in the same category.                                                                  | 99  |
| Table 3.4 The statistical methods used in each analysis in this thesis.                                                                                                                                                                            | 100 |
| Table 4.1 Variance explained results for NFBC1966                                                                                                                                                                                                  | 111 |
| Table 4.2 Variance explained results for NFBC1986                                                                                                                                                                                                  | 111 |
| Table 4.3 Results (adjusted R <sup>2</sup> ) from testing cohorts. Covariates as fitted in GWAS (age, age <sup>2</sup> , BMI, gender, PCs (if needed)).                                                                                            | 114 |
| Table 5.1 The distinct signals at the identified loci for FAs.                                                                                                                                                                                     | 124 |
| Table 6.1 A comparison of the seven imputation approaches applied to the simulated datasets in the present study.                                                                                                                                  | 142 |
| Table 6.2 The performance comparison between missing data imputation methods which indicates that RF is the best imputation in term its performance, the ability to handle mix type of data, and the easy setting that require no parameter tuning | 143 |
| Table 7.1 Computational time comparison between missForest and imputeSCOPA in a different number of individuals between 5,000 – 500,000 individuals.                                                                                               | 155 |
| Table 8.1 A summary of multicollinearity of the 31 metabolites used in the multi-phenotype analysis (MPA).                                                                                                                                         | 164 |
| Table 8.2 – Signals detected in the full-model of meta-analysis MP-GWAS, consist of NFBC1966 and NFBC1986 cohorts using complete case analysis and imputeSCOPA imputed data.                                                                       | 168 |

## List of Figures

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |    |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 1.1 Types of Cholesterol: HDL, LDL, and Triglycerides. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 34 |
| Figure 1.2 Types of FAs according to the number of carbon double bonds that represent the saturation. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 35 |
| Figure 1.3 Fine-mapping workflow. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 44 |
| Figure 1.4 The intersection of established GWAS loci with cardiometabolic traits. (source:(Marullo et al., 2014)). ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 50 |
| Figure 1.5: Several possible pleiotropy scenarios (source: Solovieff et al., 2013)....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 51 |
| Figure 1.6 Illustration of the proportion of missing data increase in MP-GWAS as compared with single-phenotype GWAS.....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 54 |
| Figure 1.7 An example of missing data patterns in multivariate data. Blue indicates observed data, and grey indicates missing data (van Buuren, 2018). ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 56 |
| Figure 2.1 The workflow of the selected literature topics used in the thesis. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 58 |
| Figure 2.2 (A) The interactions and additive effects of multiple genetic variants and environmental factors affects common diseases and traits. (b) Polygenic risk scores (PRS) are usually computed by summing independent (i.e., uncorrelated) single nucleotide polymorphisms (SNPs) weighted allele counts observed in a genome-wide association analysis (GWAS) to be associated with a phenotype or disease. The weights are the linear regression effect sizes for quantitative characteristics, i.e., The regression coefficients ( $\beta$ ). The weights are the normal logarithm of the odds ratio for case/control phenotypes, $\ln(OR)$ . .... | 62 |
| Figure 2.3 Workflow for the conditional MP-GWAS analyses.....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 72 |
| Figure 2.4 Different simple imputation (SI) approaches. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 74 |
| Figure 2.5 The limitation of regression imputation or single imputation (SI) where the predicted values lie on the regression line will overestimate the correlations but underestimate the variances. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 76 |
| Figure 2.6 Regression coefficient ( $\beta$ ) and Standard Error (SE) formulated for multiple imputation (MI), either using the Bayesian approach or Expectation-Maximisation (EM) bootstrapping, based on Rubin's rule. The number of imputed datasets is M; the coefficient of regression is $\beta_i$ for phenotype i, SE is the standard error.....                                                                                                                                                                                                                                                                                                     | 78 |
| Figure 2.7 Workflow of the random forest (RF) algorithm.....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 84 |
| Figure 2.8 Workflow of bootstrap sample generation with replacement. "Sampling-with-replacement" here implies that if a data point is selected in the first random subsample, another random subsample will follow with equal probability remains in the original sample for selection. This replacement sampling can be seen on the picture as "C", which is chosen twice in the second bootstrap subsample.....                                                                                                                                                                                                                                           | 85 |
| Figure 2.9 Illustration of the k nearest neighbour (kNN) algorithm. The oval area reflects, in this case, the adjacent points of the green square data point. The calculation of space is used using Euclidean distance to define neighbours.....                                                                                                                                                                                                                                                                                                                                                                                                           | 87 |
| Figure 2.10 The calculation of the distance space in the presence of missingness of the k nearest neighbour (kNN) algorithm.....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 88 |

|                                                                                                                                                                                                                                                            |     |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Figure 2.11 The data fitted with quantile regression.....                                                                                                                                                                                                  | 90  |
| Figure 3.1 Geographical location of the Northern Finland Birth Cohort (NFBC)1966 and 1986. Blue represents the targeted area of the cohorts. ....                                                                                                          | 93  |
| Figure 4.1 Comparison of PRS derived using EA variants versus EA plus transethnic variants.....                                                                                                                                                            | 112 |
| Figure 5.1 The workflow for the conditional MP-GWAS analyses .....                                                                                                                                                                                         | 122 |
| Figure 5.2 Regional plots of novel loci attaining statistical significance of $P < 5 \times 10^{-8}$ in our MP-GWAS meta-analysis of FAs. A) LPXN, and B) RPS6KA4.....                                                                                     | 125 |
| Figure 5.3 Regional plots of the APOE region before and after conditional MP-GWAS. A) unconditional, B) conditional on rs7412, C) conditional on rs7412 and rs429358. ....                                                                                 | 125 |
| Figure 5.4 Regional plots of previously established lipid loci attaining statistical significance of $P < 5 \times 10^{-8}$ in our MP-GWAS meta-analysis of FAs. A) PCSK9, B) GCKR, C) FADS1/2, D) ZNF259, E) LIPC, F) PDXDC1, G) SUGP1, and H) APOE. .... | 126 |
| Figure 5.5 Regional plots of the LIPC region before and after conditional MP-GWAS. A) unconditional, B) conditional on rs1532085, C) conditional on rs1532085 and rs2070895, and D) conditional on rs1532085, rs2070895 and rs117376818. ....              | 127 |
| Figure 5.6 Regional plots of the SUGP1 region before and after conditional MP-GWAS. A) unconditional, B) conditional on rs8100204, C) conditional on rs8100204 and rs12463074. ....                                                                        | 127 |
| Figure 5.7 Regional plots of the FADS1/2/3 region before and after conditional MP-GWAS. A) unconditional, B) conditional on rs174564, C) conditional on rs174564 and rs174590. ....                                                                        | 128 |
| Figure 6.1 Workflow diagram of genotype and phenotype data simulations and missing data imputation methods .....                                                                                                                                           | 134 |
| Figure 6.2 The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis in MP-GWAS analysis. ....                                                     | 140 |
| Figure 6.3 The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis in MP-GWAS analysis in 5,000 and 50,000 combined. ....                        | 141 |
| Figure 6.4 The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis in MP-GWAS analysis in 5,000 and 50,000 combined. ....                        | 148 |
| Figure 7.1 Workflow of an imputeSCOPA, including input file, commands, and output file.....                                                                                                                                                                | 153 |
| Figure 7.2 imputeSCOPA, command run on terminal .....                                                                                                                                                                                                      | 153 |
| Figure 7.3 imputeSCOPA input file containing missing data before imputation.....                                                                                                                                                                           | 154 |
| Figure 7.4 imputeSCOPA output file after imputation.....                                                                                                                                                                                                   | 154 |
| Figure 7.5 Classification of random forest (RF) software, based on their abilities to handle high-dimensional data, imputation support, and whether their codes were written in C++ or R. ....                                                             | 157 |

Figure 8.1 – Manhattan plot of MP-GWAS results using complete case data. The X-axis indicates the genomic position with the association P-value for each tested SNP on the Y-axis. The genome-wide significance threshold of  $5 \times 10^{-8}$  and the suggested significance threshold of  $5 \times 10^{-6}$  are represented with the red line and the blue line.

..... 167

Figure 8.2 – Manhattan plot of MP-GWAS results using imputeSCOPA imputed data. The X-axis indicates the genomic position with the association P-value for each tested SNP on the Y-axis. The genome-wide significance threshold of  $5 \times 10^{-8}$  and the suggested significance threshold of  $5 \times 10^{-6}$  are represented with the red line and the blue line..... 167

## List of Supplementary Tables

|                                                                                                                                                                                                                                                    |     |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Supplementary Table 1 The variables of NFBC cohorts. ....                                                                                                                                                                                          | 204 |
| Supplementary Table 2 The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis of 5,000 individuals in MP-GWAS analysis. ....             | 208 |
| Supplementary Table 3 The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis of 50,000 individuals in MP-GWAS analysis. ....            | 209 |
| Supplementary Table 4 Root mean squared error (RMSE) boxplot obtained from the regression coefficient of seven imputation methods and those from full data analysis of 5,000 individuals in MP-GWAS analysis. ....                                 | 210 |
| Supplementary Table 5 Root mean squared error (RMSE) boxplot obtained from the regression coefficient of seven imputation methods and those from full data analysis of 50,000 individuals in MP-GWAS analysis. ....                                | 211 |
| Supplementary Table 6 The SNP list contains 394 SNPs that represent signals across any of the phenotypes analysed in the meta-analysis, and a few of them did not pass imputation quality and did not present either in NFBC1966 or NFBC1986. .... | 212 |
| Supplementary Table 7 The RMSE of regression coefficient between different imputation methods in 5,000 individuals and three phenotypes. ....                                                                                                      | 221 |
| Supplementary Table 8 The RMSE of regression coefficient between different imputation methods in 5,000 individuals and 30 phenotypes. ....                                                                                                         | 225 |
| Supplementary Table 9 The RMSE of regression coefficient between different imputation methods in 5,000 individuals and 120 phenotypes. ....                                                                                                        | 229 |
| Supplementary Table 10 The RMSE of regression coefficient between different imputation methods in 50,000 individuals and three phenotypes. ....                                                                                                    | 233 |
| Supplementary Table 11 The RMSE of regression coefficient between different imputation methods in 50,000 individuals and 30 phenotypes. ....                                                                                                       | 237 |
| Supplementary Table 12 The RMSE of regression coefficient between different imputation methods in 50,000 individuals and 120 phenotypes. ....                                                                                                      | 240 |
| Supplementary Table 13 The RMSE of regression coefficient between different imputation methods in 500,000 individuals and three phenotypes. ....                                                                                                   | 243 |
| Supplementary Table 14 The RMSE of regression coefficient between different imputation methods in 500,000 individuals and 30 phenotypes. ....                                                                                                      | 246 |
| Supplementary Table 15 The RMSE of regression coefficient between different imputation methods in 500,000 individuals and 120 phenotypes. ....                                                                                                     | 249 |

## Abbreviations

|          |                                                |
|----------|------------------------------------------------|
| 1KGP     | 1000 Genomes Project                           |
| BMI      | Body mass index                                |
| CC       | Complete cases                                 |
| CVD      | Cardiovascular diseases                        |
| DT       | Decision tree                                  |
| FA       | Fatty acid                                     |
| FG       | Fasting glucose                                |
| FI       | Fasting insulin                                |
| GWAS     | Genome-wide association study                  |
| GRS      | Genetic Risk Score                             |
| HDL      | High-density lipoprotein                       |
| HWE      | Hardy-Weinberg equilibrium                     |
| KNN      | K-nearest neighbour                            |
| LD       | Linkage disequilibrium                         |
| LDL      | Low-density lipoprotein                        |
| MAF      | Minor allele frequency                         |
| MAR      | Missing at random                              |
| MCAR     | Missing completely at random                   |
| MDS      | Multidimensional scaling                       |
| MetS     | Metabolic syndrome                             |
| MI       | Multiple imputation                            |
| MNAR     | Missing not at random                          |
| MPA      | Multi-phenotype analysis                       |
| MP-GWAS  | Multi-phenotype genome-wide association study  |
| MTAG     | Multi-Trait Analysis of GWAS (MTAG)            |
| NFBC1966 | Northern Finland Birth Cohort 1966             |
| NFBC1986 | Northern Finland Birth Cohort 1986             |
| NMR      | Nuclear magnetic resonance                     |
| OOB      | Out of bag                                     |
| PC       | Principle component                            |
| RF       | Random forest                                  |
| SE       | Standard error                                 |
| SI       | Single imputation                              |
| SNP      | Single nucleotide polymorphism                 |
| SP-GWAS  | Single-phenotype genome-wide association study |

|     |                           |
|-----|---------------------------|
| QC  | Quality control           |
| T1D | Type 1 diabetes           |
| T2D | Type 2 diabetes           |
| TG  | Triglycerides             |
| VIF | Variance inflation factor |

# The overall aim, hypotheses, research questions, scope, and structure of the thesis

## The overall aim of the thesis

This thesis aims to elucidate the genetic architecture that underlies cardiometabolic traits, and particularly focuses on addressing the missing heritability issue.

The first approach was to enlarge the sample size by applying the meta-analysis single-phenotype genome-wide association studies (SP-GWAS) ([Mahajan et al., 2018](#); [Ng et al., 2014](#); [Prokopenko et al., 2009](#); [Wheeler, Leong, et al., 2017](#)).

Our team previously contributed to a large meta-analysis SP-GWAS of glycaemic traits, fasting glucose (FG), and fasting insulin (FI). We collaborated with other Meta-analyses of Glucose and Insulin-related traits Consortium (MAGIC) investigators combined data from multiple GWAS using a trans-ethnic approach to pinpoint additional sites that affect glycaemic and metabolic properties. I contributed to this analysis as part of the consortium collaboration to calculate the variance explained by FG and FI in Northern Finland Birth Cohort (NFBC) 1966 and 1986.

However, many of the genetic variants identified in the SP-GWAS are associated with more than one trait ([Marullo, El-Sayed Moustafa, & Prokopenko, 2014](#)). Therefore, joint multi-phenotype analysis in high-dimensional omics data sets serves as a better approach to move beyond the traditional SP-GWAS.

A better power of multi-phenotype genome-wide association studies (MP-GWAS) has been proposed by our team ([Mägi et al., 2017](#)) to detect multi phenotypes effects that

can be applied even in a small dataset. Therefore, the development and implementation of this MP-GWAS approach are the main focus of this thesis.

After conducting direct conditional MP-GWAS on FA metabolism, I found that the proportion of missing phenotype data increased gradually in the MP-GWAS framework. This missing data problem leads to my investigation to find the least biased and efficient imputation method and imputation software development. The specific aims in this thesis are the following:

**Aim 1:** To define the variance explained from glycaemic trait-associated loci detected within the largest to date single-trait trans-ethnic GWAS meta-analysis. This thesis focused on calculating the variance explained by fasting glucose (FG) and fasting insulin (FI) in NFBC1966 and NFBC1986.

**Aim 2:** To improve the power of locus discovery by implementing the MP-GWAS approach within the analysis of blood metabolite FAs. My role in this analysis was to conduct conditional analysis to detect the independent signals identified in our team's MP-GWAS of FA metabolism.

**Aim 3:** To define and implement an efficient approach for MP imputation of epidemiological data for high-dimensional multi-omics studies. This aim consisted of three objectives.

**Objective 3.1:** To find the most efficient and least biased imputation method in the multi-phenotype analytical framework.

**Objective 3.2:** To implement the least biased imputation method into an imputation software.

**Objective 3.3:** To apply the imputation method within MP-GWAS in real data analyses to boost the power of association and elucidate the genetic architecture of cardiometabolic traits.

### Study-specific hypotheses

The hypotheses tested in this thesis were:

1. Increasing the sample size by incorporating large-scale trans-ancestry SP-GWAS meta-analysis of glycaemic traits will detect more additional loci; thus, increasing the variance explained by the glycaemic trait-associated loci.
2. Improved power of the MP-GWAS approach within the analysis of blood metabolite fatty acids (FAs) will detect novel associated loci and multi-phenotype effects associated with loci.
3. Subsequent conditional analysis within the MP-GWAS framework may enable us to refine the genetic architecture underlying identified loci by identifying the multiple distinct signals affecting FA metabolism.
4. Comparing advanced imputation methods within the MP-GWAS framework using simulated data under different scenarios will provide the least biased and most efficient imputation method.
5. When the least biased imputation method is inefficient in imputing high-dimensional omics data, developing an imputation software in the fastest low-level programming language C++ will decrease its computational time.
6. Imputing phenotype data with the least biased imputation method within the MP-GWAS framework will improve the power to detect novel loci associated with lipids compared to the standardly used complete case (CC) analysis.

## Research questions for the thesis

1. Does increasing the sample size in trans-ethnic meta-analysis GWAS of glycemic traits increase the number of identified loci and variance explained?
2. Does moving to the MP-GWAS approach increase the power to discover susceptibility loci and suggest a multi-phenotype effect on FA metabolisms? More specifically, does subsequent conditional analysis within the MP-GWAS framework refine genetic loci affecting FA metabolism by suggesting multiple distinct signals at the identified loci?
3. Does moving to high dimensional multi-phenotype analyses of complex, high-dimensional datasets raise many statistical problems, i.e., missing data problem? Does remove the missing data reduce the statistical power to detect loci in the CC analysis within the MP-GWAS framework?
4. What is the best imputation method within the MP-GWAS framework that gives the least biased parameter estimates under different missing data proportions, mechanisms, sample sizes, and the correlation between samples?
5. Is the least biased imputation method recommended in the simulated study applicable and efficient in imputing a mixture of complex, high-dimensional data, such as UK Biobank? If not, does implementing this imputation method in the fastest programming language C++ increase the efficiency in imputing high-dimensional data?
6. Does imputing the phenotype data with the least biased imputation method improve the power to detect novel loci associated with lipids compared to the complete case (CC) analysis within the MP-GWAS framework?

## Scope of the thesis

GWAS have yielded a wealth of biological insight in uncovering the genetic architecture of cardiometabolic traits. However, the traditional GWAS conducted in this field only focus on one trait at a time, failing to capture the metabolic complexity underlying the correlated cardiometabolic traits.

Joint analysis of correlated phenotypes in high-dimensional datasets serves as an advanced methodology to move beyond traditional SP-GWAS. Therefore, this thesis focuses on developing and applying the MP-GWAS framework to improve the power of locus discovery.

The first project in this thesis is in the SP-GWAS framework setting, which I contribute to calculating the variance explained for both fasting glucose (FG) and fasting insulin (FI) in NFBC1966 and NFBC1986 data as part of the cohorts contributing to the meta-analysis GWAS in MAGIC. The polygenic risk score (PRS) calculation method that was used in this analysis was chosen primarily due to this method was chosen by the MAGIC principal investigator.

The next project in this thesis applied the multi-phenotype methodology based on the in-house software SCOPA developed by our team that implemented a reverse regression approach as opposed to the commonly used SP-GWAS.

I extended the MP-GWAS methodology's applicability by designing and implementing a direct conditional analysis within the MP-GWAS framework as a subsequent analysis to detect distinct signals at the identified loci. This conditional analysis is the first method designed in the MP-GWAS framework, although the MP-GWAS methodology has been recently applied in several studies.

As MP-GWAS tends to increase the proportion of missing data compared to the SP GWAS, the missing data problem needs to be handled properly within this framework. I investigated seven imputation methods and compared their performance in an extensive simulated study. The simulated data represented different sample sizes, missing data mechanisms, and the proportion of missing data. The imputation methods were selected due to their available R software packages, limited to only seven imputation methods commonly used in the literature.

The missing data mechanisms were generated separately for each simulated data to observe whether the imputation methods predict different performance with different missing data mechanisms. However, the real dataset may contain a mixture of missing data mechanisms, which was not generated in this simulated study due to the time constraint. The confounding variables were also not implemented in this simulated study.

The imputation method developed in this thesis was intended to be used in high-dimensional data like UK Biobank; instead, the imputation for a similar sample-size was performed on simulated data. The imputation in the real dataset in this thesis was performed in NFBC1966 and NFBC198 phenotype data, which detected novel signals associated with lipids within the MP-GWAS framework.

The least biased imputation method developed as an independent imputation software in this thesis outperformed existing state-of-the-art methods from the statistics and machine learning (ML) literature. It can boost signals of the association on various simulated data sets from various scenarios and real datasets.

## Structure of the thesis

The thesis is divided into nine chapters that correspond to the ‘background and literature study’, ‘materials and methods’, ‘four main projects’, and ‘overall conclusion’.

Chapter 1 gives a brief introduction to the project and the motivations behind it, along with general concepts of metabolic health, GWAS, and missing data problems.

Chapter 2 introduces the methods used in the analyses compared to the other available methods from the literature. Several imputation methods, which are the central theme of this thesis, are also discussed.

Chapter 3 introduces the contributing study populations and methods used in each of the analyses. This chapter describes the Northern Finland Birth Cohort (NFBC) 1966 and 1986 data collection, including genetic and phenotype data.

Chapter 4 describes the variance explained analysis within MAGIC for signals detected in meta-analysis GWAS of glycaemic traits using PRS calculation.

Chapter 5 presents the multi-phenotype fine-mapping through direct conditional MP-GWAS of FA metabolism and their role in complex diseases.

Chapter 6 compares the performance of several imputation methods to simulated datasets within the MP-GWAS framework.

Chapter 7 describes software development imputeSCOPA based on the best imputation method from the study in Chapter 5, including the prototype development of software and the method.

Chapter 8 presents phenotype imputation using imputeSCOPA in real data analyses.

Chapter 9 summarises the principal study observations, conclusions, and suggestions for future work.

# 1. Introduction

*In this chapter, the main concepts relevant to this thesis are introduced. The chapter introduces complex traits and metabolic health and moves through genome-wide association studies to multi-phenotype genome-wide association studies (MP-GWAS). Finally, I discuss the missing data problem in large-scale genomic studies.*

## 1.1. Complex traits

The term trait or phenotype relates to the body's physio-physical features, including its appearance, growth, and behavioural characteristics ([Nature Education, 2014](#)). Height, weight, hair colour, and observable features that can be measured in the laboratory, such as the content of hormones or blood cells, are examples of phenotypes. The phenotypes of an organism are controlled by their genotypes, i.e., the number of genes carried by the organism and environmental conditions affecting these genes ([Nature Education, 2014](#)).

In many cases, phenotypes are quantitative and complex, containing combinations of several ecological and genetic causes ([J. Craig, 2008](#)). Such a combination is in line with the experience of many researchers seeking to identify the roots of genetic and environmental causes associated with complex forms of human disease. Thus, complex diseases are largely developed due to a mixture of genetic and ecological factors, most of which remain unidentified ([J. Craig, 2008](#)).

To understand the subtleties of complex diseases, Mendel's three important inheritance laws must first need to be understood: the principle of segregation, the principle of dominance, and the principle of independent assortment ([Sandler & Sandler, 1986](#)). These principles are used to explain the way inherited traits,

particularly those underlying disease, are transmitted across generations ([Sandler & Sandler, 1986](#)).

However, since the mid-18th century, researchers have identified various situations in which Mendel's rules of inheritance have been violated, which may complicate pinpointing the determinants that affect complex genetic diseases or common diseases ([A Century Mendelism Hum. Genet., 2000](#)).

Modern researchers are aware that complex diseases are not governed by the dominant or recessive Mendelian pattern of a single gene ([van Heyningen & Yeyati, 2004](#)). Although genetic factors are inherited, they only contribute to complex disease risks ([J. Craig, 2008](#)). Predisposing genetically indicates that an individual is genetically susceptible for a particular disease to develop, but this does not necessarily imply that an individual with inherited genes will develop the disease ([van Heyningen & Yeyati, 2004](#)). Genuine progression of the common disease-related phenotype is significantly influenced by the environment and lifestyle of the individual. We can change our living habits and improve the environment to anticipate or control such disease's onset, yet we cannot change our genes ([Dempfle et al., 2008](#)).

Genetic and ecological factors' co-existence is a constant challenge for scientists in elucidating the architecture of complex diseases ([Dempfle et al., 2008](#)). Therefore, dissecting the associations between genes and environment could help understand the causes of complex diseases and help researchers develop effective therapies.

Nowadays, scientists are aware that genes and the influence of environmental damage can interact at the molecular level. They now study the interaction of genes and external factors that can be measured at a molecular level as metabolic traits rather than studying them separately. Researchers can examine the overall picture to

identify genetic risk factors that can be modified for environmental reasons (Dempfle et al., 2008).

### **1.1.1. Metabolic health**

Over a decade, genome-wide association studies (GWAS) have uncovered the genome influences in normal metabolic processes that can also be influenced by external factors, such as diet, exercise, etc. However, existing literature shows poor agreement on metabolic health definitions (Stefan, Häring, Hu, & Schulze, 2013). Meanwhile, this topic has attracted a lot of attention recently, with more and more studies being conducted with definitions of metabolic health in the Body Mass Index (BMI) classifications, both as a risk factor or as a result (Bell, Kivimaki, & Hamer, 2014; Kramer, Zinman, & Retnakaran, 2013). Therefore, assessing the predictive value of the current definitions of metabolic health was extremely crucial as an accurate risk classification that can validate successful prevention in high-risk or vulnerable groups (Stefan et al., 2013).

In a study, 8,721 participants were examined by National Health and Nutrition Examination Survey (NHANES), resulting in that just 1 in 8 Americans have optimal metabolic health (Araújo, Cai, & Stevens, 2019). The study mentioned that lifestyle factors affect metabolic health, measured by normal BMI, blood pressure, lipids concentrations, and blood sugar. These factors directly relate to a person's risk for cardiometabolic diseases; therefore, their blood levels can determine whether one can have metabolic health.

Overweight and metabolic syndrome (MetS), two very common and often coexisting diseases, are important determinant conditions for type 2 diabetes (T2D) and cardiovascular diseases (CVD) (Grundy, 2016; Grundy, Hansen, Smith, Cleeman, &

Kahn, 2004; Zimmet, Alberti, & Shaw, 2001). MetS results predominantly from an unbalance between calorie supply and energy demand. Still, they are also influenced by the genetic or epigenetic composition of the individual, the superiority of the sedentary lifestyle over physical activity, and other factors such as the nutritional quality and composition, and the composition of intestinal microbes (Oresic, 2012). MetS is a group of traits that occur together and heighten the risk of T2D, hypertension, and heart disease. Common conditions include elevated blood pressure, blood sugar levels, excessive waist fat, and deviating triglyceride or cholesterol or measurements (Grundy, 2016). Worldwide, MetS contributes to the major global health, which has been diagnosed in 23% of American adults in the US and 29% of Australia's adult population (Saklayen, 2018). MetS was used as a reference for several definitions of metabolic health, a clinically useful risk indicator for future T2D and avoid disease and fatality.

There have been several sets of diagnostic criteria for MetS defined by several organisations. Some of them are defined by the World Health Organization (WHO) in 1999, The National Cholesterol Education Program Adult Treatment Panel III (NCEP-ATP III) in 2001, Revised NCEP (NCEP-R) in 2004, International Diabetes Federation (IDF) in 2006, and American Association of Clinical Endocrinologists (AACE) in 2003 as described in Table 1.1 (Ogedengbe & Ezeani, 2014). The common risk factors among these different sets of criteria for Met-S are obesity, serum TG, serum HDL, and BP, although the threshold and interpretation are not the same.

*Table 1.1 Different diagnostic criteria of the metabolic syndrome by five organisations (source: (Ogedengbe & Ezeani, 2014)).*

|                        | WHO (1999)                                                                                  | NCEP-ATP III (2001)                       | NCEP-R (2004)                             | IDF (2005)                                                | AACE                                                                                                                                                                                      |
|------------------------|---------------------------------------------------------------------------------------------|-------------------------------------------|-------------------------------------------|-----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Obesity                | WHR<br>>0.90(male)<br>>0.85(female) or<br>BMI>30 kg/m <sup>2</sup>                          | WC<br>≥ 102 cm (male)<br>≥ 88 cm (female) | WC<br>≥ 102 cm (male)<br>≥ 88 cm (female) | [REQUIREMENT]<br>WC<br>≥ 94 cm (male)<br>≥ 80 cm (female) | Overweight/ Obesity<br>BMI ≥ 25 kg/m <sup>2</sup>                                                                                                                                         |
| Serum triglycerides    | ≥ 150mg/dl                                                                                  | ≥ 150 mg/ dl                              | ≥ 150 mg/ dl<br>or medication             | ≥ 150 mg/ dl<br>or medication                             | ≥ 150 mg/ dl                                                                                                                                                                              |
| Serum HDL              | < 35 mg/ dl (male)                                                                          | < 40 mg/ dl (male)                        | < 40 mg/ dl (male)                        | <40 mg/ dl (male)                                         | <40 mg/ dl (male)                                                                                                                                                                         |
| Cholesterol            | < 39 mg/ dl (female)                                                                        | < 50 mg/ dl (female)                      | < 50 mg/ dl (female)<br>or medication     | <50 mg/ dl (female)<br>or medication                      | <50 mg/ dl (female)                                                                                                                                                                       |
| Blood pressure         | ≥ 140/ 90 mmHg                                                                              | ≥ 130/85 mmHg<br>or medication            | ≥ 130/85 mmHg<br>or medication            | ≥ 130/85 mmHg<br>or medication                            | ≥130/ 85 mmHg<br>or medication                                                                                                                                                            |
| Fasting plasma glucose | [REQUIREMENT]<br>FPG≥110 mg/ dl                                                             | ≥ 100mg/ dl                               | ≥ 100mg/ dl                               | ≥ 100mg/ dl<br>or previously diagnosed<br>T2DM            | 110-126 mg/ dl                                                                                                                                                                            |
| Other risk factors     | Urinary albumin excretion<br>rate ≥ 20μg/ min<br>or albumin / creatinine<br>ratio ≥ 30 mg/g |                                           |                                           |                                                           | Family history of T2DM,<br>HTN, or CVD. Polycystic<br>ovary syndrome, sedentary life<br>style, Advancing age and ethnic<br>groups having high risk for DM<br>or CVD Physician's judgement |
| Diagnosis              | Impaired FPG+any 2 criteria                                                                 | Any 3 criteria                            | Any 3 criteria                            | WC+any 2 criteria                                         |                                                                                                                                                                                           |

T2DM: Type 2 diabetes mellitus, HTN: Hypertension, CVD: Cerebrovascular accident, DM: Diabete mellitus, WC: Waist circumference. NCEP ATP III : National cholesterol education program adult tretment panel III, NCEP-R: NCEP-R: revised NCEP, IDF: International diabetes federation, ACE: American association of clinical endocrinologists, WHO: World Health Organization, WHR: Waist-to-hip ratio, BMI: Body mass index, HDL: high density lipoprotein

MetS predicted about 25% of new CVD cases, as was reported from a study on the Framingham district in 2004 (Grundy, Brewer, Cleeman, Smith, & Lenfant, 2004). Therefore, these MetS criteria, including weight, weight to hip ratio, and BMI, have been treated as risk factors for CVD in many GWAS (Emdin et al., 2017; Karaderi, Drong, & Lindgren, 2015). Glycaemic properties such as fasting glucose (FG) and fasting insulin (FI) measurements that are usually used to diagnose T2D have also been included to diagnose CVD. They are also risk factors for CVD. Thus, studying them could yield additional insights into the pathophysiology of T2D and CVD. An extensive GWAS have been performed to date, leading to the detection of over 97 sites affecting glycaemic traits (Wheeler, 2017).

#### 1.1.1.1. Blood metabolites

Clinical diagnoses such as blood pressure and anthropometric measurements are usually used as risk factors in GWAS to study cardiometabolic diseases (Emdin et al.,

2017; Karaderi et al., 2015). However, metabolic molecules, known as metabolites, are traits that provide a new compendium of associations between genetic regions and metabolite levels that also could be used in GWAS. Metabolites are small molecules that makeup or are the products of all biological pathways within cells, which can help us understand the metabolic reactions used in diagnostic tests for a wide range of metabolic disorders (Illig et al., 2010; Suhre & Gieger, 2012). These metabolites are also used in clinical diagnosis, for example, high fasting blood glucose levels to detect T2D risk ((Zimmet et al., 2001), as cited in Inzucchi (2013)). Therefore, complex diseases associated with metabolic health issues can be diagnosed by changes in metabolite levels in our body (Illig et al., 2010; Suhre & Gieger, 2012).

Metabolites as metabolism products include different molecules such as lipids, vitamins, glucose, and nucleotides. Metabolomic and genomic technologies allow the accurate measurements of these substances to be used as risk factors of cardiometabolic diseases. However, studying genetic diversity and metabolites together is hindered by the traditional genomic and metabolomic analyses performed separately. Studies focused on the interaction between metabolic change associated with genetic variants have recently been expanded into larger studies that analyse gene-metabolite associations. Metabolomic GWAS have been introduced to analyse high-dimensional metabolomic and genomic data to conduct statistical associations between metabolite levels and genetic variants. This association is because metabolic diseases are characterised by changes in metabolite levels associated with questions of metabolism and metabolic processes in the body (Illig et al., 2010; Suhre & Gieger, 2012).

Blood metabolites represent small molecules as intermediates products of different metabolic reactions and cellular pathways regulated by gene-environment interactions

(Suhre & Gieger, 2012). Advanced metabolomic techniques enable identifying and measuring small molecules between 50 – 1500 Daltons (Da) in human blood. Their levels are widely used in diagnosis, e.g., high low-density lipoprotein (LDL) and low high-density lipoprotein (HDL) levels are used to diagnose the risk of cardiovascular disease (CVD) (Grundy, Hansen, et al., 2004). Some examples of blood metabolites include lipids, fatty acids (FAs), amino acids, etc.

The metabolic profile can be measured using high-throughput platforms of nuclear magnetic resonance metabolomics (NMR) as well as mass spectrometry (MS) (Würtz et al., 2016). NMR spectroscopy has the advantage of being quantitative, which does not require preparing the sample. However, the MS-based metabolomics approach serves as an excellent technic that provides a combination of sensitivity and selectivity for quantitative metabolomics research (Emwas, 2015). Several subtypes of metabolites can be quantified by MS spectroscopy: FAs, lipoproteins, and low-molecular-weight metabolites (LMWMs; size < 900 Da) (Würtz et al., 2016). These variant metabolites identified by MS are important as they have different cardiometabolic properties that have been linked to various health outcomes. High levels of unsaturated FAs, for example, can improve metabolic health, while a high concentration of saturated FAs can be a risk factor for cardiometabolic diseases (P. C. Calder, 2015).

Omega-3 FAs also other metabolite examples, which have advantages to lower blood pressure, reduce triglycerides, and increase insulin sensitivity (Kalupahana, Claycombe, & Moustaid-Moussa, 2011). In addition, omega-6 FAs in the form of linoleic acid also have a benefit in decreasing low-density lipoproteins (LDLs) (Cho, Dokko, & Chung, 2002).

#### 1.1.1.1.1. Lipoproteins

The role of lipoprotein is to transmit hydrophobic lipids (also known as fat), such as triglycerides (TG) and cholesterol, through the body's extracellular fluid or blood. Lipoproteins are of different types with different functions. In general, they can be divided into HDL and LDL. The liver absorbs the majority of LDL located at the end of the endogenous channel. Still, a small proportion of LDLs can penetrate and oxidise into the endothelium, leading to high blood cholesterol levels and the formation of atherosclerotic plaques inside the artery walls, increasing CVD risk (Brunzell et al., 2008).

Cholesterol levels are one of the risk factors that can be used to measure heart health. As opposed to LDL, having a higher HDL level is better, associated with a lower risk of heart disease. HDL involves the opposite route of cholesterol delivery to the liver by removing other cholesterol excess forms from the bloodstream (Kingsbury & Bondy, 2003).

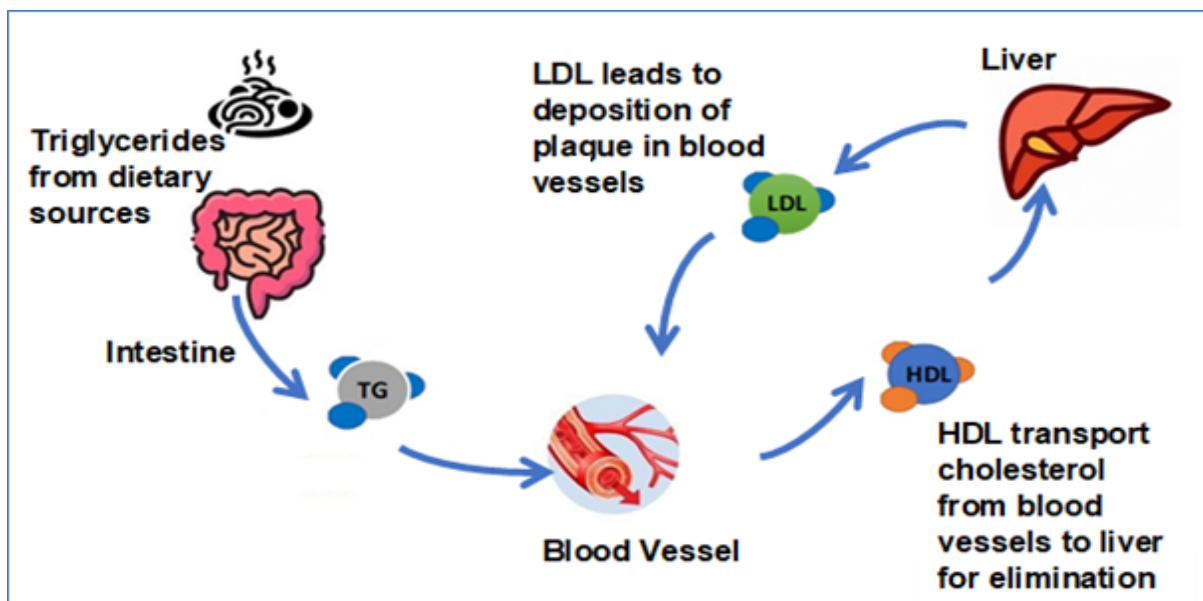


Figure 1.1 Types of Cholesterol: HDL, LDL, and Triglycerides.

Some cardiometabolic diseases have been associated with lipoprotein changes, particularly dyslipidemia, one of the components of MetS (Würtz et al., 2014). Furthermore, lipoproteins are widely known to be a risk factor predicting CVD, whereby higher LDL values and lower HDL levels indicate a greater risk (Brunzell et al., 2008).

#### 1.1.1.1.2. Fatty acids

One of the major risk factors of cardiometabolic diseases is obesity, closely related to an increased free fatty acid (FA) levels (Simopoulos, 2002, 2016). FAs can be categorised into two categories based on their chemical bonding: saturated and unsaturated FA, which the latest can be divided into monounsaturated and polyunsaturated FA, as seen in Figure 1.2.

Figure 1.2. Elevated saturated FAs represent a greater risk for CVD, with elevated unsaturated FAs representing a lower CVD risk (Philip C Calder, 2015).

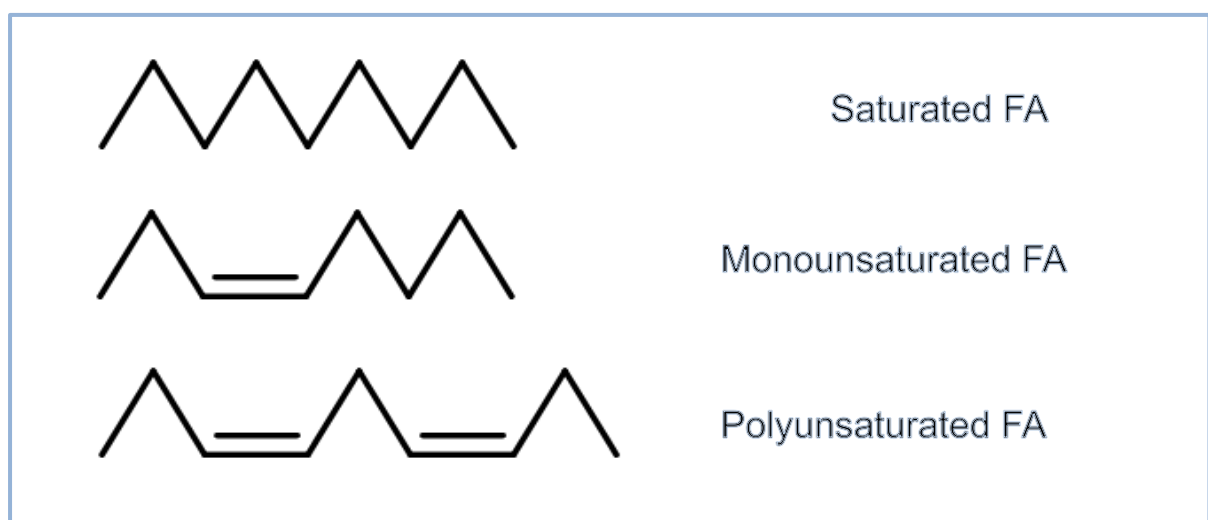


Figure 1.2 Types of FAs according to the number of carbon double bonds that represent the saturation.

FAs can be grouped as non-essential and essential FAs. Non-essential FAs can be synthesised in the body, whereas essential FAs must be obtained from the diet (Simopoulos, 2016).

The non-essential FAs are omega-7 and omega-9, which are monounsaturated FAs with one double bond. Omega-7s improves insulin sensitivity and cardiovascular inflammation by improving pancreatic production of insulin and lowering blood total and LDL cholesterol (Philip C Calder, 2015). Oleic acid, which is the most common omega-9 FA, has been reported to increase HDL cholesterol and decrease LDL cholesterol by eliminating plaque in the arteries, which causes myocardial infarction (Philip C Calder, 2015).

Two essential FAs are alpha-linolenic acid (an omega-3 FA) and linoleic acid (an omega-6 FA). Both of them are important in reducing the risk of heart disease by lowering blood pressure, total cholesterol and LDL cholesterol levels, and increasing HDL cholesterol levels (Philip C Calder, 2015). The supplements or diets rich in omega-3 FAs showed mildly but significantly decreased heart rate (Hidayat et al., 2018).

#### ***1.1.1.1.3. Low-molecular-weight metabolites***

LMWM are all metabolites with a molecular weight below 900 Da (Macielag, 2012). Many other non-lipid-based metabolites, characterised as LMWM, also contribute to metabolism in humans. For example, amino acids as organic molecules with an average molecular weight of 135 Da are essential to building blocks of proteins and enzymes. Amino acids, considering the large differences in their structures, include various properties and functions, also inside metabolic processes (Wu, 2009). They play many critical roles, like a synthesis of proteins, hormones, neurotransmitters, and metabolic signalling (Wu, 2009). Due to their functions, they are closely linked to the health of cardiac metabolism. For instance, branched-chain amino

acids (leucine, valine, and isoleucine) are important for signal transduction, glucose homeostasis, and muscle protein synthesis, as described in [Lynch & Adams \(2014\)](#).

The application of metabolomics technologies has surprisingly shown that branched-chain amino acids (BCAA) and related metabolites are more strongly associated with insulin resistance than many common lipids, although FA and FA-derived metabolites have a long-standing association with the development of insulin resistance and T2D ([Newgard, 2012](#)). In addition, the BCAA-related signature is predictive of the occurrence of diabetes and outcomes of interventions and responds in a unique way to therapeutic interventions. Nevertheless, BCAA supplementation in animal feeding studies requires a high-fat diet background to support insulin ([Newgard, 2012](#)).

BCAA as strong predictors of T2D is also positively associated with CVD risk ([Tobias et al., 2018](#)). It was discovered that total BCAAs were associated with a 23% greater risk of total CVD ([Tobias et al., 2018](#)).

BCAAs are essential amino acids that are converted to energy in the muscles. Although their plasma values are frequently found in high concentrations in individuals with CVD risk, their potential role with cardiovascular disease (CVD) is uncertain whether this is dependent on an intermediate diagnosis of T2D ([Grajeda-Iglesias & Aviram, 2018](#)). In recent years, they have proven to be potent biomarkers for cardiometabolic diseases as they have been linked to MetS ([Yamakado et al., 2015](#)).

## **1.2.Genome-wide association studies (GWAS)**

Genetic polymorphism across the genome associated with complex traits has been studied by geneticists to dissect their interactions ([Stranger, Stahl, & Raj, 2011](#)). Technology for long-read sequencing allows access to genomic areas formerly

obscure to human genetic variation studies ([Chaisson et al., 2015](#)). The result over the years has been a surge of complex genetic variation discovery. One analysis of 15 human genomes, for example, using long reads revealed about one hundred thousand common structural variants, in which half of them were not known before ([Audano et al., 2019](#)). Besides, several human reference genomes also contributed to the information about the existing human reference genome structure-allowing the discovery of over fifteen thousand sites in which either the sequence is incorrect, or which minor alleles are present (alleles that are <50%) in the population. With the reference genome's availability, resequencing the genome of individuals enables investigation of the association between sequence variation and traits.

Not surprisingly, structural variant sequence resolution has increased how well these variants can be genotyped even when using short-read sequencing, leading to a new identification of potential disease associations ([Huddleston et al., 2017](#)).

Despite the assumption that there are about 10 million common SNPs in the genome, arrays of SNPs covering the entire genome should be used ([LaFramboise, 2009](#)). However, no genotyping is required of each of these SNPs as human populations have migrated, and genetic variants have emerged, resulting in many variants being linked ([Dudley et al., 2012](#)). Thus, SNP genotyping can only be a small part of genetic variation in humans that the variants studied on SNP arrays are selected based on the structure of the linkage disequilibrium (LD) of the human genome; thus, it can reduce the costs of genotyping ([Das, Abecasis, & Browning, 2018](#)). LD is an attribute of SNPs outlining their degree of correlation ([Bush & Moore, 2012](#)). The correlation is associated with recombination events within a population. Finally, all alleles should theoretically be separate after a certain number of recombination events ([Bush & Moore, 2012](#)).  $R^2$ , a statistical correlation metric, allows the analysis of LD. Strong  $R^2$

values across multiple closely related SNPs indicate the provision of similar information. (Deigner & Kohl, 2018). Two genotyping array platforms are currently widely used, with both platforms providing over 99% genotyping accuracy (LaFramboise, 2009). As the first genotyping arrays, Affymetrix was designed to genotype 1494 SNPs to about more than one million SNPs as the current release (Lamy, Grove, & Wiuf, 2011). The other platform is Illumina that currently has up to more than five million genotyping capacity (Vergara et al., 2018), which is more from its first genotyping array of 100,000 SNPs (Human-1) (Lamy et al., 2011).

Genome-wide association studies (GWAS) propose a technique to perform genomic association studies with no previous hypothesis on which genes might be involved in a complex disease (Bush & Moore, 2012). GWAS studies a genome-wide record of genetic variants in various individuals to determine whether a variant is linked to a trait (Bush & Moore, 2012).

These approaches try to detect single nucleotide polymorphisms (SNPs), genetic variations in the genome that appear more frequently in people with a phenotype or certain disease than in people without that phenotype of interest or not suffering from the disease. Every test can view millions of SNPs simultaneously. Researchers use data to identify genes involved in a person's risk of suffering from a particular medical condition (R. Russo, Capasso, Paolucci, & Iolascon, 2011).

In a GWAS hypothesis test, the p-value suggests how likely a variant associated with a disease is due to an arbitrary chance. The GWAS community has long taken p-values seriously as protection against false-positive findings (Type 1 errors). The Bonferroni correction method is considered the most common method used for determining the correct p-value threshold by simply dividing the alpha value (most like

set to 0.05) by the number of tests being performed ([Kanai, Tanaka, & Okada, 2016](#)). Based on this correction method, a statistical significance limit has been suggested as the standard for common-variant GWAS with p-value  $5 \times 10^{-8}$  as a genome-wide significant threshold, which is achieved by dividing the alpha value ( $\alpha=0.05$ ) with one million independent variant tests ([Fadista, Manning, Florez, & Groop, 2016](#)). This correction is sensible because, after testing millions of variants in the genome, some random variants should produce randomly small p-values ([WTCCC, 2007](#)).

Since GWAS observes SNPs throughout the genome, it offers an excellent opportunity to investigate complex, frequent diseases involving multiple genetic mutations that lead individuals to risk. SNPs that cause several complex diseases, including CVD, Parkinson's disease, and Crohn's disease, have been discovered in this approach. It is expected that the future GWAS could detect more SNPs associated with common diseases, including variants interaction with lifestyle and response to specific medications ([Tam et al., 2019](#)).

GWAS test variants can rely on pairwise LD between an assayed SNP and a causal variant to detect an association between the assayed SNP and traits ([Das et al., 2018](#)). Nowadays, various LD structures have been mapped to be used in designing a genetic association study provided by the 1000 Genomes Project (1KGP) Phase 3, Haplotype Reference Consortium reference (HRC), and Trans-Omics for Precision Medicine (TOPMed) ([Kowalski et al., 2019](#)).

However, the causality between causal variants and an effect, such as a disease phenotype generated by GWAS, is only probability prediction and rarely a deterministic certainty. The degree of probabilistic causality is, in part, determined by the effect sizes of the SNPs identified ([Marian, 2009](#)).

Analysing a causality is challenging in complex diseases, which variants have small effect sizes and low penetrant; thus, do not show a clear co-segregation with the trait. These variants are often identified in the general population; therefore, it is even more difficult to measure it in an individual.

The effect of the genetic backgrounds (genetic modifier variants) and external exposures that are supposed to have a larger impact while the effect sizes including its causal genetic variants are very tiny, further compounds the determination of causality. Therefore, in some genetic backgrounds, a genetic mutation with a small to moderate impact size can be penetrating, and not in others.

Genetic variants in the human genome, with regards to their pathogenic role in human diseases, can be grouped into five groups ([Marian, 2014](#)):

1. **Variants causing diseases:** This group includes genetic variants for which the most robust evidence of causality is available, generally achieved in large families through genetic linkage analyses. As a rule, there is high penetrance of the variants, they have large effect sizes, and they are usually responsible for Mendelian inheritance patterns of single-gene diseases.
2. **Possible disease-causing variants:** Consist of variants characterised as those that demonstrate an association with the phenotype of interest, in association with robust mechanistic data incorporating them into the pathogenesis associated with the phenotype of interest. However, contrary to the first category, these variants cannot take advantage of data robust in human molecular genetics, including linkage evidence within families of large size.
3. **Disease-associated variants:** classified as disease variants without being implicated in the pathogenesis of the disease based on genetic or mechanistic

data. They may simply be in association with the actual causal variants in the same LD. Generally, these variants are standard variants with limited influential variables. GWAS place the vast majority of variants associated with the phenotype in this group. Causal involvement of "disease-associated variants" in the pathogenesis of the phenotype of interest, detected via GWAS or otherwise, is usually not elucidated, although some may be functional or located in genes involved in the phenotype's pathogenesis. In cases of LD with the actual pathogenic variants, it is expected that many are simply DNA markers. The power of the evidence of the association observed with the phenotype, its replication in independent sample populations, gene dosage effects, and biological plausibility must be considered in determining the causality of these variants. At last, to determine the possible causal role of this group of variants in the pathogenesis, including its associated phenotype, experimentation is required.

**4. Functional variants associated with unspecified clinical consequences:**

This group of variants is characterised as variants exhibiting effects that are biological yet not associated with clinical phenotypes, such as affecting their respective genes' mRNAs and protein levels.

- 5. Variants of unclear biological or clinical significance:** this group includes variants that have not yet been characterised and account for the large majority of each human genome's approximately 4 million sequence variants. It is not clear whether they have a biological role and are not linked to a phenotype. Characterising these variants can obviously motivate their classification. Although the initial GWAS gives statistical evidence that a region is likely to have a causal variant, there is a need for additional statistical methods to

distinguish between possible functional variants and variants that are purely associated with functional variants.

A disadvantage of the lead SNP is that it is not the causal version that has a fair chance. This may occur as GWAS microarrays are based on tag-SNPs, with the tag-SNP simply associated with the unmeasured causal SNP (Schaid, Chen, & Larson, 2018). In addition, even though the causal SNP is calculated or imputed, there is a fair probability that the causal SNP statistical association with the trait is not the most relevant association of all associated SNPs, unless the statistical power is strong (Liu et al., 2016; Schaid et al., 2018).

Fine mapping is the process of analysing a region associated with a trait from a GWAS to determine the particular genetic variants most likely to have a causal influence on the trait under study. It attempts to identify the genetic variant(s) causing complex traits when an association of a genomic region with a trait is demonstrated, and there is at least one causal variant. After at least one SNP closely correlated with a trait is identified by an initial GWAS (e.g.,  $p$ -value  $<5 \times 10^{-8}$ ), fine-mapping usually continues with the general steps illustrated in Figure 1.4.

The overall approach is to use the GWAS list of trait-related SNPs to classify regions of interest. For their LD structure and for genes in the region known to be mapped, each region is then visually examined. Since fine-mapping one causal variant at a time is easier, each region is divided into sub-regions with roughly independent effects on the phenotype. Then, a method for fine mapping is chosen. The selected SNPs will be evaluated further for their likely function based on an openly available genomic annotation to prioritise expensive and time-intensive lab-based functional studies. The

result can then be used for follow-up research, such as large-scale specific candidate replication studies or laboratory functional studies.

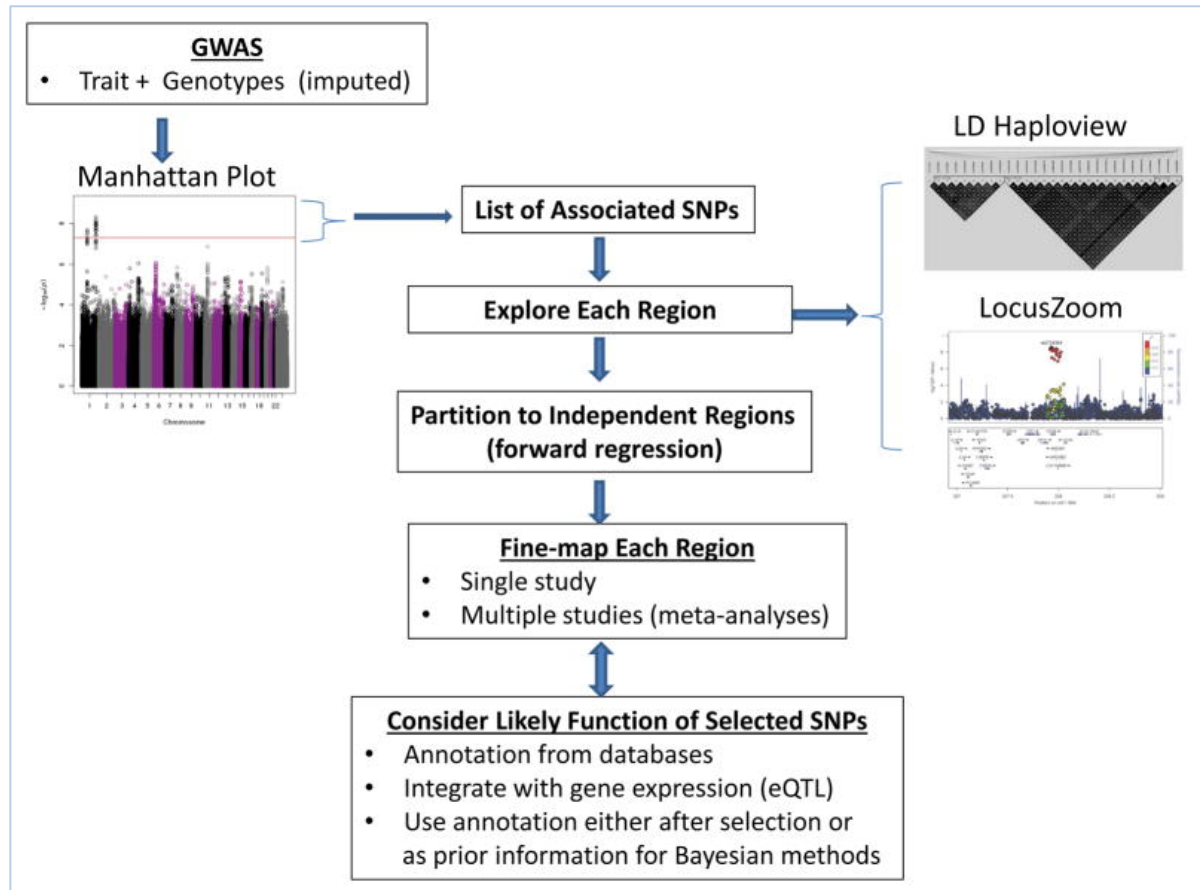


Figure 1.3 Fine-mapping workflow.

Human health and observational disease studies (basic, clinical, and epidemiological) are prone to methodological issues that make causal inferences problematic, such as selection bias and confounding. Gene-disease associations are much the same, as observational designs are widely used to examine them. However, relative to environmental exposure studies, genetic differences are less likely to influence subject selection (e.g., cases and controls in a case-control study) in genetic studies. Excluding linkage disequilibrium (i.e., attribution of a genetic influence to a particular gene rather than an adjacent one) and population stratification, confounding is often

less probable (when cases and controls are drawn from different ethnic populations). There is currently some empirical evidence indicating that associations of gene-disease are less vulnerable to misunderstanding (e.g., socioeconomic status) than correlations between genes and variables of the environment and lifestyle. In interpreting the causal significance of gene-disease associations, there are several well-known methodological challenges; these include linkage disequilibrium, epistasis, and gene-environment interactions (GEI).

In 1965, to help assess whether observed epidemiological associations are causal, Sir Austin Bradford Hill released nine "viewpoints." Ever since, in epidemiological research, the 'Bradford Hill Criterion' has become the most widely cited paradigm for causal inference. When considering data obtained utilising modern molecular methods, such as epigenetics and causation biomarkers the Hill's criteria can be adapted ([Fedak, Bernal, Capshaw, & Gross, 2015](#)).

In term of the causation, a rich body of information exists on the evaluation of causal associations involving environmental and personal causes of disease; it involves Austin Bradford Hill's causal criteria developed and more lately applied Hill's method is based on nine criteria: 1) association strength; 2) consistency; 3) association specificity; 4) temporality; 5) biological gradient ( dosage-effect) relationship; 6) biological plausibility; 7) coherence; 8) Experimental evidence (e.g., replicability in animal models); and 9) Analogy ([Fedak et al., 2015](#)).

Statistical importance was not identified but discussed separately by Hill. One of Hill's main problems is that it considers a single causal factor at a time and is not designed to discuss complex relationships and interactions as found in contemporary molecular medicine and genomics, which deal with mediation and not just exposures that have

a direct effect (Geneletti, Gallo, Porta, Khoury, & Vineis, 2013). However, even complicated problems can sometimes be broken down into simpler components, and Hill's criteria can be applied fruitfully in such cases. This is a crucial inspiration behind the work at hand. Genetic science, however, has fundamental distinctions from a clinical and epidemiological study. For illustration, confounding in genetics may be the outcome of events at other levels that may not be clearly explained, like overdominance, allelic heterogeneity, haplotype blocks, and epistasis (Glazier, Nadeau, & Aitman, 2002). In genomic studies, selection bias is more easily observable since the null hypothesis was used described by Hardy-Weinberg equilibrium (HWE); i.e., in the population, an independent assortment of alleles is predicted, although a similar rationale cannot be extended to exposures to everyday life (Glazier et al., 2002). The equilibrium of Hardy-Weinberg is based on population genetics assumptions related to lack of selection, inbreeding, migration; departure from HWE will therefore raise the possibility of gross biases (for example genotyping errors or selection bias) (Glazier et al., 2002). There are, however, genetic rules on causality and they actually have interesting parallels to Hill's criteria (Geneletti, Gallo, Porta, Khoury, & Vineis, 2013: (a) relation to a particular area of the human genome ( $LOD > 3$ ); (b) one or more independent mutations that are associated with disease status in affected families (specificity, the strength of association); (c) defects leading to macro protein changes (specificity, coherence); (c) defects leading to macro protein changes (specificity, coherence); Putative mutations never found in a control population sample (specificity); or (e) existence of certain other biological evidence lines (including expression, knockout data, etc.). Criteria (a), (b) and (c) apply to knowledge of the history. However, criterion (e) in particular, supports the causal relation by conferring coherence with prior knowledge (Glazier et al., 2002).

### 1.2.1. Missing heritability

Over a decade, many genetic loci in humans have been mapped, leading to early success in dissecting the genetic architecture of diseases with complex traits. However, the results usually explain only a moderate proportion of inherited disease genes in complex traits (Bansal, Libiger, Torkamani, & Schork, 2010), although GWAS have been able to detect common variants a decade. This issue continues to be a major challenge in understanding the role of genetic risk factors in health-related diseases, referred to as “missing heritability” (Li & Keating, 2014).

A predominant opinion that explains missing heritability is the limitation of heritability genes with moderate effects that cause common disease (S. Lee, Abecasis, Boehnke, & Lin, 2014); therefore, it is likely that genetic inheritance occurs in rare variants of greater penetrance that have emerged over the years (Schork, Murray, Frazer, & Topol, 2009). When identifying associations, looking for variants with >5% minor allele frequency (MAF) leads to underpowered associations. Moreover, lower frequency variants (MAF <5%), as described in (Auer & Lettre, 2015), are rarely detected by genotyping arrays. Another way to address the missing heritability issue in the GWAS setting is by enlarging the sample size or joint analysis of correlated phenotypes to improve the statistical power.

### 1.2.2. Meta-analysis and Trans-ethnic meta-analysis of multiple GWAS

GWAS are usually performed in various research teams at various sites. Meanwhile, various studies use specific genotyping arrays consisting of multiple sets of variants with different genotyping platforms. For example, the Illumina 660k array only used 20% SNPs of the Affymetrix 6.0 array (Das et al., 2018). Imputation of genotypes can produce a set of common variants that can be analysed in all studies to increase

performance. It estimates genotypes unobserved from genome-wide constructors to improve genome coverage and performance for GWAS ([Vergara et al., 2018](#)).

Meta-analysis provides the statistical method for integrating data from multiple trials of GWAS. Meta-analysis allows this common effect to be identified when the treatment effect (or effect size) is coherent across studies. ([Haidich, 2010](#)). Genotype imputation plays a key role in GWAS meta-analyses to increase statistical power and facilitate fine mapping of causal variants ([Das et al., 2018](#)). It estimates genotypes unobserved from sequencing to increase the performance of GWAS by completing the genome coverage ([Vergara et al., 2018](#)). A sequence of individuals consists of variants that previously observed can be used as a reference panel to inexpensively impute whole-genome sequence data ([Das et al., 2018](#)).

The use of international collaboration, such as the 1KGP, allows the imputation to estimate the genetic information at unobserved variants by providing the most comprehensive map of genetic variation data across different ethnic groups ([Stranger et al., 2011](#)). It was first introduced in 2008, facilitating genetic variation in human genetics by sequencing about 2,500 genomes from about 25 populations worldwide. The 1KGP provided support for SNP queries and deletions/insertions from the phase-1 integrated release (March 2012), including more than 1,000 selections in 14 populations. 1KGP was completed in 2015, sequenced individuals from 26 populations in Africa, East Asia, Europe, South Asia, and America, resulting in 2,054 genomes were recreated to the end of the project. ([Panagiotara et al., 2019](#)).

Reference panels for the European ancestry population have been established for genetic imputation, which was successfully used in two meta-analyses GWAS in 2008 ([Barclay et al., 2008](#)). New association signals were detected in these two studies that

used different arrays and imputations from reference panels, while the signals cannot be detected in the individual study (Barclay et al., 2008).

GWAS also showed that common SNPs promote susceptibility to complex traits are distributed across different ethnic groups (Li & Keating, 2014; Marigorta & Navarro, 2013). Replication of discovery association signals in independent population and population-wide meta-analyses can be used to improve performance. This replication is in line with a paradigm of sharing causal variants across populations. The transethnic meta-analysis increases the sample size to boost the power to detect new loci (Liu et al., 2016).

Trans-ethnic studies have been invaluable in studying the genetic architecture of complex diseases, such as type 1 diabetes (T1D) and T2D (Karaderi et al., 2015), Parkinson's disease (Pandey, Singh, & Verma, 2013), coronary artery disease (Teslovich, T. Musunuru, K. Smith, 2010), different types of cancer (Schifano, Li, Christiani, & Lin, 2013; Solovieff, Cotsapas, Lee, Purcell, & Smoller, 2013), height (Grundy, Brewer, et al., 2004), body mass index (Karaderi et al., 2015), and lipid levels (Tada et al., 2014).

### 1.2.1. Multi-phenotype genome-wide association studies

#### 1.2.1.1. Multi-phenotype effects

Usually, the traditional GWAS are conducted on single phenotype analysis (Mägi et al., 2017), while many previously discovered GWAS loci show overlap to multiple phenotypes (Marullo et al., 2014). For example, Loci associated with glycemic traits are also intersecting with those for lipids, as shown in Figure 1.4.

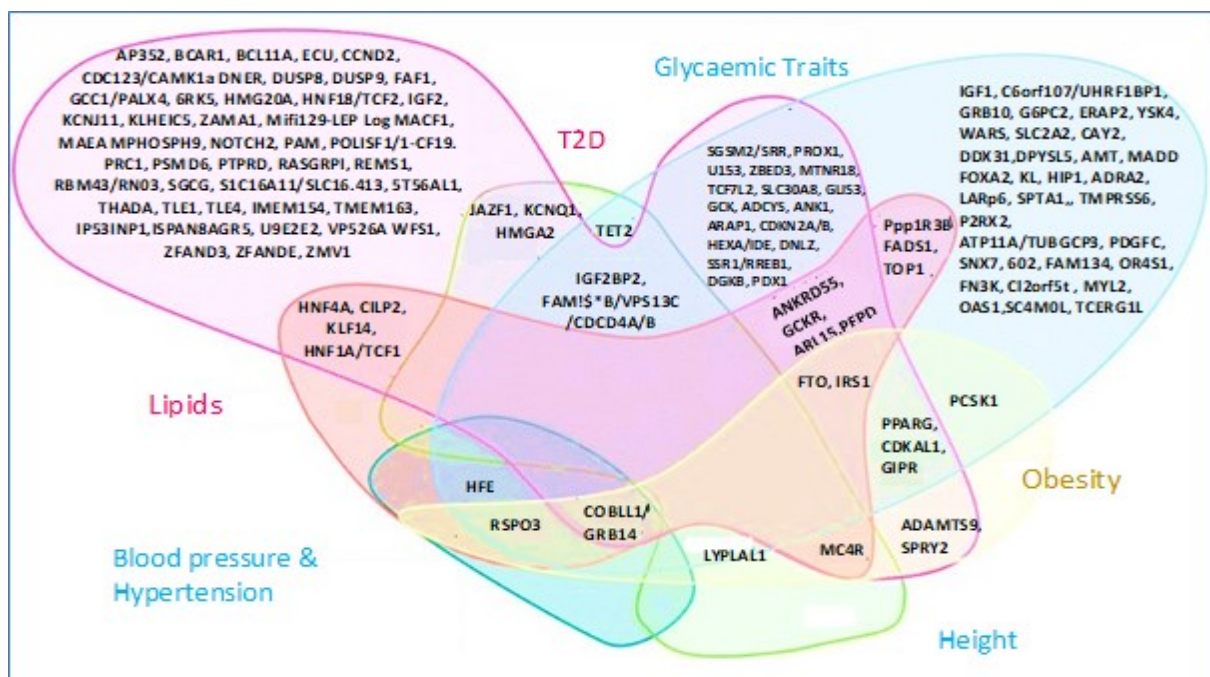


Figure 1.4 The intersection of established GWAS loci with cardiometabolic traits. (source:(Marullo et al., 2014)).

The combined evaluation of the single trait GWAS for cardiometabolic phenotypes has revealed many loci that overlap, resulting in multi-phenotype effect or pleiotropy when each variant affects several phenotypes (Shriner, 2012). Pleiotropy itself has a biological basis in which a single or several variants of a gene, or a location in general, influence several characteristics (Solovieff et al., 2013). However, effects similar to pleiotropy can also be caused by multiple scenarios, e.g., closely related genes,

ascertainment bias causing spurious pleiotropy, and mediated pleiotropy (Solovieff et al., 2013).

In mediated pleiotropy, the causal variant of a phenotype (P1) indirectly associated with another phenotype (P2) due to P1 is causally related to P2. In comparison, spurious pleiotropy is usually triggered due to ascertainment bias when causal variants in different genes are associated with multiple phenotypes, as seen in Figure 1.5 (Solovieff et al., 2013).

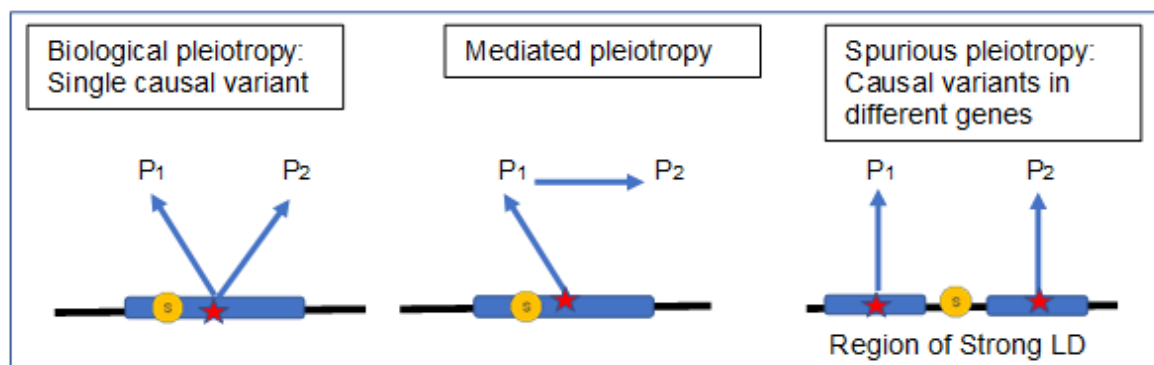


Figure 1.5: Several possible pleiotropy scenarios (source: Solovieff et al., 2013).

Thus, it is easier to identify these signals as the collaborative analysis of multiple traits provides additional information from the relationship between the traits used (Wang, Sha, & Zhang, 2016). Joint analysis of correlated phenotypes has increased the power to detect genetic variants and the ability to detect pleiotropy phenomenon in complex traits over-analysis of traits one by one, suggesting what is called MPA (Wang et al., 2016).

MPA offered statistical and biological benefits compared to the univariate methods. By analysing multiple phenotypes at once, correction for multiple traits can be avoided; thus, it can increase the power of locus discovery. It also takes advantage of the correlations between phenotypes that can also help boost the power of associations.

However, high correlations between phenotypes may suffer from multicollinearity problems; thus, we should carefully select phenotypes with  $|r^2| < 0.8$  in MPA.

GWAS methodology using MPA has been demonstrated to improve the ability to discover loci (Galesloot, Van Steen, Kiemeney, Janss, & Vermeulen, 2014; Mägi et al., 2017; O'Reilly et al., 2012). Multi-phenotype GWAS (MP-GWAS) conduct individual association tests on each variant with multiple phenotypes at once, optimising resource usage and time (Mägi et al., 2017).

One of the proposed MP-GWAS approaches uses the reverse regression model, as opposed to the commonly used single-phenotype GWAS. This methodology uses the phenotypes as the predictors and the genotypes as the regression outcome (O'Reilly et al., 2012). Both binary and continuous traits are accepted in this approach, as the reverse regression model does not need any phenotype distribution assumption (O'Reilly et al., 2012). The reverse regression model can be described into this formula for the genotype,  $G_i$ :

$$G_i = \alpha + \beta_1 y_1 + \beta_2 y_2 + \dots + \beta_i y_i + \epsilon_i$$

The intercept is  $\alpha$ , the coefficient of regression is  $\beta_i$  for phenotype  $i$ ,  $G_i$  is the  $i$ th genotype,  $y_i$  is the  $i$ th phenotype, and  $\epsilon_i$  indicates an error.

### 1.3.Missing data

After more than ten years of accumulated efforts, GWAS have discovered many loci associated with interest traits; however, the power to discover these loci relies on the sample size. Meanwhile, technical limitations or difficulty in collecting the data regarding many phenotypes are missing. Consequently, as the number of phenotypes

increases, the chance of missing data will also increase gradually in the MP-GWAS setting. Complete cases (CC) analysis using the completed data without missing data means removing the individuals who have missing data, even if data only have the missing phenotype. Thus, it can drastically reduce the sample size. However, most studies use this simple CC analysis, for example, studies that removed between 3%-31% of samples (Y. H. Choi, Chowdhury, & Swaminathan, 2014; Cui, Sha, Zhang, & Chen, 2009; Park, Lee, & Kim, 2011; Piccolo et al., 2009; Scutari, Howell, Balding, & Mackay, 2014).

The vast majority of high-dimensional datasets include some degree of missing data. This missing data can be for various reasons, e.g., untraceable, left blank, unreadable. Although missing data is usually ignored in many analyses, it can reduce the statistical power due to shrinking the sample size (Rubin, 1976). Therefore, instead of removing the missingness and only using the completed data, it is essential to adequately characterise and manage missingness (Rubin, 1976; Schafer & Graham, 2002).

While having an advantage for its simplicity, CC analysis may be useful for univariate analysis with only a few missing values. However, it is not the case with MPA. Analysing multiple phenotypes at once will also increase the proportion of missing data, leading to the standardly implemented CC analysis's inefficiency, as illustrated in Figure 1.6.

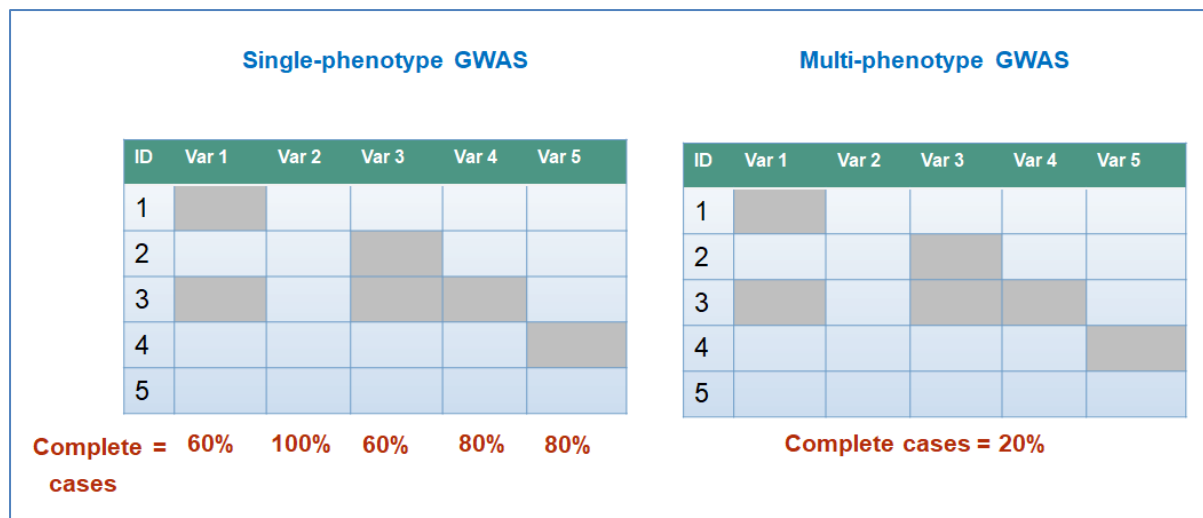


Figure 1.6 Illustration of the proportion of missing data increase in MP-GWAS as compared with single-phenotype GWAS.

When analysing multiple phenotypes, each with some missing values, the complete cases remove all the individuals with a missing value even in only one phenotype. Therefore, many observed values are not analysed because the standard is to exclude individuals with missing data. Likewise, genotype data, defined as a fraction of the genome's total set of testable polymorphisms, pose a similar problem. However, the imputation of genotype data is commonly implemented, whereas the imputation of phenotype data is not.

Therefore, the imputation of missing phenotypes provides an important approach for MP analysis not to lose the power to detect new associations (Dahl et al., 2016).

### 1.3.1. Missing data patterns and mechanisms

It is useful for theoretical and practical reasons to identify missingness patterns and mechanisms before choosing the best imputation approach. Missing data patterns can be divided into (van Buuren, 2018):

1. *Univariate and multivariate*. The univariate is a missing data pattern if only one variable contains missingness, whereas in the multivariate case, the data has more than one variable which is missing.

2. *Monotone and non-monotone*.

The data can be illustrated by the  $n \times p$  matrix  $Y$ . If data is missing,  $Y$  is observed partially. The notation  $Y_j$  is the  $j^{\text{th}}$  column in  $Y$ , and  $Y_{-j}$  denotes the addition of  $Y_j$ . A missing data pattern is monotonous when the variables  $Y_j$  satisfy the requirement that if  $Y_j$  is absent, variables  $Y_k$  are unobserved when  $k > j$ . This pattern happens, e.g., in length studies with failure. The pattern is *general* or *non-monotone* if it is not monotonous ([van Buuren, 2018](#)).

3. *Connected and unconnected* missing data.

A missingness pattern is considered *connected* if a data point observed from another observed data point is accessible by a series of vertical or horizontal moves (such as in chess) ([van Buuren, 2018](#)).

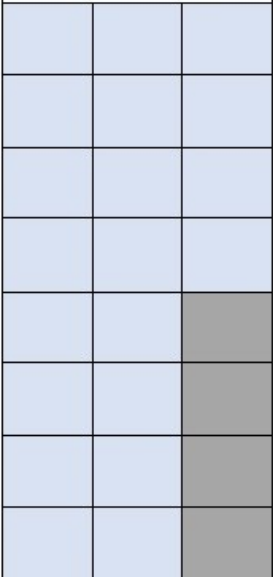
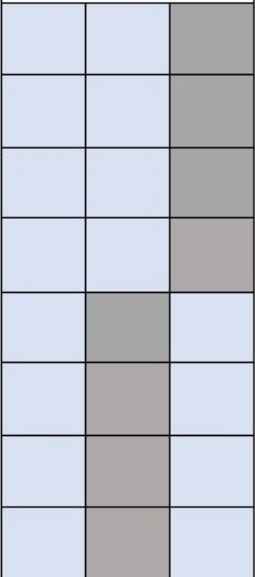
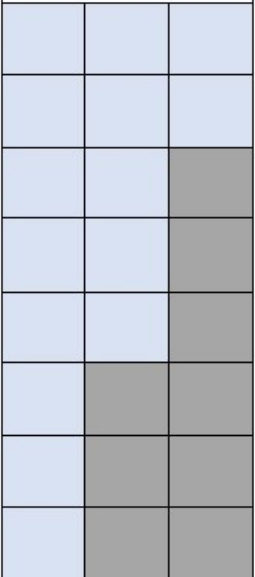
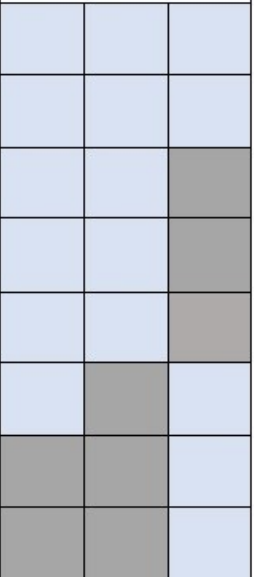
| Univariate                                                                          | Multi-variate                                                                       | Monotone                                                                             | Non-Monotone                                                                          |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
|  |  |  |  |

*Figure 1.7 An example of missing data patterns in multivariate data. Blue indicates observed data, and grey indicates missing data (van Buuren, 2018).*

Monotonous patterns may be created by the setting of longitudinal studies. For monotonous patterns, variables can be arranged according to the % of missing data. If univariate data is missing, a specific monotonous pattern is created. Major mathematical efficiencies are possible when the data is monotonous.

It is necessary to figure out the reason why the missingness is generated to predict it, which can be differentiated based on their missing data mechanisms from the naïve assumption into the special case: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (Rubin, 1976; Schafer & Graham, 2002). Missing values in a data set have an MCAR pattern when the missingness is not caused by the observable and the non-observable parameters of interest and are completely random (Rubin, 1976; Schafer & Graham, 2002). If the values are MCAR, the sampling is probably unbiased and still representative of the population, but the data are rarely MCAR.

When missingness can be completely considered by the observed variables, the missing data will turn into MAR (Bhaskaran & Smeeth, 2014). As MAR is a hypothesis that cannot be confirmed from a statistical point of view, it is necessary to rely on its appropriateness (R. J. A. Little & Rubin, 2002). For example, male participants tend to be reluctant to complete a depression study, albeit this is not related to their level of depression. Thus, it creates a missing data of variables observed based on the difference of sex. These data may still cause parameter distortions in the analysis due to the conditional emptiness of the variables, depending on the analysis method. MAR serves asymptotically unbiased estimates when the parameter is assessed with a Full Information Maximum Likelihood (FIML). (R. J. A. Little & Rubin, 2002).

MNAR is the special case of MAR when missingness relies on the unobserved data (Schafer & Graham, 2002). If the data, which are MNAR, are not explicitly considered and modelled, this leads to the study's bias (Schafer & Graham, 2002). Think of a clinical trial in which the participants' blood pressure is measured twice, but only the >140 mmHg is recorded by the second measurement (Schafer & Graham, 2002). This missing data will turn into MNAR. Suppose the cause of the error is unidentified. In that case, it is important to form rational estimates of the error's origin and the distribution of the error, although these estimates are not testable (Schafer & Graham, 2002).

## 2. Review of the literature

*This chapter reviews the relevant literature on the single-phenotype and multi-phenotype approaches to dissect the genetic architecture of cardiometabolic traits. I am focusing on the previous studies to serve the insight into the topic carried out in this thesis and offer a picture of the existing knowledge in this area. This literature review provides a theoretical context and a methodological basis to clarify the importance of the work carried out in this report,*

As the thesis already focused on a specific field; therefore, the selected literature topics during this thesis can be found in [Figure 2.1](#).

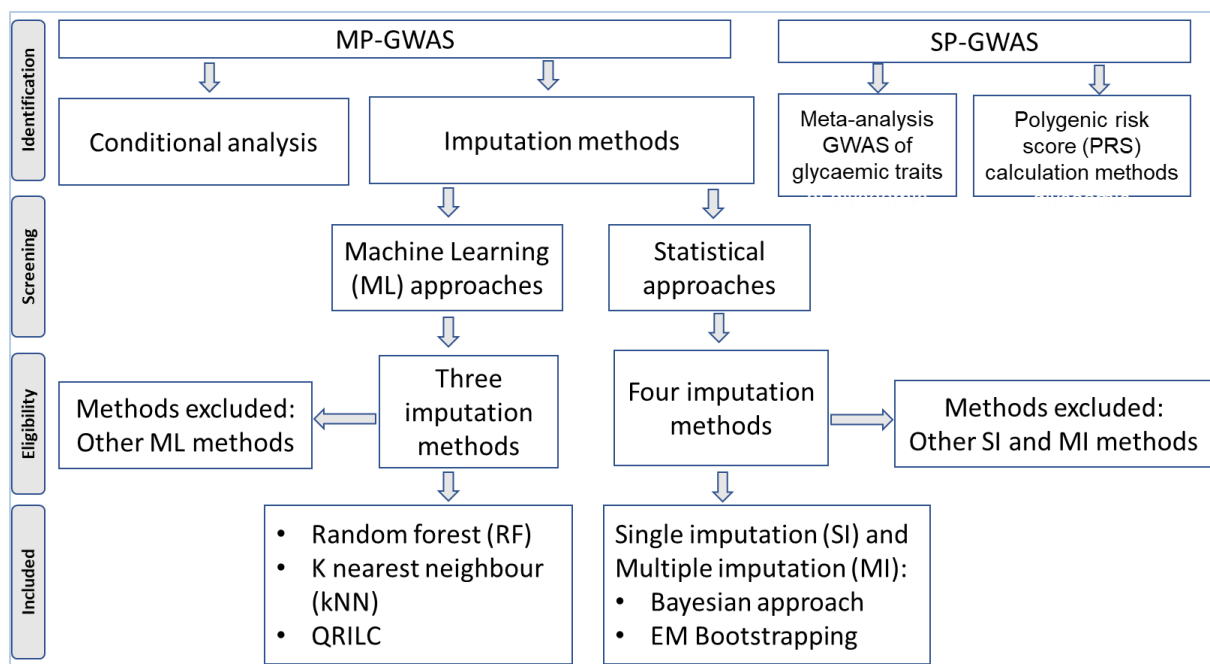


Figure 2.1 The workflow of the selected literature topics used in the thesis.

The main literature used in this thesis is described in [Table 2.1](#) and [Table 2.2](#).

Table 2.1 The key literature used on the method development and implementation of SP-GWAS and MP-GWAS on cardiometabolic traits.

| Year | Author         | Title                                                                                                                                | Comments                                                                                                                                                                                                                                                                                                                            |
|------|----------------|--------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2012 | O'Reilly et.al | MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS                                                         | Joint analysis of multiple-phenotype in the MP-GWAS framework using reverse-regression approach was proposed in this study yet did not accommodate directly typed and imputed SNPs (under an additive dosage model); moreover, association summary statistics cannot be aggregated across GWAS through fixed-effects meta-analysis. |
| 2017 | Mägi et.al     | SCOPA and META-SCOPA: software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes | Join-analysis of multiple phenotypes in the MP-GWAS framework, and summary statistics through meta-analysis was proposed in this study, yet missing multi-phenotype data was not addressed in this framework.                                                                                                                       |
| 2017 | Jakobsen et.al | When and how should multiple imputation be used for handling missing data                                                            | Multiple imputation (MI) has been addressed in this study as a better imputation method compared to single-imputation (SI), yet MI demand computational time in high-dimensional data                                                                                                                                               |
| 2018 | Wright et.al   | ranger: A Fast Implementation of Random Forests                                                                                      | This study claimed to develop the fastest implementation of random forest (RF), yet it focused on regression and survival analysis without imputation.                                                                                                                                                                              |

|             |             |                                                                                 |                                                                                                                                                 |
|-------------|-------------|---------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
|             |             | for High Dimensional Data in C++ and R                                          |                                                                                                                                                 |
| <b>2012</b> | Zhang et.al | Nearest neighbour selection for iteratively kNN imputation                      | This study claimed that kNN is able to deal with missing values in heterogeneous datasets while maintaining the rational structure of the data. |
| <b>2018</b> | R Wei et.al | Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data | This study found that RF performed the best for MCAR/MAR missing mechanisms, and QRILC was the most suitable one for left-censored MNAR.        |

Table 2.2. The key literature used the analysis and implementation of SP-GWAS and MP-GWAS on cardiometabolic traits.

| Year | Author           | Title                                                                                                                                      | Comments                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|------|------------------|--------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2010 | Prokopenko et.al | Variants in the melatonin receptor 1B gene (MTNR1B) influence fasting glucose levels                                                       | Meta-analysis GWAS of glycaemic traits was performed in this study which involved a total of 36,610 individuals of European descent, yet only 1.5% of variance explained was found for fasting glucose (FG).                                                                                                                                                                                                                                                |
| 2012 | Scott et.al      | Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways | Meta-analyses GWAS of glycaemic traits was performed in this study involving a total of 133,010 individuals of European ancestry without diabetes, including individuals newly genotyped using the Metabochip. They identified the variance, which explained 4.8% by FG and 1.2% by FI.                                                                                                                                                                     |
| 2016 | Willer et.al     | Discovery and refinement of loci associated with lipid levels                                                                              | Trans-ethnic meta-analyses GWAS of lipid traits was performed in this study involving a total of 188,577 individuals, identifying multi-phenotype effects where the detected loci were associated with multiple cardiometabolic traits. Blood lipid levels were found to be associated with cardiovascular and metabolic traits, including coronary artery disease, type 2 diabetes (T2D), blood pressure, waist-hip ratio (WHR), and body mass index (BMI) |

## 2.1. Polygenic risk score (PRS) calculation

Although SP-GWAS has identified genetic variants associated with cardiometabolic traits, each of these variants only has a small effect (S. J. C. Craig et al., 2019).

The polygenic risk scores (PRS) can explain part of the total variance in liability (Vega & Bustamante, 2018). When PRS provides greater power of prediction by combining more variants with small effects rather than individual variants that reach the statistically significant genome-wide threshold, it implies that the traits of interest are caused by a combination of more variants than those variants that hit the threshold. Thus, PRS can predict traits caused by multiple genetic variants by summing the genetic effects of these variants.

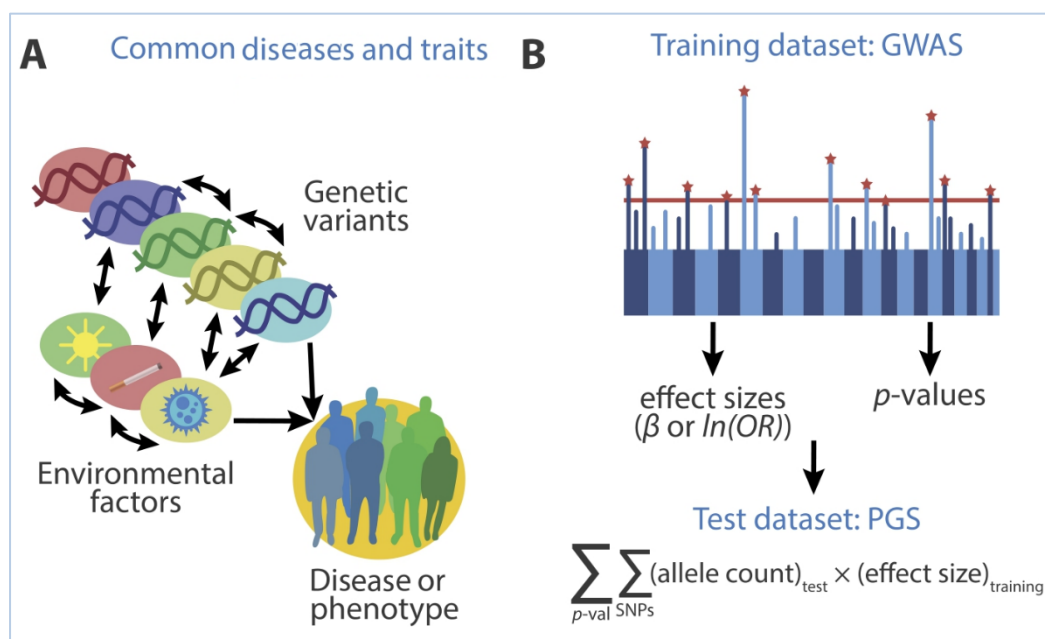


**Figure 2.2** (A) The interactions and additive effects of multiple genetic variants and environmental factors affects common diseases and traits. (b) Polygenic risk scores (PRS) are usually computed by summing independent (i.e., uncorrelated) single nucleotide polymorphisms (SNPs) weighted allele counts observed in a genome-wide association analysis (GWAS) to be associated with a phenotype or disease. The weights are the linear regression effect sizes for quantitative characteristics, i.e., the regression coefficients ( $\beta$ ). The weights are the normal logarithm of the odds ratio for case/control phenotypes,  $\ln(OR)$ .

However, PRS can not currently be used to report about the predisposition of an individual or risk of a disease in a person. One potential future application of PRS, however, would be, for example, to give more frequent screening to people with a very high PRS for a particular condition. Besides, PRS may be useful in treatment prediction models incorporating genetic and non-genetic data.

PRS is usually estimated using the weighted or unweighted sums of the disease-causing alleles as risk factors for each individual (Vega & Bustamante, 2018). It can use simple counts (unweighted) or employ the GWAS coefficients derived from discovery as a basis for weighting each SNP allele to compute an overall risk value for each individual in an independent population. The unweighted PRS treats that the traits are affected by each SNP equally, which is unlikely as some SNPs have a larger influence on trait heritability. Whereas each SNP in the weighted sum method is calculated better by its predicted disease effect size, it, therefore, accounts for its specific contribution to disease risk (Vega & Bustamante, 2018).

PRS is usually derived from summary statistics of GWAS meta-analyses. Thus, PRS represents individual genetic predictions of phenotypes, which were calculated for each cohort contributing to the meta-analysis by computing the sum of risk alleles from the discovered loci corresponding to each individual's phenotype, weighted by the effect size estimate from the meta-analysis GWAS performed on that phenotype.

After the calculation of PRS that represents at least some of the variance of a phenotype, it can be expected that it efficiently integrates a large part of the genetic variants associated with that phenotype. There are several possibilities to use PRS: as a baseline to test whether heritability predictions can be skewed; as an indicator of the genetic intersection of traits (genetic correlation) to show common genetic bases

for clusters of cardiometabolic traits; serving as a tool for assessing cluster differences in a trait as a function of natural selection, such as size, or for studying changes in a trait over time that indicate a soft selective sweep (such as for intelligence, for which the frequency changes appear insufficient to detect on each hit but not on the entire PRS) ([B. Zhao & Zou, 2018](#)).

Two important methodological aspects for the development of PRS are a selection of SNP and measurement of weight ([Dudbridge, 2013](#)).

#### 1. Selection of SNPs.

The choice of SNP candidates is important since they constitute the building blocks of the PRS. Without filtering, its basic technique is to maintain all SNPs. For genetically underexplored diseases with several low to medium SNP effects, this may be successful. However, by adding many non-informative or far too poorly associated SNPs, the PRS's output can suffer. Optionally, a subset of SNPs based on predefined parameters should be used (e.g., The ones that pass an arbitrary p-value threshold in the GWAS results). However, this ad-hoc cut-off can lose certain informative SNPs with a tiny effect size. Therefore, at different thresholds, the PRS-disease association may vary significantly ([Dudbridge, 2013](#)).

There are two approaches to select SNPs. The first approach is using clumping/pruning and thresholding, where a reduced collection of SNPs is recognised by pruning on LD and reserve as evidence of association with the phenotype being examined (clumping). The downside of this approach; however, is that certain variants used may be false positives, and selection bias could be subject to the raw estimates of impact sizes from them. In addition, it involves checking over a number of thresholds

for p-values, which are sometimes randomly picked. A positive bias can also be subject to the prediction error calculated from the optimized threshold.

A further issue is the redundancy of variants' informativity, occurring in the case of linkage disequilibrium (LD) where nearby SNPs have highly similar associations. SNP filtering strategies such as LD pruning accompanied by p-value thresholding will resolve this. Random pruning or clumping eliminates the bulk of the SNPs in an LD block. PRS, as implemented in PRSice (Euesden, Lewis, & O'Reilly, 2015) and PLINK (Purcell et al., 2007), are then can be used, to sum up, all variants reach a p-value threshold. It enables the selection of SNP under a set of p-value thresholds that give a more specific option. One concern is that, based on threshold selection criteria, overfitting problems can occur (Euesden et al., 2015).

The second approach is by specifically modelling the association structure between SNPs to predict on a genome-wide scale without attempting to select a limited subset of SNPs. The Bayesian LDpred (Vilhjálmsdóttir et al., 2015) approach is the most commonly used implementation of this last approach.

## 2. SNP-calculation for weight.

The selection of SNP weights is another key factor for PRS results. GWAS-derived statistics or risk assessments (i.e., ORs) are widely used as PRS weights on an independent sample is an expansion of this method that has been promising.

PRS is typically constructed from univariate regression tests as a weighted sum of allele counts, the weights being given by log odds ratios or linear regression coefficients. A simplified solution (Dudbridge, 2013) is also feasible by measuring

each marker by +1 or -1 (S. H. Lee et al., 2013). It is normal to ask whether we could strengthen the process of weighting to increase predictive power.

The polygenic hazard scores (PHS) (Tan et al., 2019) are derived as a weighted total of SNPs, but the weight of each SNP is represented in this case by an HR calculated using a survival model where SNPs are included as predictors. In a PRS discovery sample, GWAS genotypes may not be part of those that lead to the attenuated output of the PRS in the validation or application series. LD and regression to the mean or "winner's curse" are other variables that affect results. Adjusting SNP weights will help tackle these issues.

The two major optimized SNP re-weighting approaches are 1) inference based on Bayesian inference and 2) inference based on frequentist inference. LDpred (Vilhjálmsen et al., 2015) uses the known LD structure without requiring raw genotype data or p-value thresholds as a precursor to new SNP weights. LDpred achieved better trait prediction accuracy when applied to simulation data compared to conventional methods without LD information (Vilhjálmsen et al., 2015)]. AnnoPred (Hu et al., 2017) further strengthened LDpred by considering that the biological identity of each SNP leads to heritability unique to the SNP.

A study found that the Bayes approach improves the genomic risk prediction recovered the underlying effect sizes by used them as weights to construct PRS (So & Sham, 2017).

## **2.2. Correlated phenotype analysis**

Many traditional SP-GWAS in cardiometabolic traits has discovered that some loci identified in this method associated with more than one phenotype, a phenomenon known as pleiotropy when the phenotypes are correlated (Marullo et al., 2014).

There are several pleiotropy association methods have been suggested to confirm the GWAS finding. Phenome-wide association studies (PheWAS) is a reciprocal method to SP-GWAS by implementing a reverse directional approach using a genotype to multi-phenotype perspective. PheWAS focus on loci that SP-GWAS already detected. PheWAS assess whether a locus is associated with a wide range of multiple phenotypes, the phenome. The first study that implemented this approach was carried out in 2010, demonstrated its ability to uncover pleiotropy, identifying the phenotypes associated with the target loci ([Denny et al., 2010](#)).

However, PheWAS are hypotheses generating method that have the limitation of GWAS, which needs multiple correction testing ([Hebbring, 2014](#)). A Bonferroni correction may not be an appropriate option to correct the multiple testing when the phenotypes are correlated with each other. Like GWAS, subsequent analysis targeting the specific phenotype still need to be carried out after the PheWAS screening to validate the findings ([Hebbring, 2014](#)).

Another method, Multi-Trait Analysis of GWAS (MTAG), proposed incorporating summary statistics from GWAS to enable joint-analysis of related multiple phenotypes ([Turley et al., 2018](#)). Thus, it can be applied directly without accessing individual-level data of the GWAS for each phenotype. Moreover, this approach applied a regression score of bivariate linkage disequilibrium to consider overlapping cohorts between the GWAS analysis ([Turley et al., 2018](#)).

The inputs in MTAG are GWAS summary statistics for multiple phenotypes with potential each phenotype is analysed in SP-GWAS meta-analysis. This summary statistics give a set of association combinations among all the phenotypes that constitute the best of the information combined. The P-values generated in this result

can be treated in the same way as GWAS P-values to precedence identified loci for further analyses, such as PRS calculation or annotation ([Turley et al., 2017](#)).

Instead of relying on the summary statistics of SP-GWAS, joint analysis of correlated phenotypes into one analysis has been proposed to increase the power to detect genetic variants and the ability to detect pleiotropy phenomenon in complex traits over an analysis of traits one by one, suggesting what is called multi-phenotype analysis (MPA) ([Wang et al., 2016](#)).

MPA offered statistical and biological benefits compared to the univariate methods. By analysing multiple phenotypes at once, correction for multiple traits can be avoided; thus, it can increase the power of locus discovery. It also takes advantage of the correlations between phenotypes that can also boost the power of associations. However, high correlations between phenotypes may suffer from multicollinearity problems; thus, we should carefully select phenotypes with  $|r^2| < 0.8$  in MPA.

GWAS methodology using MPA has been demonstrated to improve the ability to discover loci ([Galesloot et al., 2014](#); [Mägi et al., 2017](#); [O'Reilly et al., 2012](#)). Multi-phenotype GWAS (MP-GWAS) conduct individual association tests on each variant with multiple phenotypes at once, optimising resource usage and time ([Mägi et al., 2017](#)).

One of the proposed MP-GWAS approaches uses the reverse regression model, as opposed to the commonly used SP-GWAS. This methodology uses the phenotypes as the predictors and the genotypes as the regression outcome ([O'Reilly et al., 2012](#)). Both binary and continuous traits are accepted in this approach, as the reverse regression model does not need any phenotype distribution assumption ([O'Reilly et](#)

al., 2012). The reverse regression model can be described into this formula for the genotype,  $G_i$ :

$$G_i = \alpha + \beta_1 y_1 + \beta_2 y_2 + \dots + \beta_i y_i + \epsilon_i$$

The intercept is  $\alpha$ , the coefficient of regression is  $\beta_i$  for phenotype  $i$ ,  $G_i$  is the  $i$ th genotype,  $y_i$  is the  $i$ th phenotype, and  $\epsilon_i$  indicates an error.

The advantage of this reverse regression methodology is that it can simultaneously incorporate both quantitative traits and categorical phenotypes in the same model increase in power over univariate analyses in many scenarios while controlling false positive error rates. Besides, the reverse regression of the MP-GWAS method with the full model does not require correction for multiple testing as it used the same model (Kaakinen et al., 2017); thus, it will increase the statistical power. However, the disadvantage is that model parameter estimates cannot be directly interpreted in terms of an SNP's effect on each phenotype.

There are many software implementations for the multi-phenotype approach, including the SCOPA (Software for Correlated Phenotype Analysis) developed by our team and collaborators (Mägi et al., 2017). It uses the reverse regression approach as previously introduced by MultiPhen (O'Reilly et al., 2012). In this method, genotypes are treated as the outcomes, and the phenotypes are the predictors of regression.

SCOPA, as a multi-phenotype approach, accepts the genetic data that was converted into genotype dosage, a transformation of the genotype probabilities from imputation. The dosage used in SCOPA's method is an advantage over MultiPhen, which accommodates directly typed and imputed SNPs, accounting for uncertainty in the

imputation. These dosages are regressed on all possible linear combinations of multi-phenotypes at once (Mägi et al., 2017). For example, analysing multi-phenotypes *A*, *B*, and *C* means that all of the possible combinations *A*, *B*, *C* (univariate models), *AB*, *BC*, *AC* (two-phenotype combinations), and *ABC* (full model) are regressed on each variant. This linear combination is also one of the advantages of this software. The majority of available MP methods, such as MultiPhen, do not implement all of these possible combinations, including univariate models; thus, they usually need single-phenotype analysis to be performed separately to compare with MP results. The Bayesian Information Criterion (BIC) is calculated for each model to observe each SNP variance to select the best fitting model (G. Schwarz, 1978). The BIC estimates the maximum likelihood of model fit for each phenotype combinations (G. Schwarz, 1978). The model that has the minimum BIC is selected as the best model by filtering out the other models that tend to over-fit (G. Schwarz, 1978).

The association summary statistics generated by SCOPA can also be combined using the implementation of META-SCOPA. The META-SCOPA extended SCOPA software which has not been implemented in the other MPA software. META-SCOPA implement fixed-effects meta-analysis to perform reverse regression analyses in large-scale consortia when individual-level datasets are not shareable (Mägi et al., 2017).

### **2.3. Conditional analysis on multi-phenotype GWAS**

When the only statistically relevant SNPs at a locus is reported in GWAS, only the total SNP is captured, i.e., with no missing heritability at that locus, given two hypotheses: (1) a single causal underlying SNP exists at the locus and (2) an absolute LD exists between the causal SNP and the top SNP.

However, these two hypotheses can also be challenged when multiple causal SNPs or alleles may exist at the locus, such that not all genetic variation at the locus can be accounted for by a single SNP (Ning et al., 2017). The phenomenon in which multiple causal SNPs or alleles for a given phenotype are found at the same locus is defined in a recent analysis as allelic heterogeneity, the prevalence of which is recorded in multiple complex diseases (Hormozdiari et al., 2017). Besides, an individual top SNP will not capture all of the genetic variances even if there is only one associated causal variant at the locus if the LD between the top variant and the causal variant is not complete (Ning et al., 2017).

Many SP-GWAS meta-analyses have used conditional analysis like GCTA-COJO (Yang, Lee, Goddard, & Visscher, 2011) to classify secondary interaction signals. GCTA-COJO conducts a secondary association analysis based on identified top SNPs using summary statistics of SP-GWAS meta-analysis.

In the SP-GWAS setting, conditional analysis has been proposed to identify the independently associated SNPs in a locus of interest (Deng, 2017). However, subsequent conditional analysis has never been done in many studies in the MP-GWAS setting since the MPA is a relatively new methodology.

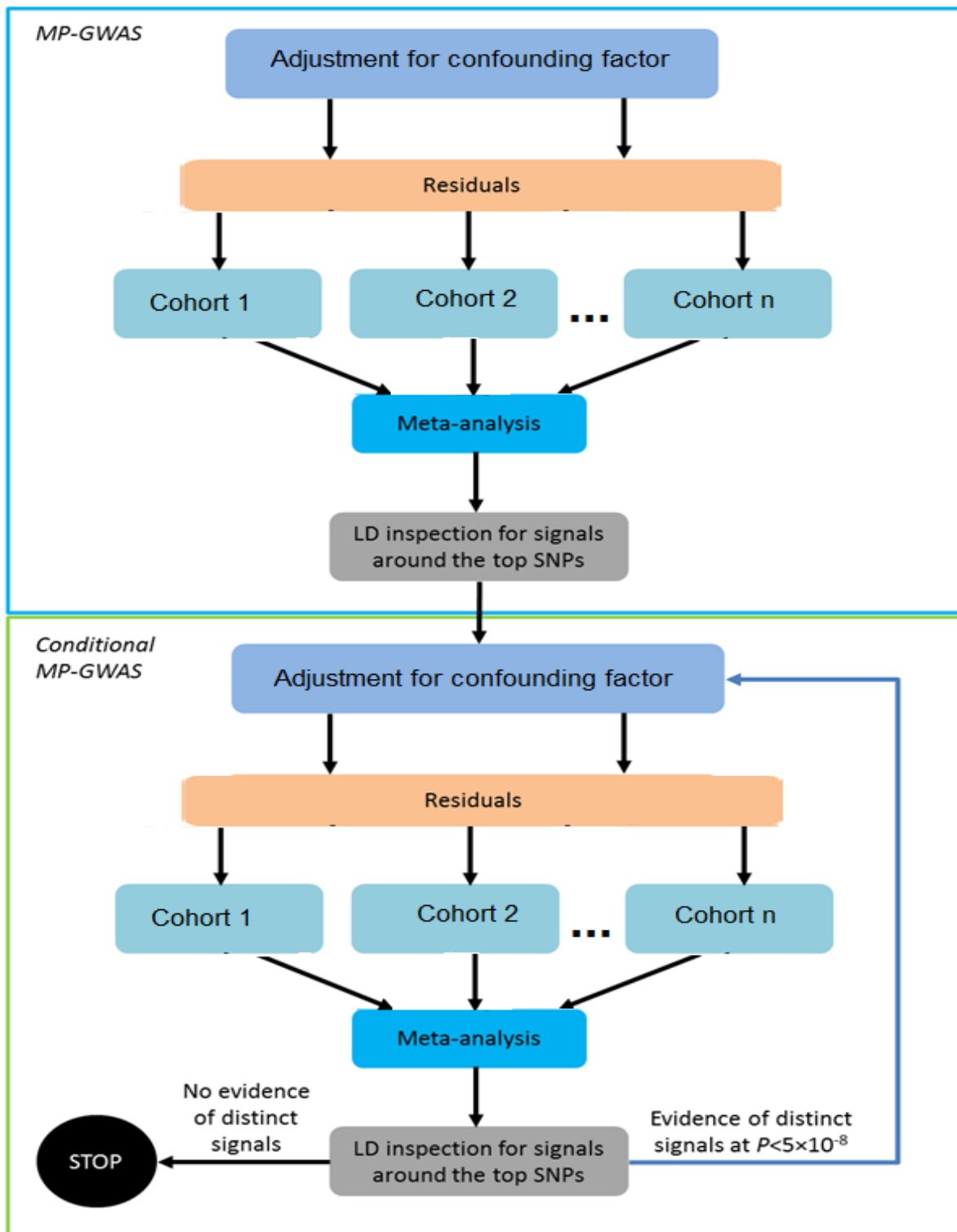


Figure 2.3 Workflow for the conditional MP-GWAS analyses.

A multi-phenotype conditional analysis has been proposed in this thesis to dissect the genetic architecture of the phenotype-associated loci identified in the MP-GWAS framework in individual-level data of the sample. Within MP-GWAS, this can be done

by examining the LD structure and genetic architecture of loci and LD within each locus between variants that reach genome-wide significance that can suggest whether evidence for several different signals is observed.

For the analysis of each identified signal, the effect of the most strongly associated SNP needs to be regressed out from the traits of interest, in addition to the confounding factors, followed by normalisation of the data, e.g., by inverse-normal transformation; followed by meta-analysis results of the contributing studies. An additional distinct signal would be defined if, upon conditioning, there is still an association signal reaching the genome-wide threshold. When distinct signals are observed, conditioning on the original lead SNP and the newly observed distinct SNP(s) is needed to detect further signals. This process is repeated until no additional signal is identified at genome-wide significance, as illustrated in Figure 2.3

#### **2.4. Several approaches to tackle missing phenotype data**

Missing data are a common issue in multi-omics analyses whose standard approach is to base inference only on observed data, the so-called "complete case (CC)" analysis, or listwise deletion. However, it is often disadvantageous in most cases as it only assumes the missing data is MCAR, which is typically rare to occur. As a result, the CC analysis method tends to produce biased parameters and estimates, while this assumption is violated. Besides, using only the observed values will reduce the sample size, reducing the statistical power.

It is well known that inference based on imputed missing values often performs better than the one based only on the observed data. There are mainly two imputation methods coming from mainstream statistics, i.e., single-imputation (SI) and multiple imputation (MI); both can be implemented using the Bayesian approach (implemented

in MICE package in R) and maximum-likelihood approach (implemented in AMELIAII package in R). Besides, there are also imputation approaches based on more advanced machine learning approaches, i.e., imputation using Random Forest (RF), k nearest neighbour (k-NN), and left-censored method QRILC. There is no imputation method available within the MP-GWAS framework as usually CC analyses are used. Employing these more advanced methods within the MP-GWAS framework will boost the power of association testing.

### 2.4.1. Imputation methods based on statistical approaches

#### 2.4.1.1. Single imputation (SI)

The simple imputation (SI) is implemented using several different approaches, categorised, whether it used explicit or implicit modelling, as seen in [Figure 2.4](#).

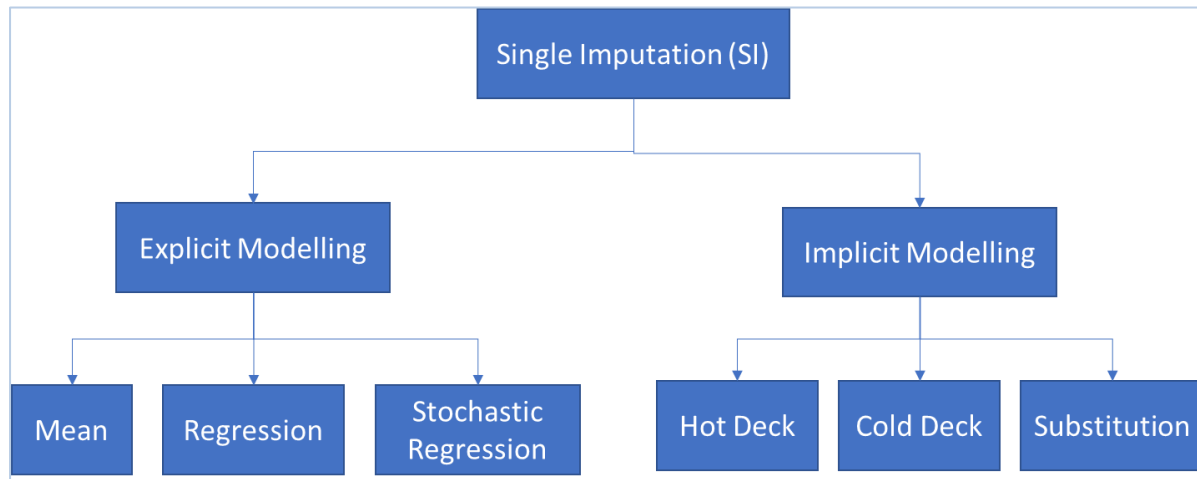


Figure 2.4 Different simple imputation (SI) approaches.

#### **Mean value imputation**

The simplest SI method commonly used is based on mean imputation by replacing all the missing data with the mean of the observed values. If the data are MCAR, the

estimate of the mean remains unbiased as imputing the mean preserves the mean of the observed data (Stasny, 2015). However, this method imputes the missing values with the same value that distorts the correlations between variables. Moreover, a large proportion of missing values leads to underestimated variance estimates (Jamshidian & Mata, 2007).

### ***Stochastic regression***

The random regression imputation applies the most similar techniques employed in predictive regression imputation and adds a stochastic component via the residual terms. Multiple alternative assumptions are available to describe how to generate these residual terms in the process of imputation - normally distributed, randomly selected from residuals of the respondent, or randomly selected from respondents who are similar to the subsidiary variable. A disadvantage that is also unique to the other regression imputations is that they can lead to results that are unlikely to be true.

### ***Regression imputation***

The more modern SI approach is by imputing values from a regression equation that predicts the missing values using the information from the observed data. This approach is a better way to impute the missing values and retain the sample size compared to the CC analysis and mean imputation as information from the individual is used to predict the missing values

This approach considers the correlated phenotypes information as the benefit to estimate the missing values. The assumption is slightly better than the mean imputation, as it assumes MCAR and slightly regulatory condition of MAR. However, this approach still has a less underestimated variance and covariance than the mean imputation approach.

However, the regression analysis used in SI always treats the imputed values as the same as the values on a regression equation, which will overestimate the correlations but underestimate the variances. Besides, the estimates are all treated as if they were true, which is never the case. Therefore, multiple imputation (MI) is proposed to solve this uncertainty (Yuan, 2010).

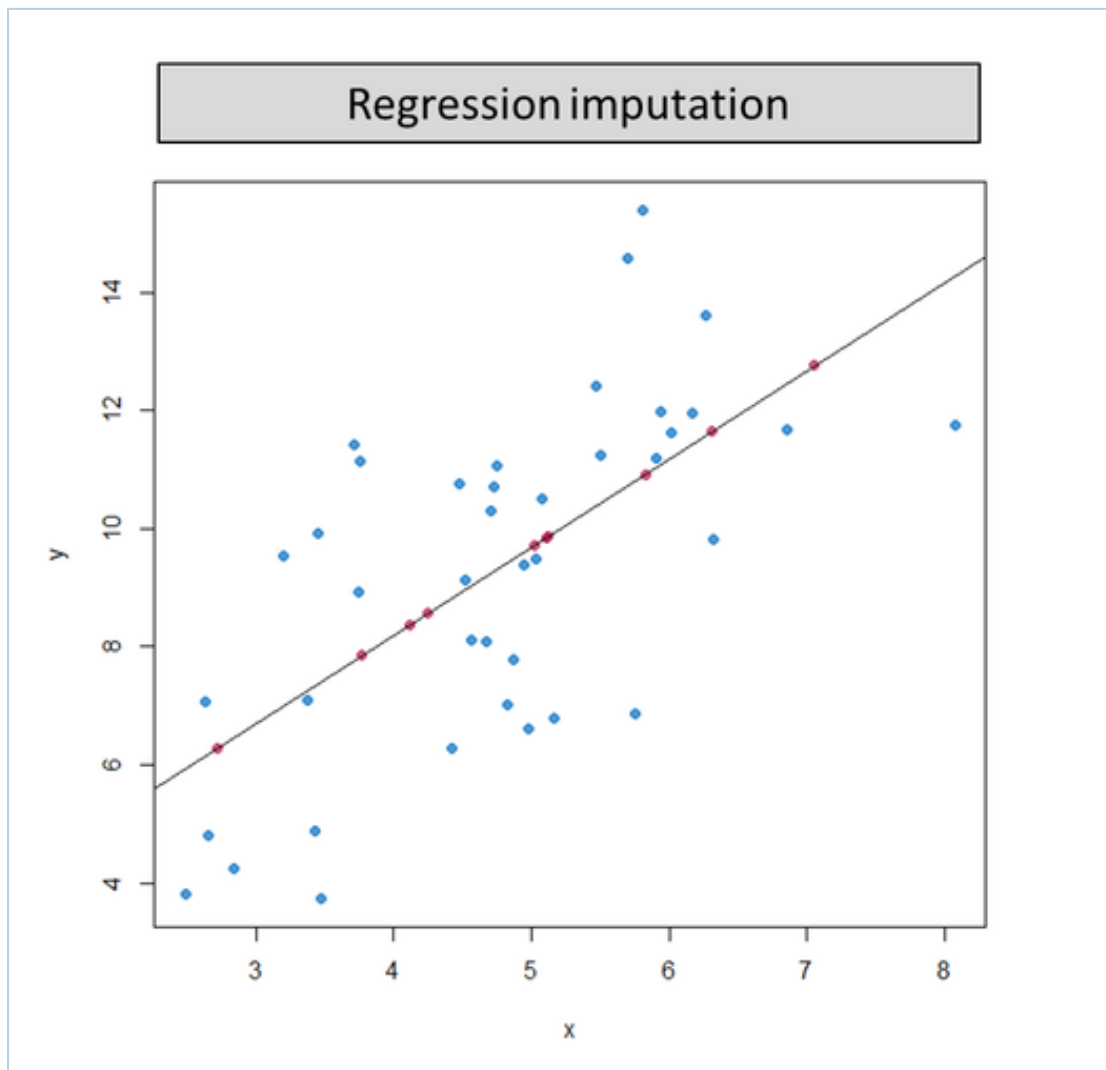


Figure 2.5 The limitation of regression imputation or single imputation (SI) where the predicted values lie on the regression line will overestimate the correlations but underestimate the variances.

### **Hot-deck imputation**

Hot-Deck originated from the decks of cards used in data file processing, where the term "hot" refers to the data file itself. Unobserved data in this imputation are substituted by values observed on a group of matching variables with a very similar

pattern. The regression imputation predicted the unobserved data for a target variable using the target variable's regression over every other variable or a subset of all other variables (Gelman, Hill, Gelman, & Hill, 2010). By preserving the distribution of estimates, this method raises a higher variance than the mean imputation method. This method may reduce the underestimation of estimate variance.

### ***Cold-deck imputation***

Cold-deck imputation imputes a constant value from a source outside of the present data collection, usually an earlier survey. Practically, like in a CC analysis, imputation using this method is ignored due to the lack of a sufficient theory for data analysis generated by this imputation approach (R. J. A. Little & Rubin, 2002). This method was referred to be a historical imputation which is not commonly used in real-world studies.

#### 2.4.1.1. Multiple imputation (MI)

MI provides a useful strategy compared to SI such that instead of imputing a single value, it imputes the missing data with multiple predicted values (e.g., 3–5) as several scenarios of the uncertainty imputed values (Kasner, Hunter, Ph, Kariko, & Ph, 2013). MI generates  $m$  datasets that filling the missingness in  $m$  different values, then analysing each of the datasets separately and pooling the results using Rubin's rule by calculating the pooled regression coefficients and standard errors (SEs) (Rubin, 1988), which is shown in Figure 2.6.

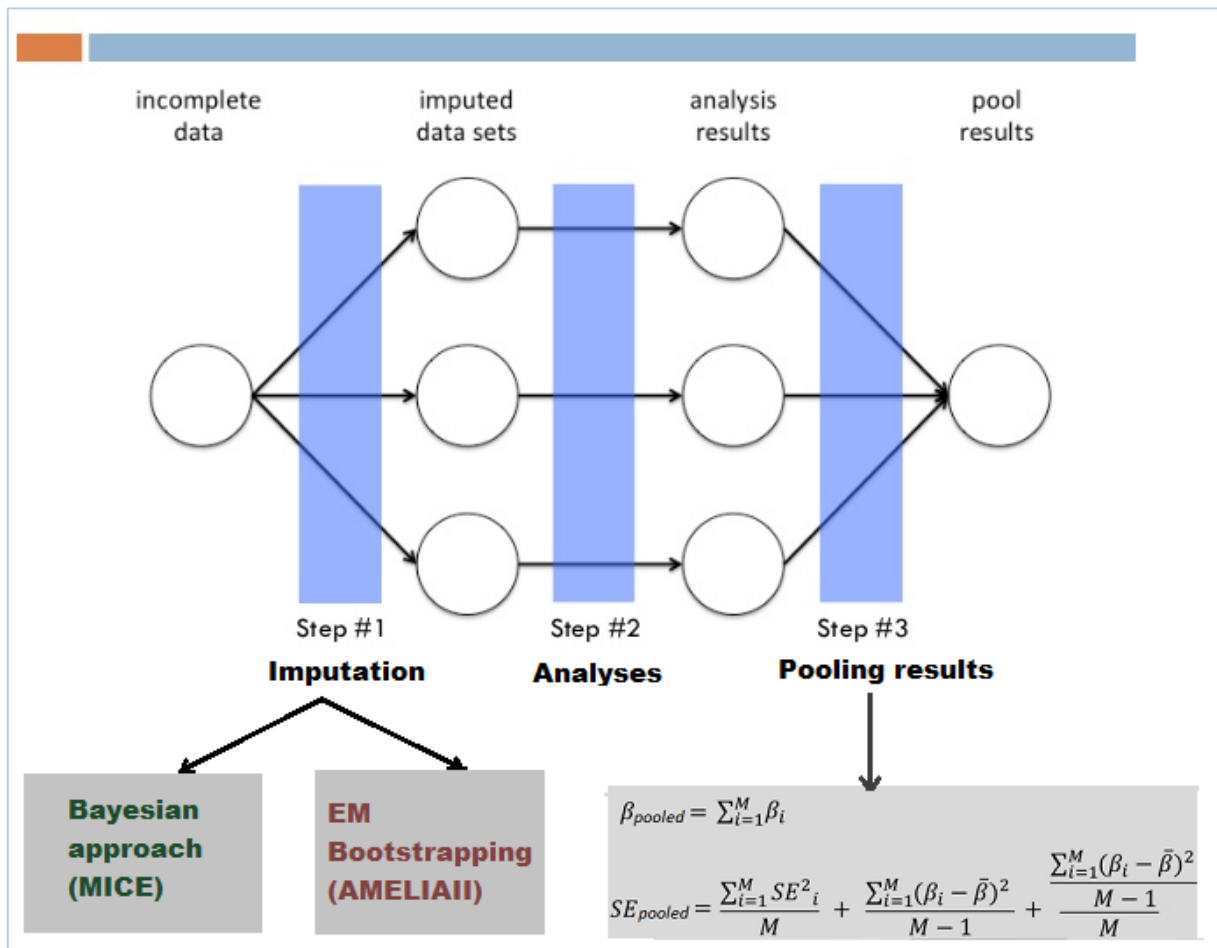


Figure 2.6 Regression coefficient ( $\beta$ ) and Standard Error (SE) formulated for multiple imputation (MI), either using the Bayesian approach or Expectation-Maximisation (EM) bootstrapping, based on Rubin's rule. The number of imputed datasets is  $M$ ; the coefficient of regression is  $\beta_i$  for phenotype  $i$ , SE is the standard error.

The SE is the statistical term to measure the accuracy of a sample distribution in representing a population using the standard deviation of its sampling distribution. Referring to Rubin's rule, the pooling regression coefficient is carried out on the analyses' results for the separate completed datasets by taking the average over the parameter estimates from all imputed datasets. The SE are pooled by combining them within imputation variance, as seen in [Figure 2.6](#).

### ***Pooling Effect estimates***

Once Rubin's rule is used ([Rubin, 1988](#)), this is expected that the estimates of repeated parameters are normally distributed. For all statistical test data, i.e., correlation coefficients, this cannot be inferred. Transformations are performed first for these test statistics before RR can be implemented.

The following formula is used to measure the pooled mean differences ([Rubin, 1988](#)):

$$\bar{\theta} = \frac{1}{m} \left( \sum_{i=1}^m \theta_i \right)$$

$\bar{\theta}$  is the pooled parameter estimate in this formula, m denotes the number of imputed datasets,  $\theta_i$  implies taking the sum of the estimate of the parameter (i.e., mean difference) in each imputed dataset i. This formula is similar to the basic formula where the mean value of a number of series is taken.

### ***Pooling Standard errors***

The pooled standard error is extracted from various components that in the multiple-imputed datasets represent the within and between sampling variance of the mean difference. Below, the computation of these elements is described.

### ***Within-imputation variance***

This would be the mean in each imputed dataset of the mean of the calculation of within variance, e.g., squared standard error. This represents the variance of the sampling, e.g., the precision of the interest parameter in each dataset finished. In small samples, this value will be high and in large samples, small.

$$V_w = \frac{1}{m} \left( \sum_{i=1}^m SE_i^2 \right)$$

$V_w$  is the within imputation variance for this formula,  $m$  denotes the number of imputed datasets,  $SE_i^2$  represents the amount of the squared Standard Error (SE), measured in each imputed dataset,  $i$ .

### ***Between-imputation variance***

This represents the extra variance owing to the missing data. This is calculated by taking the variance over imputed datasets of the parameter of interest estimated. This formula is equal to the formula that is widely used in statistics for the (sample) variance. Where the level of missing data is high, and when the level of missing data is small, this value is large.

$$V_B = \sqrt{\frac{\sum_{i=1}^m (\theta_i - \bar{\theta})^2}{m - 1}}$$

$V_B$  is the between imputation variance for this formula,  $\bar{\theta}$  represents the pooled estimate,  $\theta_i$  measured the estimate of parameters in each imputed dataset  $i$ .

Using additional variables in the imputation process leads to the improvement of the MI model. However, it can also decrease the precision when the correlations are low

as the regression coefficients can be biased with too many variables (Tang & Ishwaran, 2017). MI can be applied using the Bayesian approach implemented by the MICE package in R (Van Buuren & Groothuis-Oudshoorn, 2011) or using Expectation-Maximisation (EM) bootstrapping as implemented by the AMELIAII package in R (Honaker, King, & Blackwell, 2011).

MICE impute data on a variable-by-variable basis, whereas AMELIAII uses a joint modelling approach based on a multivariate normal distribution. So AMELIAII uses means and covariances to summarize data. First, it takes  $m$  bootstrap samples and applies the EM bootstrapping algorithm to each sample. Then apply the regression of these samples to obtain the estimates.

With that, MICE is capable of handling different types of variables, including categorical and continuous data without a prior assumption of normality (Van Buuren & Groothuis-Oudshoorn, 2011).

In contrast, the variables in AMELIAII need to be normally distributed or transformed to approximate normality (Honaker et al., 2011). However, MICE have a significant limitation since it will need more computational time when the data is huge. In this case, AMELIAII is a way faster.

MI performs reasonably at its best missing data are MAR ((Eekhout et al., 2014). The variables associated with absence may be used in the imputation model. In this way, bias is decreased, and predictions are more reliable.

Despite the benefits, MI may have poor performance in high-dimensional data, e.g., omics data (Yuan, 2010). It is recommended that all variables be used in imputation with MI to avoid bias in estimating a correlation. However, it can also create

overparameterization when the variables are greater than the number of individuals (Tang & Ishwaran, 2017).

## **2.4.2. Imputation methods based on Machine Learning (ML) approaches**

### **2.4.2.1. *Random forest (RF)***

Random Forest (RF), based on Breiman's RF algorithm (Breiman, 2001), is a promising machine learning approach that can adjust to the high dimensionality of data while avoiding overfitting and has recently been proposed to replace MI (Package & Lazar, 2015). Both RF and MI are capable of processing different types of data.

It is possible to summarise the RF algorithm as follows (Hong & Lynn, 2020):

1. Initialisation. The missing values will be replaced with their mean (for continuous variables) or their highest frequent class for a variable containing missing values (for categorical variables).
2. Imputation. For the data variables, the imputation method is performed sequentially in the increasing (or decreasing if assigned) order of missing observations to every variable. For building the RF model, the variable under imputation is used as the response. The observations of the dataset are split into two sections, depending on whether the variable in the original dataset is observed or absent. As the training set, the observed observations are used, and as the prediction set, the missing observations are used. The missing portion of the imputed variable is replaced by the RF classifier (Breiman, 2001; Liaw & Wiener, 2002).

3. Finalisation. Imputation is finished when all variables with missing data have been imputed. A further iteration of the imputation process continues until missForest produces the previous imputation as the final result and the relative sum of squared differences (or the fraction] of misclassified entries for categorical variables) raises between the actual and the prior imputation result (Stekhoven & Bühlmann, 2012a).

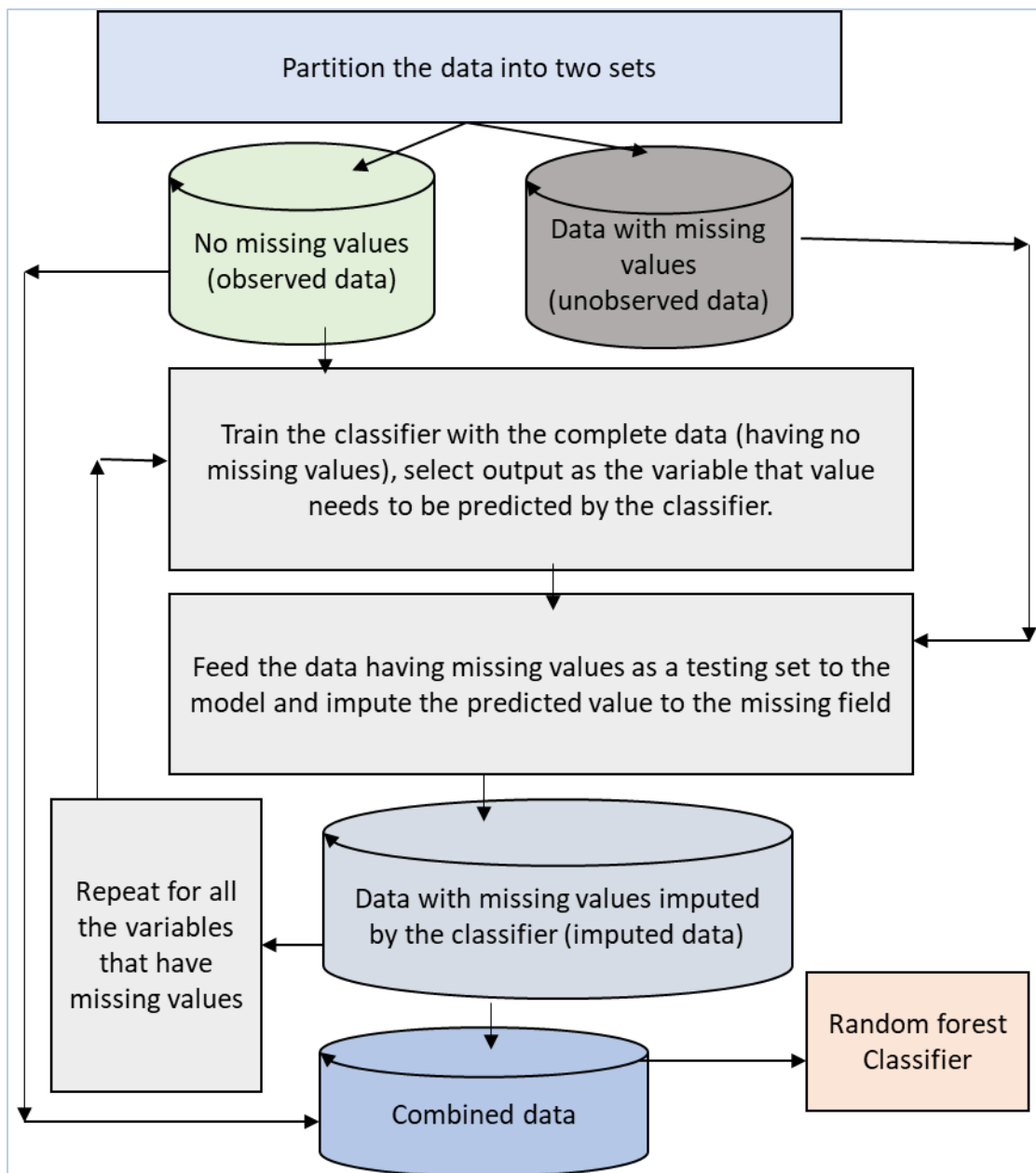


Figure 2.7 Workflow of the random forest (RF) algorithm.

The out-of-bag (OOB) approach has become a method of verifying the RF model and is specified in the RF classifier (Breiman, 2001). The original training data in the model is arbitrarily sampled-with-replacement, producing partial data subsets. These subsets were therefore defined as collections of bootstraps.

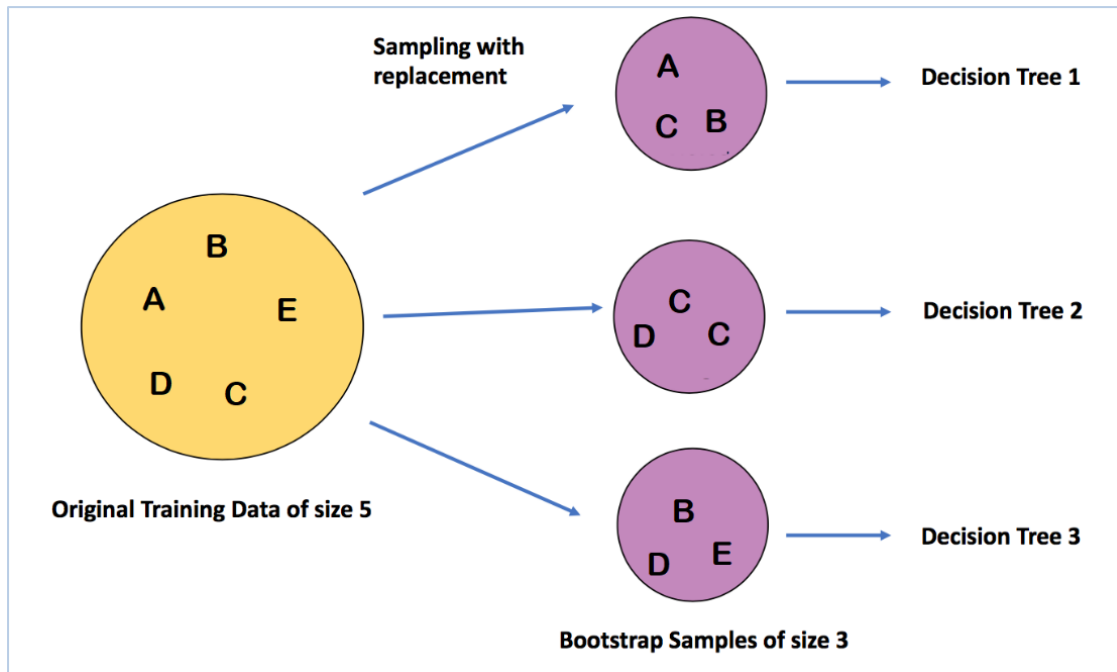


Figure 2.8 Workflow of bootstrap sample generation with replacement. "Sampling-with-replacement" here implies that if a data point is selected in the first random subsample, another random subsample will follow with equal probability remains in the original sample for selection. This replacement sampling can be seen on the picture as "C", which is chosen twice in the second bootstrap subsample.

These bootstrap samples are then fed to several broad depth decision trees (DTs) as training data. Each of these DTs on these bootstrap samples is separately trained. The RF ensemble is called this aggregation of DTs. By counting a consensus vote from all the DTs, the final result of the ensemble model is decided. This notion is referred to as Bootstrap Aggregation or Bagging.

The deviations in the original training dataset do not affect the final result obtained from the aggregation of DTs, since each DT takes as input a different collection of training data. Bagging as a term, therefore, reduces variance without modifying the complete ensemble's bias.

Every subsample of the OOB sample rows is run through each DT in its bootstrap training data that do not include the OOB sample row, and a majority prediction for each row is noted.

If the training dataset contains N rows. Then, in a random draw, the probability of not selecting the same a row is (Breiman, 2001):

$$\frac{N - 1}{N}$$

That becomes equivalent to, at the limit of large N (Breiman, 2001):

$$\lim_{N \rightarrow \infty} \left(1 + \frac{1}{N}\right)^N = e^{-1} = 0.368$$

Therefore, approximately 36.8% of the total training data is available for each DT as an OOB sample and can therefore be used to test or validate the RF model. And finally, the OOB score is determined from the out-of-bag sample as the number of correctly predicted rows.

#### **2.4.2.2. *K nearest neighbour (kNN)***

Another machine learning imputation approach is k-Nearest Neighbour (kNN), which is applicable in any of the three missing mechanisms when there is a link between the missing value variable and the other variables. The assumption to impute missing values using kNN is that the missing variable's value can be inferred by the observed values of other variables closest to it (Beretta & Santaniello, 2016).

In this method, the missing values are imputed with observed data from the nearest neighbours calculated by Euclidean distance.

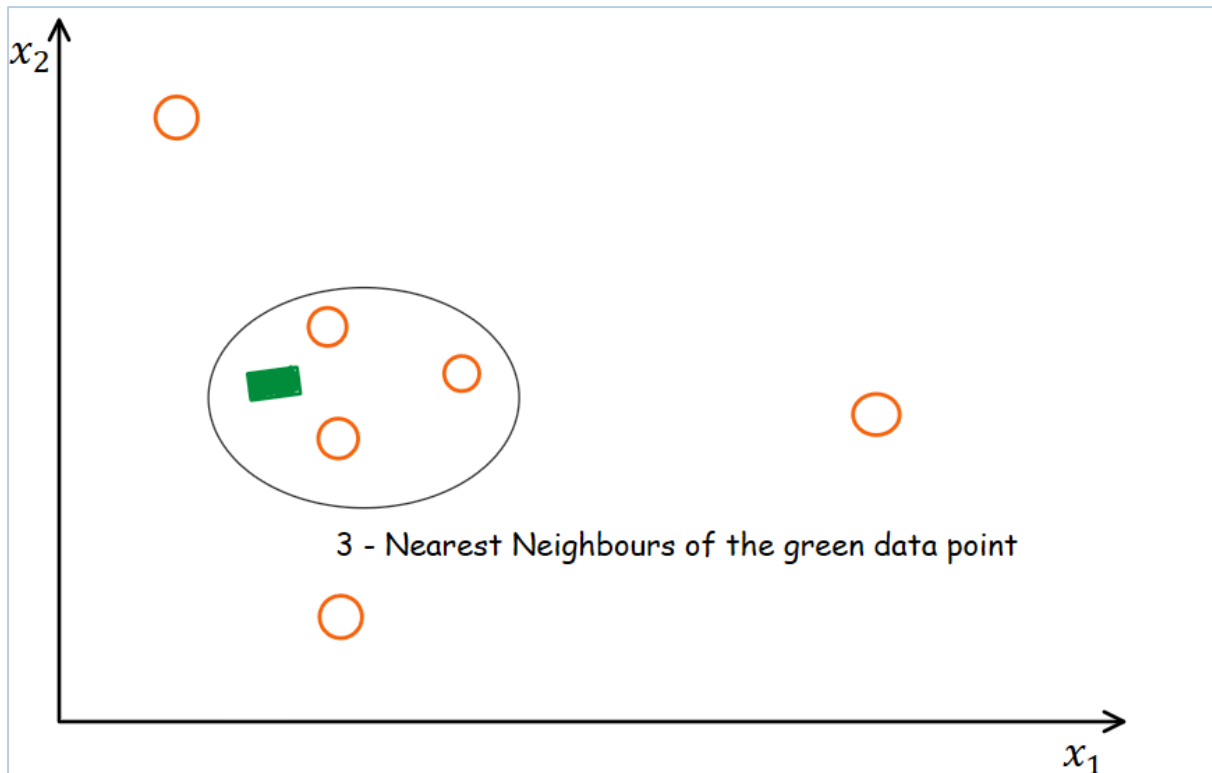


Figure 2.9 Illustration of the  $k$  nearest neighbour (kNN) algorithm. The oval area reflects, in this case, the adjacent points of the green square data point. The calculation of space is used using Euclidean distance to define neighbours.

The idea in kNN methods is to identify 'k' samples in the dataset that are similar or close in the space. Then these 'k' samples are used to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbours found in the dataset.

### ***Distance calculation in the presence of missing values***

Consider the two-dimensional space of a pair of observations (2,0), (2,2), (3,3). A graphical representation of the following points is given in [Figure 2.10](#):

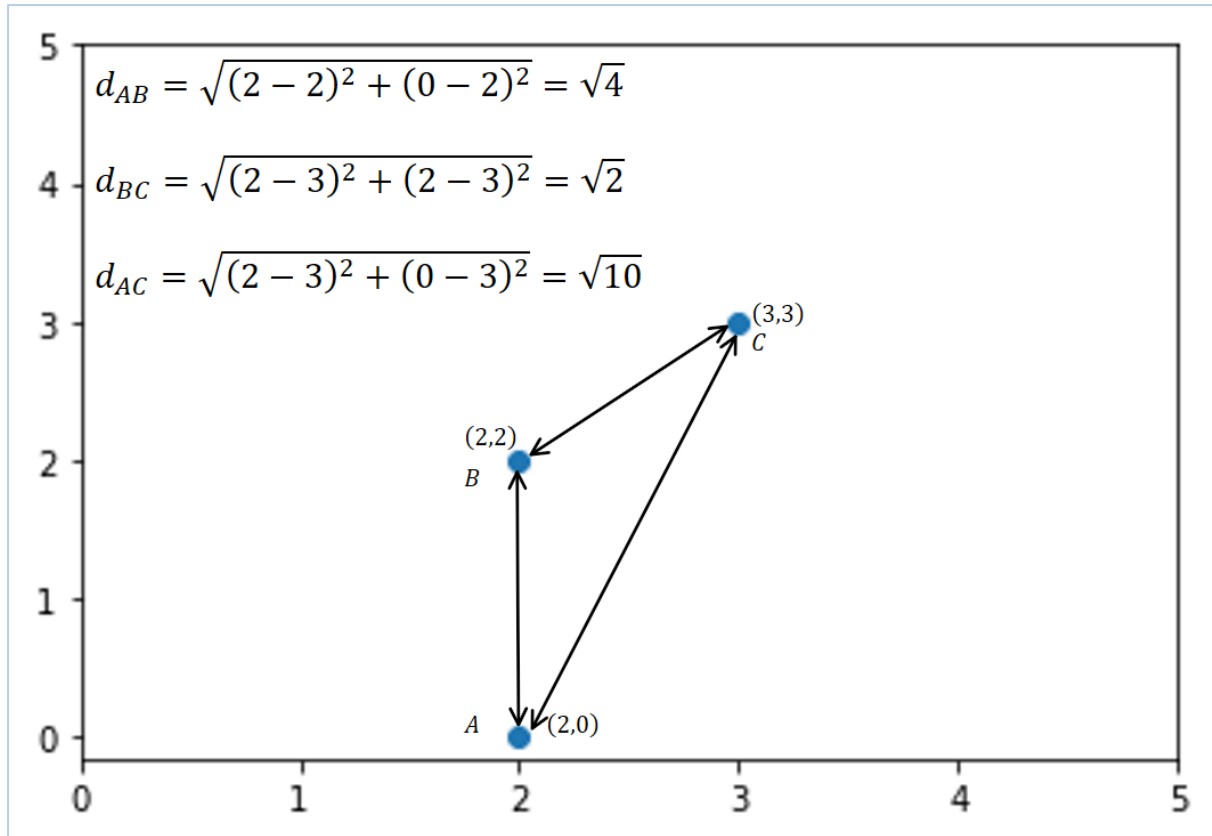


Figure 2.10 The calculation of the distance space in the presence of missingness of the  $k$  nearest neighbour ( $k$ NN) algorithm.

In this example, the shortest distance points based on Euclidean distances are considered as the nearest neighbours. The first closest neighbour of point A, for example, is point B. The first closest neighbour for point B is point C.

When the coordinates are missing, the Euclidean distance is computed by excluding the missing values and upscaling the weight of the coordinates that are not missing.

$$d_{xy} = \sqrt{\text{weight} * \text{square distance from present coordinate}}$$

In summary, how  $k$  is chosen to impute the missing values with the  $k$ NN algorithm is potentially contentious. In addition, with different values of  $k$ , it was suggested for the model to be tested using cross-validation after imputation is performed. However,

imputation of missing values is a continuously advancing research area with kNN serving as a convenient and effective algorithm.

#### *2.4.2.1. Left-censored imputation*

In metabolomic data generated by mass spectrometry (MS), missing values usually present due to the limit of detection/quantification (LOD/LOQ), which could be categorised as MNAR. This missingness could decrease the statistical power in the analysis as MNAR leads to distorting the data distribution. Although several imputation approaches have been designed for omics data, only a limited method has been developed specifically for the data already known as having a missing mechanism of MNAR.

Left-censored missing data using QRILC is also one of the promising ML approaches that have been designed to handle missing values in high throughput metabolomics data ([Package & Lazar, 2015](#)), which usually have missing values due to both technical and biological reasons which can generate MNAR, the most difficult type of missing data which generally cannot be handled properly by SI or MI. Typically, missing values in metabolomics data are substituted with a minimum value, leading to bias in downstream analyses ([Package & Lazar, 2015](#)).

It has been suggested that log-transformation of the variables should be performed before QRILC imputation to improve the imputation accuracy and positive value on the original scale.

Left-censored imputation method imputed the missing data from a truncated distribution with parameters calculated using quantile regression using random draws. Quantile regression calculates the conditional median of the target, unlike standard

linear regression that uses the least-square approach to measure the conditional mean of the target over different values of the features (P. Deelen et al., 2014).

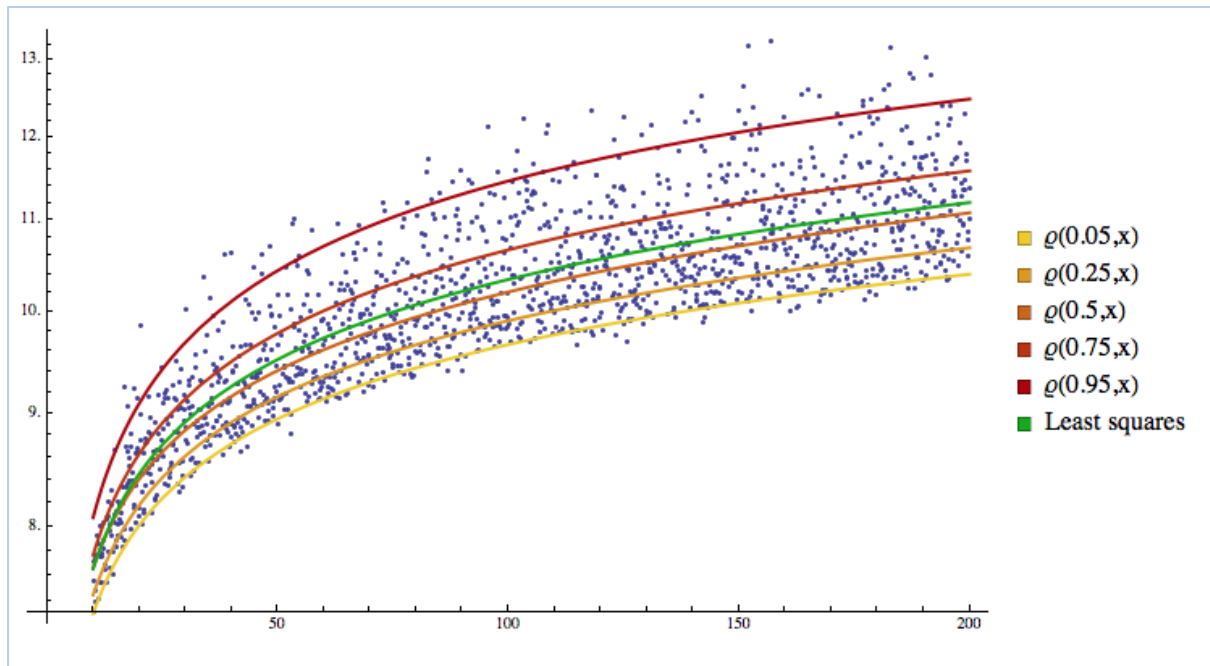


Figure 2.11 The data fitted with quantile regression

Quantile regression is a linear regression extension used when the linear regression requirements are not satisfied (i.e., linearity, homoscedasticity, independence, or normality).

The formula of the linear regression model to calculate the mean is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \quad i = 1, \dots, n$$

The linear regression model to computing the mean involves a process in which  $p$  is equal to the number of variables in the formula, and the number of data points is  $n$ . By minimizing the mean square error (MSE), which is found with the formula, the perfect linear regression line is obtained (Succop, Clark, Chen, & Galke, 2004):

$$MSE = \frac{1}{n} \left( \sum_{i=1}^n \left( y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \right)^2 \right)$$

It is not limited to only finding the median for quantile regression, and it can measure any quantile (percentage) in the function variables for a specific value. For instance, if we find the 25<sup>th</sup> quantile for the missing values, it implies that the true value has a 25% chance of being below the prediction, while the value has a 75% chance of being above it.

The equation of the quantile regression model for the  $\tau$ th quantile has a structure similar to the linear regression model (Succop et al., 2004):

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \dots + \beta_p(\tau)x_{ip} \quad i = 1, \dots, n$$

The equation implies that the beta coefficients now operate with a dependence on the quantile, rather than constants. It takes the same procedure as the standard linear quantization to find the values for these betas at a given quantile value, with the exception that we now need to reduce the median absolute deviation (MAD) (Succop et al., 2004):

$$MAD = \frac{1}{n} \sum_{i=1}^n \left( y_i - (\beta_0(\tau) + \beta_1(\tau)x_{i1} + \dots + \beta_p(\tau)x_{ip}) \right)$$

Here, the function  $\rho$  is the test function that, based on the quantile and the overall sign of the error, assigns asymmetric weights to the error. Mathematically,  $\rho$  has the form (Succop et al., 2004):

$$\rho_\tau(u) = \tau \max(u, 0) + (1 - \tau) \max(-u, 0)$$

In this case, if the single data point error, then the max function returns the largest value. In other words, when the error is positive, it is multiplied by  $\tau$  using the check function, and when the error is negative, it is multiplied by  $(1 - \tau)$  using the check function.

For instance, if the median of the 10<sup>th</sup> percentile is what we want, it means that 90% of the errors must be positive, and 10% must be negative. For this statement to be valid, weights must be added to the errors to find the smallest MAD. In the 10<sup>th</sup> magnitude, the negative weights are added to a weight of 0.9, while the positive ones are added to a weight of 0.1. (Wei, Wang, Su, Jia, Chen, & Chen, 2018)

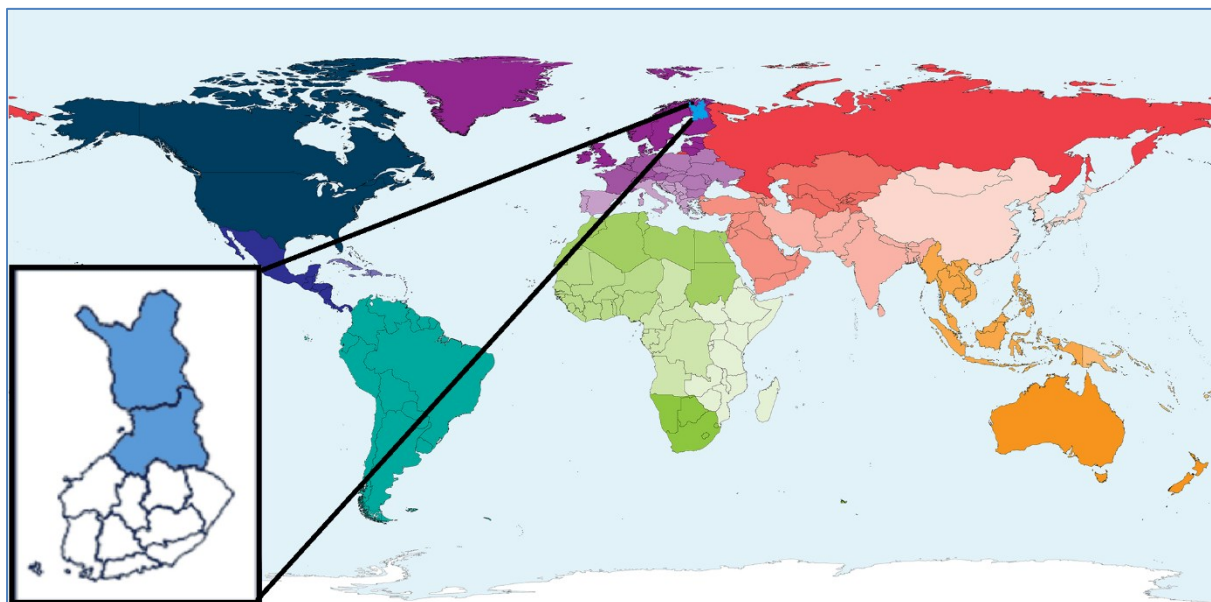
### 3. Materials and Methods

*This chapter introduces the studies of Northern Finland Birth Cohort (NFBC) 1966 and 1986; their genetic and phenotypic data are described. In the next subchapter, I also listed statistical methods that contributed to my thesis work analyses.*

#### 3.1. Study populations

##### 3.1.1. Northern Finland birth cohort (NFBC) 1966 & 1986

The Northern Finland Birth Cohorts (NFBC) include two independent studies: NFBC1966 and NFBC1986, cohorts defined by expected birth dates for 1966 and January 1<sup>st</sup>, 1985 to June 20<sup>th</sup>, 1986). These studies provide longitudinal and prospective birth cohorts over 20-year intervals of over 20,000 children born from the same two of Finland's northernmost regions (see [Figure 3.1](#)) ([Faculty of Medicine University of Oulu, 2015b](#)).



[University of Oulu, 2015b](#)).

*Figure 3.1 Geographical location of the Northern Finland Birth Cohort (NFBC) 1966 and 1986. Blue represents the targeted area of the cohorts.*

Based on their study design, both cohorts have several advantages:

Firstly, the high percentages of the same age participants providing no age-related bias in the data ([Sabatti et al., 2009](#)). Age usually was analysed as confounding in some research, which is a bias because it can skew the correlation between exposure and the predicted result. In any research design (e.g., observational, case-control, cohort, ecological), confounding could be occurring, mainly because it is not a product of the design of the study. Participants were born in the same year, removing age as a possible confounder, and mitigating the impact of inter-individual heterogeneity due to secular shifts in environmental factors; longitudinal study collected over many decades ([Marjo Riitta Järvelin et al., 2004](#); [Laitinen, Pietiläinen, Wadsworth, Sovio, & Järvelin, 2004](#); [Saari et al., 2005](#)), promotes research into the relationships between genetic variation, early life experiences, and disease risk progression.

The waist-hip ratio (WHR) was also found to increase with a person's age ([Lahti-Koski & Gill, 2004](#)). The age effect on abdominal obesity may be confounded by the overall adiposity level, considering that body mass index (BMI) typically tends to rise with age ([Läärä & Rantakallio, 1996](#); [Lahti-Koski & Gill, 2004](#); [Molarius, Seidell, Sans, Tuomilehto, & Kuulasmaa, 1999](#)).

Secondly, the participants came from a genetic isolate area to minimize the effects of population substructure ([Sabatti et al., 2009](#)). NFBC1966 collected 96% of all babies (N=12,231) born with a birth weight of  $\geq 600$  grams in the Finnish provinces of Oulu and Lapland in 1966 ([M R Järvelin et al., 1997](#); [Rantakallio P., 1969](#)). Originally, the total number of children born was expected to 12,231; however, it was decreased to 12,058 live-born children from 12,068 deliveries ([M R Järvelin et al., 1997](#); [Rantakallio P., 1969](#)).

Like NFBC1966, NFBC1986 covers 99% (N=9,479) of the babies born from 1<sup>st</sup> of July 1985 to 30<sup>th</sup> of June 1986 (M R Järvelin et al., 1997; Rantakallio P., 1969). The NFBC1986 data were collected from 9,362 births, with 9,432 were live-born (M R Järvelin et al., 1997; Rantakallio P., 1969).

The data of both cohorts are comprised of metabolomics, metabolic trait, and epidemiologic data. It was obtained from participants' clinical examination and medical records from hospitals with their national register data and postal questionnaires sent to certain follow-up studies. To be approved by the ethical committees of Northern Ostrobothnia Hospital District and the University of Oulu, all participants were asked for their written informed consent. In contrast, the participants of NFBC1986 were asked to sign a written informed consent both from the participants and their parents. Both studies also have been approved by Imperial College London to be used in this project.

*Table 3.1 Numbers of invited, participated, and gave participants consent that their data were provided from clinical data collection and questionnaire (area of catchment: NFBC1966 / NFBC1986).*

|                                      | Numbers of participants (N) |                           |                      |
|--------------------------------------|-----------------------------|---------------------------|----------------------|
|                                      | N <sub>invited</sub>        | N <sub>participated</sub> | N <sub>consent</sub> |
| <b>NFBC1966 31-year follow-up</b>    |                             |                           |                      |
| Clinical examination                 | 8,463                       | 6,033                     | 6,007                |
| Questionnaire                        | 11,541                      | 8,767                     | 8,690                |
| <b>NFBC1986 15–16-year follow-up</b> |                             |                           |                      |
| Clinical examination                 | 9,215                       | 6,798                     | 6,795                |
| Questionnaire (parents/adolescents)  | 9,215                       | 6,985 / 7,344             | 6,866 / 7,182        |

The participants of NFBC1966 were examined in longitudinal studies from the early age of 1 year with the follow-up at 14, 31, and 46 years (Faculty of Medicine University of Oulu, 2015a). In comparison, those of NFBC1986 were followed up at the ages of

0 to 1 year (growth data), 7, 8, 15-16 year (and subsample studies until over 20 years) (see [Table 3.1](#)) ([Faculty of Medicine University of Oulu, 2015b](#)). In this project, we used the data from the followed-up study at the age of 31 (NFBC1966) and 15-16 (NFBC1986) as both data were the most recent.

Medical check-up and blood sample collection were conducted in Finland's capital city (Helsinki) and the Finnish provinces of Oulu and Lapland ([Faculty of Medicine University of Oulu, 2015a](#)).

In NFBC1966, the clinical examination was performed with consent by 6007 participants, while 8690 participants filled in the questionnaire. In NFBC1986, due to their age, not only were the adolescents asked the consent, but also their parents. In total, 7182 adolescents gave consent, and only 6866 parents gave their consent for their children's data to be used for research. For clinical examination, 6795 individuals participated and consented to the use of their data.

#### ***3.1.1.1. Genetic data***

The genetic data were extracted from blood samples of 31-year clinical examination of NFBC1966 and 16-year clinical examination of NFBC1986. Genomic DNA was genotyped from blood samples of 5,753 NFBC1966 participants by utilizing the Illumina Infinium HumanCNV370-Duo array. Genotyping of 3,815 NFBC1986 subsamples (random sample and "exposure" sample) was performed on the Illumina Infinium HumanOmniExpressExome Chip. In the end, 5,402 individuals from the NFBC1996 cohort were genotyped successfully, followed by 3,743 NFBC1986 individuals.

The genotype data of the NFBC1966 were imputed for the whole genome genotyping with 1000G high-density reference panel of March 2012 with IMPUTE2 (Marchini & Howie, 2010), after applying exclusion criteria and quality control (QC) procedures (Sabatti et al., 2009). Similarly, with NFBC1986 genotype data, IMPUTE2 was used for imputation; however, with the newer version of 1000G Phase 3 reference panel launch of August 2015 as its reference. Non-monomorphic SNPs with imputation quality score > 0.4 in both cohorts were selected for analysis. In the meta-analysis, only if the SNP was present in both of the cohorts were considered.

The imputed data were also selected based on a minor allele frequency (MAF) > 1.0%. The choice of MAF cut-point was selected to reflect the genotyping accuracy of lower frequency SNPs. Additional QC was also applied based on the Hardy-Weinberg equilibrium (HWE) principle that excluded the deviated allele frequencies from the binomial distribution (p-value < 0.0001).

The standardized QC procedures of the 1000G imputed NFBC1966, and NFBC1986 datasets are summarised in Table 3.2 (Sabatti et al., 2009).

*Table 3.2 Genotype Quality Control criteria applied in the NFBC1966 and NFBC1986 studies*

| Inclusion criteria                       | NFBC1966 | NFBC1986 |
|------------------------------------------|----------|----------|
| Minor allele frequency (MAF)             | > 1.0%   | > 1.0%   |
| Hardy-Weinberg equilibrium (HWE) p-value | > 0.0001 | > 0.0001 |
| Call rate                                | > 95.0%  | > 99.0%  |
| Genotyping completeness                  | > 95.0%  | > 98.9%  |
| Total number of individuals              | 5402     | 3743     |

Although all participants live in the same area of a genetic isolate, which is relatively homogeneous, population structure can still be detected in their genetic data.

Therefore, to avoid spurious associations, the genetic data needs to be adjusted for principal components (PCs) or multidimensional scaling (MDS) (Price et al., 2006). Both of the methods infer genetic ancestry are usually used in standard GWAS.

In NFBC1966, we used PCs composed of a limited number of dimensions trying to define the variation spectrum, which was previously computed for the DNA data of the NFBC1966. In NFBC1986, we used MDS, a method similar to PCs that reduce the highly-dimensional genetic data (Mardia, Kent, & Bibby, 1979), for a subset of the data genotyped on the MethaboChip. Due to time constraints, we did not re-calculate the MDS for all NFBC1986 data. In total, only 3,743 NFBC1986 genotyped data were used. For NFBC1966, 5,402 genotyped data were used for PCs calculation.

#### *3.1.1.2. Phenotype data*

Anthropometric measurements, glycaemic traits (fasting glucose and insulin), and 149 micro molecules essential for metabolisms (metabolites) obtained from blood, were collected from both NFBCs. Measurements of size and weight were used to calculate the Body Mass Index (BMI), which was preferred to self-reported BMI measurements by participants. For measuring the metabolite levels, the samples were taken after at least 12 hours of fasting before the assessment to ensure dietary consumption did not affect the metabolic levels (Groop, Mäkiperna, Stenman, DeFronzo, & Teppo, 1990; Kettunen et al., 2016).

The metabolite levels measurements were precisely quantified by proton ( $^1\text{H}$ ) high throughput nuclear magnetic resonance (NMR) spectroscopy. In a metabolomic study, metabolites can be precisely identified and quantified by NMR (Soininen, Kangas, Würtz, Suna, & Ala-Korpela, 2015).

Using different settings of the NMR metabolomics platform, particular metabolites were detected (see [Supplementary Table 1](#)) in three molecular windows: low-density lipoproteins (LIPO), lipid extracts (LIPID), and other low-molecular-weight metabolites (LMWMs) ([Soininen et al., 2015](#)). A ratio phenotype was used for the apolipoprotein ratio between apolipoprotein A and apolipoprotein B and eight FA ratio as the proportion of FAs represent the influence of FAs on each other ([Schwertner & Mosser, 1993](#)). Taken together, these three windows and the ratios captured 125 metabolites (see [Supplementary Table 1](#)). The proportion of missing phenotype data in NFBC1966 and NFBC1986 are shown in [Table 3.3](#).

*Table 3.3 The proportion of missingness for each category of phenotypes in NFBC1966 and NFBC1986. The means and medians were calculated from the phenotypes in the same category.*

|                   | NFBC1966 (N = 5402)      |                                | NFBC1986 (N = 4173)      |                                |
|-------------------|--------------------------|--------------------------------|--------------------------|--------------------------------|
|                   | Number of missing values | The proportion of missing data | Number of missing values | The proportion of missing data |
| Anthropometric    |                          |                                |                          |                                |
| <b>Mean</b>       | 203.9                    | 3.8%                           | 12.7                     | 0.3%                           |
| <b>Median</b>     | 276.0                    | 5.1%                           | 12.0                     | 0.3%                           |
| Insulin/Glucose   |                          |                                |                          |                                |
| <b>Mean</b>       | 324.8                    | 6.0%                           | 128.5                    | 3.1%                           |
| <b>Median</b>     | 338.0                    | 6.3%                           | 121                      | 2.9%                           |
| LIPID window      |                          |                                |                          |                                |
| <b>Mean</b>       | 341.9                    | 6.3%                           | 18.3                     | 0.4%                           |
| <b>Median</b>     | 331.0                    | 6.1%                           | 16.0                     | 0.4%                           |
| LIPO window       |                          |                                |                          |                                |
| <b>Mean</b>       | 252.0                    | 4.7%                           | 2.0                      | 0.05%                          |
| <b>Median</b>     | 252.0                    | 4.7%                           | 2.0                      | 0.05%                          |
| LMWM window       |                          |                                |                          |                                |
| <b>Mean</b>       | 257.8                    | 4.8%                           | 29.0                     | 0.07%                          |
| <b>Median</b>     | 254.0                    | 4.7%                           | 27.5                     | 0.07%                          |
| Other metabolites |                          |                                |                          |                                |
| <b>Mean</b>       | 349.4                    | 6.5%                           | 47.6                     | 1.1%                           |
| <b>Median</b>     | 261.0                    | 4.8%                           | 11                       | 0.2%                           |
| All phenotypes    |                          |                                |                          |                                |
| <b>Mean</b>       | 327.1                    | 6.1%                           | 17.5                     | 0.4%                           |
| <b>Median</b>     | 254.0                    | 4.7%                           | 2.0                      | 0.04%                          |

Overall, the proportion of missing data in NFBC186 is lower than in NFBC1966. *The other metabolites* category, i.e., metabolite ratios, in NFBC1966, has the highest number of missing values with an average of 6.5%.

### 3.2. Statistical analyses

In this thesis, different statistical methods in the SP-GWAS and MP-GWAS setting were used, with specific methods described in [Table 3.4](#).

*Table 3.4 The statistical methods used in each analysis in this thesis.*

| Analyses                                            | The statistical method used in the analyses                                                                                                                                                                                                                                                              |
|-----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>GWAS OF glycaemic health</b>                     | Polygenic risk score (PRS) calculation                                                                                                                                                                                                                                                                   |
| <b>MP-GWAS of fatty acid (FA) metabolism</b>        | <ul style="list-style-type: none"> <li>• MP-GWAS using SCOPA</li> <li>• Phenotype imputation using single imputation (SI)</li> <li>• Conditional analysis of MP-GWAS</li> </ul>                                                                                                                          |
| <b>Phenotype imputation</b>                         | Comparison of several imputation methods: <ul style="list-style-type: none"> <li>• Single imputation (SI) and multiple imputation (MI) with two approaches: Bayesian approach and bootstrap with sampling</li> <li>• k Nearest Neighbour (kNN)</li> <li>• QRILC</li> <li>• Random forest (RF)</li> </ul> |
| <b>Software development of phenotype imputation</b> | imputeSCOPA, a random forest (RF) imputation method based on Ranger software                                                                                                                                                                                                                             |
| <b>MP-GWAS of lipids</b>                            | <ul style="list-style-type: none"> <li>• Imputation using imputeSCOPA</li> <li>• MP-GWAS using SCOPA</li> </ul>                                                                                                                                                                                          |

## 4. GWAS of glycemic traits in two healthy birth cohorts

*In this chapter, I am describing the analysis of the variance explained by both fasting glucose (FG) and fasting insulin (FI) in our NFBC1966 and NFBC1986 data as part of the cohorts contributing to the meta-analysis GWAS within MAGIC.*

### 4.1. Meta-analyses of Glucose and Insulin-related traits Consortium (MAGIC)

GWAS have dissected the genetic architecture of glycemic traits ([Wheeler, Marenne, & Barroso, 2017](#)). MAGIC combines data from multiple SP-GWAS to pinpoint additional sites that affect glycaemic and metabolic properties. FG, FI, 2-hour glucose (2hglu), and glycated haemoglobin (HbA1c) are classical polygenic traits that have provided general insights into the genetic architecture of glycaemic traits as they are also used to diagnose and prognosis in people with T2D, such as to monitor glucose control ([Wheeler, Marenne, et al., 2017](#)). All of these glycemic traits are associated with MetS, even within the non-diabetic range ([Sarwar et al., 2010](#)).

Previous SP-GWAS, Metabochip, and exome array have detected >120 loci associated with glycemic traits in individuals without diabetes ([Horikoshi et al., 2015](#); [Wheeler, Marenne, et al., 2017](#)). However, most of these glycemic traits were studied in individuals without diabetes in European ancestry with limited representation in other ancestries and inadequate resolution for the good of causal variants, although the prevalence of T2D risk factors across ancestries is different ([Spanakis & Golden, 2013](#)).

Therefore, the MAGIC investigators incorporated large-scale trans-ancestry meta-analyses of GWAS to detect novel glycaemic trait-associated loci, investigate the genetic scores across ancestries, compare and contrast the genetic architecture of these glycaemic traits to elucidate their underlying biological pathways in T2D.

Initially, MAGIC collaborators explored several phenotypes, including the most commonly used glycaemic traits (FG, FI, 2hG, and HbA1c), and analysed a more complex meta-analysis of insulin measurements, secretion, and sensitivity. Our group contributed to the meta-analysis of the two traits: FG and FI.

The additional value of using population diversity in genetic exploration was evaluated by this meta-analysis study. Therefore, to perform fine-mapping, the portability of loci and PRS across ancestries should be estimated; uncover variations in effect allele frequency (EAF), effect size, and LD across different populations; and make comparisons the genetic architecture of these glycaemic traits, then to elaborate their underlying biology and generate knowledge into T2D-related pathophysiological pathways.

The trans-ancestry and European single-ancestry meta-analysis in this study identified more novel loci compared to the previous studies ([Horikoshi et al., 2015](#); [Wheeler, Marenne, et al., 2017](#)). It identified 242 loci ( $P < 5 \times 10^{-8}$ ), 99 novel glycaemic trait loci, and 24 discovered due to diverse ancestry data, 80% without indication of inter-ancestry heterogeneity. Analyses with equal sample size, excluding individuals of non-European ancestry, resulted in 24% fewer new loci, showing the benefit of incorporating trans-ancestry meta-analysis. The genetic score then was calculated based on effect sizes at European single-ancestry index variants ( $P < 5 \times 10^{-8}$ ), after LD

pruning ( $r^2 < 0.1$ ), and assessed its utility for predicting trait variance explained in other ancestries

My contribution to this project in this study is to assess the extent to which variance is explained for these glycemic traits (especially FG and FI) by genetic factors, using a trans-ethnic study design and including individuals of both NFBC1966 and NFBC1986 cohorts. As adjusting for body mass index (BMI) provided similar results for FG and 2hGlu, but aided in new locus discovery for FI ([Scott et al., 2012](#)), this study also used analyses for FG and FI adjusted by BMI.

## 4.2. Study populations

MAGIC aggregated all individual GWAS contributed to this consortium in up to 281,416 individuals having no diagnosis of T2D with 30% of non-European ancestry (13% 58 East Asian, 7% Hispanic, 6% African-American, and 3% South Asian, and 2% sub-Saharan African). Cohorts that have contributed to the GWAS meta-analysis within MAGIC are described below:

- Amish Studies
- Atherosclerosis Risk in Communities Study (ARIC)
- The Avon Longitudinal Study of Parents and Children (ALSPAC)
- The Avon Longitudinal Study of Parents and Children (ALSPAC)
- The Microisolates in South Tyrol Study (MICROS)
- Netherlands Tweelingen Register (NTR)
- The NHLBI Family Heart Study
- **Northern Finland Birth Cohort 1966 and 1986 Studies** (NFBC1966 and NFBC1986)

- Baltimore Longitudinal Study of Aging (BLSA)
- British Women's Heart and Health Study (BWHHS)
- The Botnia Study
- The Busselton Health Study
- Caerphilly Prospective Study
- Cardiovascular Health Study
- The CoLaus Study
- The Croatian Study (CROAS)
- Diabetes Genetics Initiative
- DIAbetes GENetic Study (DIAGEN)
- deCODE Study
- DEsIR Study
- The Ely Study
- The Erasmus Rucphen Family Study
- Exeter Family Study of Childhood Health
- The Fenland Study
- Finland-United States Investigation of NIDDM Genetics Study (FUSION)
- Framingham Heart Study (FHS)
- MRC National Survey of Health and Development (NSHD)
- The Orkney Complex Disease Study (ORCADES)
- Oxford Biobank
- Partners/Roche
- Procardis
- Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS)
- Relationship Between Insulin Sensitivity and Cardiovascular Risk Study (RISC)
- The Rotterdam Study
- The SardiNIA Project
- The Segovia Study
- SORBS
- The STANISLAS Family Study (SFS)
- SUVIMAX
- Twins UK
- Umeå
- Uppsala Longitudinal Study of Adult Men (ULSAM)

- French Obese Adults Cohort
- Gene-Diet Attica Investigation (GENDAI)
- GenomEUtwin
- The GEMS Study
- The Greek Health Randomised Aging Study (GHRAS)
- Gene-Lifestyle interactions And Complex traits Involved in Elevated disease Risk Study (GLACIER)
- Whitehall II (also known as the Stress & Health Study)
- HEALTH 2000: Health and Functional Capacity of Finns - a national health survey
- The Hertfordshire Cohort
- The InCHIANTI Study
- The Inter99 Study
- Cooperative Health Research in the Region of Augsburg, Southern Germany (KORA)
- MesyBePo
- METSIM
- Haguenau

We contributed with NFBC1966 and NFBC1986 cohorts in genome-wide meta-analyses of glycaemic traits ancestry-specific and transethnic genome-wide meta-analysis summary statistics association with FG and FI. Therefore, we used the same individuals in both cohorts to calculate polygenic risk score (PRS) or genetic risk score (GRS) for variants detected in the meta-analysis GWAS associated with FG and FI phenotypes to estimate the variance explained by the traits associated with loci.

### 3.2.1. Genetic data

The genetic data used in this analysis are described in section [3.1.1.1 Genetic data](#).

We used 5,402 individuals from the NFBC1966 and 3,743 individuals from the NFBC1986 as only these numbers of individuals in both cohorts have genetic data.

In NFBC1986, I used the individuals genotyped GWAS data on the Metabochip platform, as these individuals were used in the primary meta-analysis of GWAS within MAGIC and at that time metabochip was used for genotyping before GWAS chip.

### **3.2.2. Phenotypic data**

In both NFBC1966 and NFBC1986, the meta-analysis was performed for FG and FI. The glucose measurement in both cohorts was made in the blood, which is lower than in plasma; therefore, it was multiplied by 1.13 for adjustment. All individuals with diabetes (T2D, T1D) from data sets that have status as 'diagnosed' or on diabetes treatment or fasting plasma glucose (FPG) >7 mmol/L were excluded. No imputation of missing traits and no outlier exclusions were applied. FI properties were not changed; however, natural logarithmic conversion of FI was performed.

## **4.3.Methods**

### **3.3.1. Variance explained by polygenic risk score (PRS) calculation**

In a sample from NFBC1966 and NFBC1986, our team previously performed two individual GWAS to contribute to the overall meta-analysis. Then I assessed the variance explained by genetic scores of glycaemic trait-associated SNPs constructed from the signals identified in several cohorts' meta-analysis under MAGIC investigators. The various steps of this study are detailed below.

#### *GWAS by individual studies*

Each of the MAGIC collaborators previously conducted an individual GWAS of the glycaemic traits (FG, FI, 2hG, and HbA1c) before meta-analysis. We contributed only for FG and FI traits as we only have these traits in the NFBC cohorts. Imputation of GWAS cohorts had been performed using the integrated reference panels from the March 2012 release of the 1KGP as described in section [3.1.1.1 Genetic data](#). After imputation, genome-wide association testing of the glucose and insulin-related traits with each directly genotyped or imputed variant were performed.

Each cohort was responsible for their sample and variant QC, then applied their filters as appropriate. Standard QC protocols must remove genetic data with extreme heterozygosity, low call rate, and extreme outlying samples from ethnic ancestry. The data was subsequently integrated using meta-analysis summary statistics in the European ancestry (EA) population within MAGIC and the MetaboChip-targeted variants to increase the associations.

#### *Ancestry-specific and trans-ethnic meta-analysis*

The meta-analysis of glycaemic traits using summary statistics from several studies described in section [4.2](#) was conducted by selected MAGIC investigators. We run one-ancestry-specific meta-analysis or European ancestry (EA) using METAL ([Willer, Li, & Abecasis, 2010](#)) and trans-ethnic meta-analysis (EA plus TA) with the ancestry-specific using meta-analysis of transestnic association studies (MANTRA) ([Morris, 2011](#)) for FG and FI.

#### *Identification of signals*

From the meta-analysis, 394 SNP which are associated with FG, FI, 2hG, and HbA1c that reached genome-wide significance ( $P < 5 \times 10^{-8}$ ) were identified using Genome-

wide Complex Trait Analysis (GCTA) ([Yang et al., 2011](#)), which takes LD between SNPs into account.

We received an SNP list from the central analysis group containing all the SNPs representing signals across any of the phenotypes analysed in this meta-analysis. The GWAS meta-analysis results identified 81 and 43 significant independent associations for FG and FI, respectively. Three variant files with the corresponding ancestry specific estimates for each phenotype were provided by MAGIG correspond to the cohorts involved in the meta-analysis: European Ancestry only (EA only), EA plus trans-ethnic (EA plus TE), and MANTRA analysis plus EA (Complete EA plus TE). The effects sizes (beta values) used for the PRS calculation in the NFBCs were adjusted to obtain unbiased estimates since the NFBC samples were included in the original GWAS. The adjustment was performed according to the method published previously ([Nolte et al., 2017](#)). It is because the cohort for the study described by variance should be independent of the sample of the GWAS. In this case, however, this was not possible, as most of the cohorts with the genotypic and phenotypic data needed for the study were included in the GWAS discovery. Therefore, for the contribution of each group, it is necessary to change the effect sizes (betas) from the GWAS, giving sets of cohort-specific effect sizes which were then utilised to produce the PRS. This modification includes recalibrating the effect sizes and standard errors of the variant using inverse versions of the formula for meta-analysis of inverse-variance fixed effects. Independent of whether the recalculated p-value met the significance threshold used in GWAS, all variants in the original list for that ancestry and trait were preserved; thus, there was no change in the list of variants contributing to GS in each cohort (subject to exclusions made during quality control at the cohort level).

### *Quality control (QC) for variants*

The imputation scores for all variant lists for imputed variants and call rates for genotyped variants were checked. I excluded imputed variants if the imputation quality measured by  $R^2 < 0.4$  for minimac imputation server or by  $INFO < 0.4$  for IMPUTE/Sanger imputation server. I also excluded genotyped variants if the call rate was  $< 95\%$ . In all subsequent analyses, I only included variants that passed imputation quality or call rate threshold. For NFBC1966, I also excluded individuals having T1D and T2D, and  $FPG > 7$  mmol. For NFBC1986, I excluded individuals having T1D and FPG glucose  $> 7$  mmol.

### *Calculation of the PRS from the identified signals*

The PRS was predicted using pruning and thresholding, where a reduced collection of SNPs associated with FG and FI is recognised after LD pruning ( $r^2 < 0.1$ ). LD was estimated from the collected cohort pairwise LD information, where available, or the European samples in 1000G Phase 3. The PRS then was calculated per trait (FG and FI) and ancestry, for example, for FG adjusted to BMI: European ancestry (EA only) signals, EA signals plus Trans-ethnic signals (EA plus TE), and EA signals plus all signals from the MANTRA analysis (Complete EA plus TE). Therefore, the analysis was run three times per trait, once with each file. The calculations were performed in PLINK v1.90b4.1 64-bit (30 Mar 2017) ([Purcell et al., 2007](#)) using the `--score` function and dosage format genotype data. (Chang et al., 2015). The PLINK commands using `--sum` and `--double-dosage` were used to multiply the betas by diploid allele counts and summed across all loci. Betas used were either those taken directly from the GWAS results (in the case of non-European ancestry cohorts and all independent

cohorts) or those adjusted (for European ancestry cohorts included in the discovery meta-GWAS).

### *Variance explained analyses*

Finally, I calculated the variance explained by merging each PRS with the phenotype data and fit a linear model in R v.3.3. I used raw (untransformed) trait values for FG and log-transformed trait values for FI. I fitted the model with (i) only the PRS; (ii) with the PRS plus covariates to include BMI and relevant study-specific covariates, e.g., sex, and PCs (in NFBC1966) and MDS (in NFBC1986) and (iii) with covariates only. I did not adjust with age as all the individuals are of the same age.

## **4.4. Results**

Out of the 394 signals to be used for the PRS calculation in NFBC1966, six SNPs were lost due to low imputation quality. However, these SNPs were not within the FG and FI lists and hence did not affect the analyses. In NFBC1986, eight SNPs were lost as this data does not have chromosome X imputation; they were not in the imputed data, since almost all of them are INDELs (missing for unknown reason); one SNP was lost due to low imputation quality, which was contributing to the FG signal.

The variance explained results in this study was reported in  $R^2$  output, which estimates how well one can predict FG or FI trait in NFBC1966 or NFBC1986 sample. The results of partial  $R^2$  output obtained from linear models for NFBC1966 and NFBC1986 after adjustment of BMI and principal components for FG show almost similar results either using data with trans-ethnic or without trans-ethnic variants as seen in [Table 4.1](#) and [Table 4.2](#). However, using the trait prediction using the EA plus transethnic variants has slightly greater variance explained compared to the EA only data. In our total

sample, PRS explained about 4.3% of the variance in FG trait prediction in NFBC1966, while in NFBC1986, it was slightly lower, 1.3%. For FI, the figures were much lower: 0.7 – 0.9% in NFBC1966, and 0.09 - 0.27 % in NFBC1986. The variance explained is similar when using European ancestry and trans-ethnic results, and European ancestry and MANTRA analysis. Overall, the variance explained was smaller in FI compared to FG in both cohorts.

Table 4.1 Variance explained results for NFBC1966

| PRS                       | Model                  | FG adjusted by BMI |               |                                | FI adjusted by BMI |               |                                |
|---------------------------|------------------------|--------------------|---------------|--------------------------------|--------------------|---------------|--------------------------------|
|                           |                        | N individuals      | N SNPs in PRS | Adjusted R <sup>2</sup>        | N individuals      | N SNPs in PRS | Adjusted R <sup>2</sup>        |
| <b>EA_only</b>            | Covariates only        | 5402               | 81            | 8.63x10 <sup>-02</sup>         | 5402               | 43            | 2.84 x10 <sup>-01</sup>        |
|                           | PRS only               |                    |               | <b>4.27</b> x10 <sup>-02</sup> |                    |               | <b>6.83</b> x10 <sup>-03</sup> |
|                           | Covariates plus PRS    |                    |               | 1.27 x10 <sup>-01</sup>        |                    |               | 2.92 x10 <sup>-01</sup>        |
|                           | Partial R <sup>2</sup> |                    |               | 4.52 x10 <sup>-02</sup>        |                    |               | 1.15 x10 <sup>-02</sup>        |
| <b>EA_plusTE</b>          | PRS only               |                    | 105           | <b>4.08</b> x10 <sup>-02</sup> |                    | 61            | <b>8.31</b> x10 <sup>-03</sup> |
|                           | Covariates plus PRS    |                    |               | 1.27 x10 <sup>-01</sup>        |                    |               | 2.94 x10 <sup>-01</sup>        |
|                           | Partial R <sup>2</sup> |                    |               | 4.44 x10 <sup>-02</sup>        |                    |               | 1.42 x10 <sup>-02</sup>        |
| <b>CompleteTE plus EA</b> | PRS only               |                    | 110           | <b>3.85</b> x10 <sup>-02</sup> |                    | 72            | <b>9.24</b> x10 <sup>-03</sup> |
|                           | Covariates plus PRS    |                    |               | 1.25 x10 <sup>-01</sup>        |                    |               | 2.96 x10 <sup>-01</sup>        |
|                           | Partial R <sup>2</sup> |                    |               | 4.26 x10 <sup>-02</sup>        |                    |               | 1.67 x10 <sup>-02</sup>        |

Table 4.2 Variance explained results for NFBC1986

| PRS            | Model                  | FG adjusted by BMI |               |                                | FI adjusted by BMI |               |                                |
|----------------|------------------------|--------------------|---------------|--------------------------------|--------------------|---------------|--------------------------------|
|                |                        | N individuals      | N SNPs in PRS | Adjusted R <sup>2</sup>        | N individuals      | N SNPs in PRS | Adjusted R <sup>2</sup>        |
| <b>EA_only</b> | Covariates only        | 3743               | 80            | 2.91 x10 <sup>-02</sup>        | 3743               | 43            | 1.46 x10 <sup>-01</sup>        |
|                | PRS only               |                    |               | <b>1.34</b> x10 <sup>-02</sup> |                    |               | <b>1.72</b> x10 <sup>-03</sup> |
|                | Covariates plus PRS    |                    |               | 4.31 x10 <sup>-02</sup>        |                    |               | 1.48 x10 <sup>-01</sup>        |
|                | Partial R <sup>2</sup> |                    |               | 1.48 x10 <sup>-02</sup>        |                    |               | 2.63 x10 <sup>-03</sup>        |

|                           |                        |  |            |                                          |  |           |                                          |
|---------------------------|------------------------|--|------------|------------------------------------------|--|-----------|------------------------------------------|
| <b>EA_plusTE</b>          | PRS only               |  | 105        | $1.70 \times 10^{-02}$                   |  | 61        | $9.49 \times 10^{-04}$                   |
|                           | Covariates plus PRS    |  |            | $4.59 \times 10^{-02}$                   |  |           | $1.47 \times 10^{-01}$                   |
|                           | Partial R <sup>2</sup> |  |            | $1.76 \times 10^{-02}$                   |  |           | $1.80 \times 10^{-03}$                   |
| <b>CompleteTE plus EA</b> | PRS only               |  | <b>110</b> | <b><math>1.67 \times 10^{-02}</math></b> |  | <b>72</b> | <b><math>2.73 \times 10^{-03}</math></b> |
|                           | Covariates plus PRS    |  |            | $4.54 \times 10^{-02}$                   |  |           | $1.48 \times 10^{-01}$                   |
|                           | Partial R <sup>2</sup> |  |            | $1.88 \times 10^{-02}$                   |  |           | $2.83 \times 10^{-03}$                   |

Comparison of PRS derived using EA variants versus EA plus transethnic variants can be observed from

**Figure 4.1.** It shows that the trait prediction using EA plus transethnic variants has slightly greater variance explained compared to EA only.

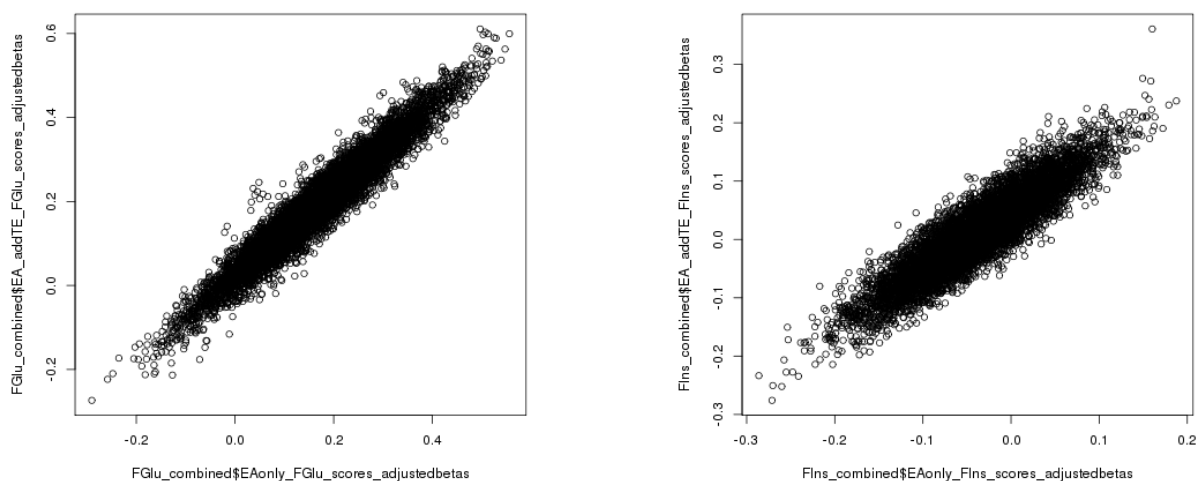


**Figure 4.1** Comparison of PRS derived using EA variants versus EA plus transethnic variants

The variance explained for FG and FI in NFBC1966 and NFBC1986 is similar to unpublished results from the other cohorts analysed by our collaborators as part of the internal report within MAGIC, as shown in [Table 4.3](#).



Table 4.3 Results (adjusted R<sup>2</sup>) from testing cohorts. Covariates as fitted in GWAS (age, age<sup>2</sup>, BMI, gender, PCs (if needed)).

| Ancestry            | Model                  | NFBC1966<br>(n=5,402)        |                              | NFBC1986<br>(n=3,743)        |                              | ALSPAC<br>(n=7,921)          |                              | RAINE<br>(n=553)             |                               |
|---------------------|------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|-------------------------------|
|                     |                        | Fasting glucose              | Fasting insulin (ln)         | Fasting glucose              | Fasting insulin (ln)         | Fasting glucose              | Fasting insulin (ln)         | Fasting glucose              | Fasting insulin (ln)          |
| EA only             | Covariates only        | 8.63x10 <sup>-02</sup>       | 2.84x10 <sup>-01</sup>       | 2.91x10 <sup>-02</sup>       | 1.46x10 <sup>-01</sup>       | 4.60x10 <sup>-02</sup>       | 2.21x10 <sup>-01</sup>       | 8.00x10 <sup>-02</sup>       | 1.87x10 <sup>-01</sup>        |
|                     | <b>PRS only</b>        | <b>4.27x10<sup>-02</sup></b> | <b>6.83x10<sup>-03</sup></b> | <b>1.34x10<sup>-02</sup></b> | <b>1.72x10<sup>-03</sup></b> | <b>5.10x10<sup>-02</sup></b> | <b>1.30x10<sup>-02</sup></b> | <b>3.40x10<sup>-02</sup></b> | <b>-4.00x10<sup>-04</sup></b> |
|                     | PRS plus covariates    | 1.27x10 <sup>-01</sup>       | 2.92x10 <sup>-01</sup>       | 1.27x10 <sup>-01</sup>       | 2.92x10 <sup>-01</sup>       | 9.90x10 <sup>-02</sup>       | 2.35x10 <sup>-01</sup>       | 1.22x10 <sup>-01</sup>       | 1.90x10 <sup>-01</sup>        |
|                     | Partial R <sup>2</sup> | 4.52x10 <sup>-02</sup>       | 1.15x10 <sup>-02</sup>       | 4.52x10 <sup>-02</sup>       | 1.15x10 <sup>-02</sup>       | 5.60x10 <sup>-02</sup>       | 1.90x10 <sup>-02</sup>       | 4.70x10 <sup>-02</sup>       | 5.00x10 <sup>-03</sup>        |
| EA plus TE          | <b>PRS only</b>        | <b>4.08x10<sup>-02</sup></b> | <b>8.31x10<sup>-03</sup></b> | <b>1.70x10<sup>-02</sup></b> | <b>2.73x10<sup>-03</sup></b> | <b>5.30x10<sup>-02</sup></b> | <b>1.40x10<sup>-02</sup></b> | <b>4.00x10<sup>-02</sup></b> | <b>2.00x10<sup>-03</sup></b>  |
|                     | PRS plus covariates    | 1.27x10 <sup>-01</sup>       | 2.94x10 <sup>-01</sup>       | 1.27x10 <sup>-01</sup>       | 2.94x10 <sup>-01</sup>       | 1.01x10 <sup>-01</sup>       | 2.36x10 <sup>-01</sup>       | 1.26x10 <sup>-01</sup>       | 1.91x10 <sup>-01</sup>        |
|                     | Partial R <sup>2</sup> | 4.44x10 <sup>-02</sup>       | 1.42x10 <sup>-02</sup>       | 4.44x10 <sup>-02</sup>       | 1.42x10 <sup>-02</sup>       | 5.80x10 <sup>-02</sup>       | 1.90x10 <sup>-02</sup>       | 5.10x10 <sup>-02</sup>       | 6.00x10 <sup>-03</sup>        |
| Complete EA plus TE | <b>PRS only</b>        | <b>3.85x10<sup>-02</sup></b> | <b>9.24x10<sup>-03</sup></b> | <b>1.67x10<sup>-02</sup></b> | <b>2.73x10<sup>-03</sup></b> | -                            | -                            | -                            | -                             |
|                     | PRS plus covariates    | 1.25x10 <sup>-01</sup>       | 2.96x10 <sup>-01</sup>       | 4.54x10 <sup>-02</sup>       | 1.48x10 <sup>-01</sup>       | -                            | -                            | -                            | -                             |
|                     | Partial R <sup>2</sup> | 4.26x10 <sup>-02</sup>       | 1.67x10 <sup>-02</sup>       | 1.88x10 <sup>-02</sup>       | 2.83x10 <sup>-03</sup>       | -                            | -                            | -                            | -                             |

#### 4.5. Discussion and conclusions

Recently, the calculation of PRS for trait prediction across different studies has been an ongoing debate, especially across various populations ([Duncan et al., 2019](#); [Khera et al., 2018](#); [Lambert, Abraham, & Inouye, 2019](#)). In this study, I estimate the utility of weighted PRS constructed from European ancestry effect sizes to predict glycemic traits variance explained.

The analysed results were obtained quantitatively by summing the contribution of all identified SNPs as variance explained in percentage, estimated using our genotype data: NFBC1966 and NFBC1986, among other studies from our collaborators. I found a similar variance explained in each study contributing to the MAGIC PRS analysis: NFBC1966 and NFBC1986, although NFBC1986 is less powered than NFBC1966. I also found a similar variance explained in my studies compared to the other studies conducted by other investigators within MAGIC, i.e., RAINE and ALSPAC.

By comparing results from each of NFBCs, I found that FG has a slightly bigger variance explained compared to FI. This finding can be related to the number of variants included in the analysis and the average effect size of each variant. FI has a smaller number of variants with smaller effect sizes. The variance explained by FG or FI is similar between all studies.

In addition, I found lower variance explained to be related to age, as demonstrated by the results from the younger cohort of NFBC1986 for both phenotypes. We could expect less variability in NFBC1986 than in NFBC1966 as the participants were younger and in the healthy range still, as participants in NFBC1986 were examined at the age of 16 compared to 31-year-old participants in NFBC1966. Age adjustment has

not been performed in the calculation as all the participants in both cohorts had the same age.

In NFBC1966, the total variance explained by FG reached 4.3% compared to the previous study that only accounted for 1.5% attributable to four signals ([Prokopenko et al., 2009](#)), as the number of signals in this current study has largely increased compared to the previous one from four to 81 SNPs. In Europeans, these estimates represent an improvement (i.e., more variance explained) relative to previous estimates, derived using similar methodology with fewer contributing SNPs, of 2.8% for FG ([Nolte et al., 2017](#)). Although, variance explained estimates by Scott et al. ([Scott et al., 2012](#)) of 4.8%(FG), 1.2% FI are beyond this study estimates. I conclude that this is probably partly due to a disparity in statistical methods.

In this study, I used the pruning and thresholding method to estimate the PRS as instructed by MAGIC's investigator. This modest method may involve false positive variants in the reduced set of SNPs, and raw estimates of impact sizes from them may be susceptible to selection bias. In addition, it involves checking over a number of thresholds for p-values, which are sometimes randomly picked. A positive bias can also be subject to the forecast error calculated from the optimized threshold.

The PRS constructed in this analysis was calculated from a weighted sum of allele counts, the weights being given from univariate regression tests by log odds ratios or linear regression coefficients. It is reasonable to ask whether we could strengthen the process of weighting to increase predictive power. An analysis showed that the Bayes strategy enhances the estimation of genomic risk and retrieves the underlying impact sizes by using them as weights to build PRS ([So & Sham, 2017](#)). In the next study,

we could probably consider using this advanced method to improve the variance explained prediction.

Although using the trans-ethnic approach has allowed us to discover more genetic variants contributing to trait variability, and increasing variance explained to about 4.3% by FG, it is still much lower than that for other related phenotypes, such as T2D, which has a variance explained of 18% by currently known genetic variants (Mahajan et al., 2018). This suggests that there is still a missing heritability issue. This is the case for many other complex traits as well. One way forward is to increase the sample size in the future. For example, the T2D study mentioned above increased the sample size up to ~900,000 European participants, detecting >400 distinct signals associated with T2D risk and achieving a variance explained ~18% (Mahajan et al., 2018). Likewise, the recent meta-analysis of GWAS for height and BMI in ~700,000 individuals identified 3,290 and 941 variants at genome-wide significance that explained ~24.6% for height variance and ~6.0% for BMI variance (Yengo et al., 2018). However, traits such as height and BMI are easily measurable, unlike FG and FI that need blood sample collection. Combining data across these traits with T2D status will further elucidate pathways driving normal physiology and pathophysiology and further develop useful predictive scores for disease classification and management (Udler, 2019).

In addition, to suggest increasing the sample size, we can use an alternative approach to boost the power of associations by performing MP-GWAS methodology.

## 5. MP-GWAS of Fatty acid (FA) metabolism

*The main goal of the work described in this chapter is to elucidate the genetic architecture of FA levels using MPA. My work focused on performing conditional MP-GWAS to detect multiple distinct signals from the previously detected loci conducted on MP-GWAS of  $\omega$ -3, 6, 7/9, and other polyunsaturated FAs (otPUFA) by our team.*

### 5.1. Introduction

The number of deaths from CVD is on the rise around the globe. The risk of CVD could be attenuated through favourable effects of diets rich in omega-3 FAs that mildly but significantly decrease heart rate ([Hidayat et al., 2018](#)). Dissecting the molecular mechanisms regulating the levels of omega-3 and other FAs will facilitate the new pharmaceutical treatments for cardiometabolic diseases.

Fatty acids (FAs) are nutrients obtained directly from the diet, such as fish oil, seeds, nuts, or the conversion of plant omega-3 FA. Essential FA, such as omega-3 FA, provides important cellular functions ([Tanaka et al., 2009](#)). The role of FAs in human health and body homeostasis is undisputable. Effects of omega-3 and other FAs on several diseases, including T2D and CVD, are well-established in epidemiological studies ([Poudyal, Panchal, Diwan, & Brown, 2011](#)). Pathophysiology of inflammation, hypertension, dyslipidaemia, and increased adiposity is influenced by the consumption of nutrients containing FAs and their metabolism.

Omega-3 and other FAs represent only a fraction of a rather wide grouping of blood lipids typically used in clinical practice, such as triglycerides (TG), total cholesterol (TC), and high- and low-density lipoprotein cholesterol levels (HDL-C and LDL-C, respectively). Therefore, the 150 blood lipid-associated genetic loci, identified through genome-wide association studies (GWAS), do not necessarily affect levels of FAs in

particular. Thus far, only a few genomic regions have been associated with FAs through GWAS, with the *FADS1/2* locus showing the most robust evidence for association (Dorajoo et al., 2015; Guan et al., 2014; Lemaitre et al., 2011; Tanaka et al., 2009). In addition, GWAS dissecting a wide range of blood metabolites, including FAs, have not proven particularly informative regarding the genetic architecture of FAs (Draisma et al., 2015; Kettunen et al., 2012; Shin et al., 2014; Suhre et al., 2032). This is likely due to lack of power, given the nature of those study designs with analysis of hundreds of individual phenotypes and subsequent correction for multiple testing. MP-GWAS is an innovative approach allowing for a substantial improvement in power for locus discovery through the joint analysis of correlated traits. Our team had previously studied the genetic contribution to FA levels via MP-GWAS of  $\omega$ -3, 6, 7/9, and otPUFA. The meta-analysis from NFBC1966 and NFBC1986 (N=8,004) conducted by our team detected 10 signals associated with FAs ( $P < 5 \times 10^{-8}$ ) at *PCSK9*, *GCKR*, *LPXN*, *FADS1*, *GPR137*, *ZNF259*, *LIPC*, *PDXDC1*, *PBX4*, and *APOE*. The LD structure of the detected loci showed evidence for multiple distinct signals at some of the loci. To dissect whether 10 omega FA loci harbour more than one distinct signal affecting FA levels, we performed subsequent, direct conditional MP-GWAS. Our MP-GWAS dataset comprised up to 4,949 individuals from the NFBC1966 and 3,055 from the NFBC1986, totalling 8,004 individuals with whole-genome genotyping, metabolomics, metabolic trait, and epidemiologic data as described in Chapter 3.

## 5.2.Methods

### 4.2.1. Phenotype imputation

Each of the FAs from Table 3.3 suffered from missing data to a certain degree (mean (range) in NFBC1966: 6.2% (6.0 to 6.8) and in NFBC1986: 14.4% (14.3 to 14.6)).

To reduce bias introduced by missing phenotypic and metabolite data, we imputed these phenotypes using single imputation (SI) implemented on multivariate imputation by chained equations (MICE) package in R ([van Buuren & Groothuis-Oudshoorn, 2011](#); [White, Royston, & Wood, 2011](#)). The approach is implemented in the R package MICE ([van Buuren & Groothuis-Oudshoorn, 2011](#); [White et al., 2011](#)), which was implemented based on conditional distribution. For each phenotype and metabolite, the imputation model included all 124 other metabolites plus other indicators of metabolic health, such as measured fasting glucose, insulin, blood pressure, and body-mass index (BMI). Predictive mean matching (pmm) was chosen for imputing continuous traits in the case of metabolite concentrations. The imputation was iterated five times to obtain a stable dataset ([van Buuren & Groothuis-Oudshoorn, 2011](#); [White et al., 2011](#)).

Individuals who were pregnant, diabetic, on lipid-lowering medication, or non-fasting prior to the clinical examination were excluded. The residuals were calculated for men and women separately before it combined as it used slightly different adjustment. They were calculated by adjusting for BMI, the three first principal components (PCs) derived from genetic data to control for population structure, and – for women – the use of oral contraceptives ([Sabatti et al., 2009](#)). We applied inverse-normal transformation for the metabolite residuals ([Kohl et al., 2012](#); [Servin & Stephens, 2007](#)). The data for men and women were then combined for the association analysis.

#### **4.2.2. Conditional analysis**

Within each locus, I conducted direct conditional multi-phenotype analyses in the two cohorts by regressing the FA phenotypes on the top marker(s) in that locus and using the resulting residuals in a subsequent multi-phenotype analysis within the 500kb

region on each side of the association peak. Finally, I performed a meta-analysis of the NFBC1966 and NFBC1986 cohort results. The workflow for the conditional MP-GWAS analysis can be observed in [Figure 5.1](#).

For the analysis of each identified signals, 1) I regressed out the effect of the most strongly associated SNP from each FA, in addition to BMI and PCs, followed by inverse-normal transformation; 2) I performed the MPA within the 500kb region on each side of association peak, and 3) I meta-analysed the results from the two contributing studies. An additional distinct signal was defined if, upon conditioning, there was still an association signal reaching the genome-wide significance threshold. Where I observed distinct signals, I then conditioned on the original lead SNP and the newly observed distinct SNP(s) to detect further signals. This process was repeated until no additional signal was identified at genome-wide significance.

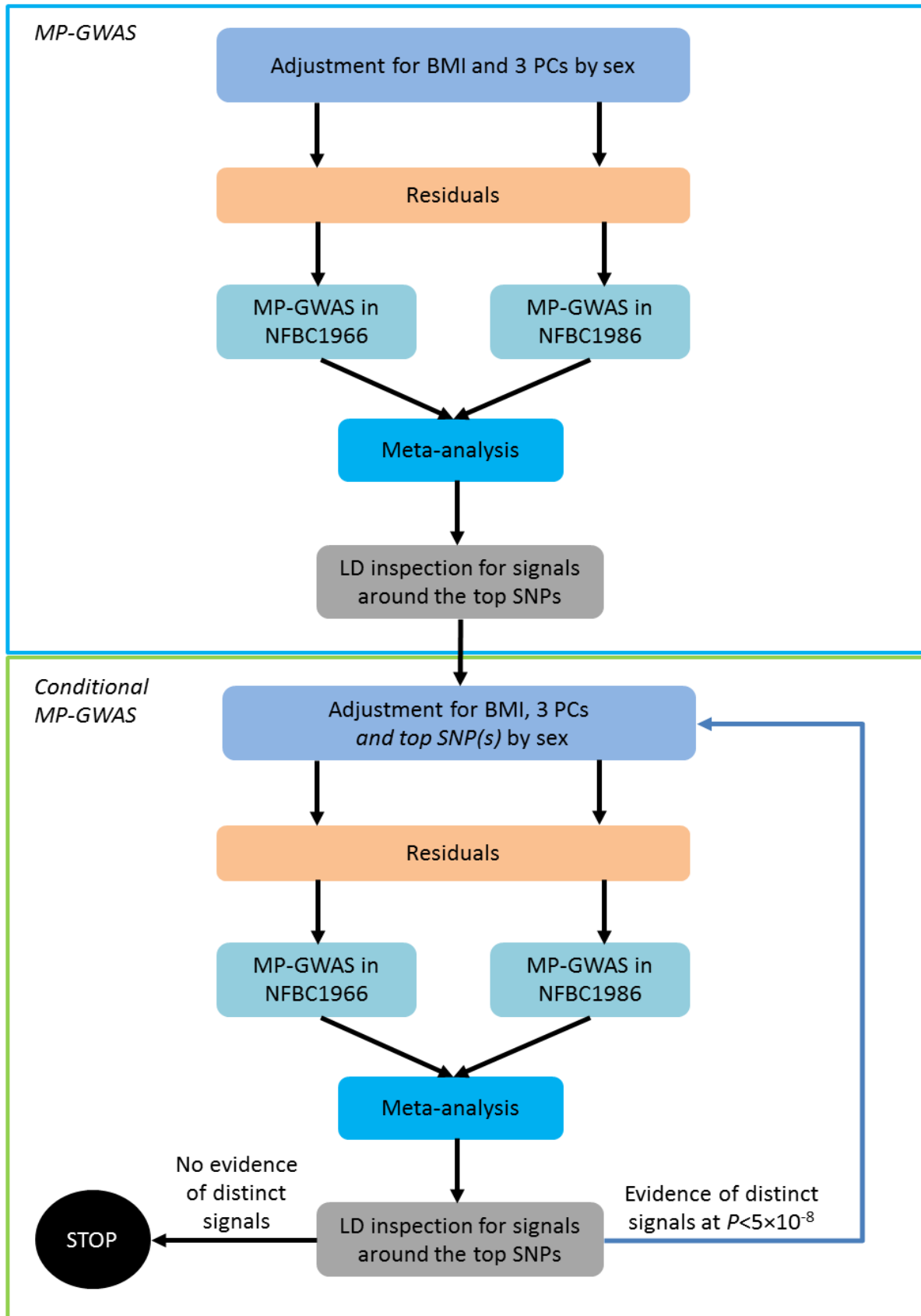


Figure 5.1 The workflow for the conditional MP-GWAS analyses

### 5.3. Results

By the boosted power through phenotype imputation and the MP-GWAS, our group has identified ten common variant loci associated with FAs ( $P < 5 \times 10^{-8}$ ) at *PCSK9*, *GCKR*, *LPXN*, *FADS1/2*, *SUGP1*, *ZNF259*, *LIPC*, *PDXDC1*, *RPS6KA4*, and *APOE*, including two novel ones at *LPXN* (rs188015703,  $P = 2.55 \times 10^{-8}$ ) and *RPS6KA4* (rs3782101,  $P = 5.51 \times 10^{-9}$ ) and eight previously established for blood lipids.

At *LIPC*, analysis through direct conditioning on the lead SNP rs1532085 demonstrated evidence for two other distinct signals in this region, at intronic rs2070895 and rs117376818 variants (Figure 5.5). In addition, rs117376818 is in low LD with other variants in the region and is rare (MAF=0.009 in NFBC1966 and 0.014 in NFBC1986; MAF=0.01 in Europeans); however, its imputation quality was over 0.69 in both cohorts, indicating that this might be a true signal at least in the Finnish population.

At *SUGP1*, the LD around the identified lead variant rs8100204 is extensive (Figure 5.6). Within this locus, the direct conditional analysis showed suggestive evidence for another distinct signal, rs12463074.

Table 5.1 The distinct signals at the identified loci for FAs.

| Chr                       | 1                 | 2                | 11                 | 11               | 11                | 11               | 15                 | 16                | 19                 | 19              |
|---------------------------|-------------------|------------------|--------------------|------------------|-------------------|------------------|--------------------|-------------------|--------------------|-----------------|
| Top position              | 1:55505647        | 2:27730940       | 11:58221445        | 11:61588305      | 11:64131280       | 11:1166448917    | 15:58683366        | 16:15172118       | 19:19393714        | 19:45412079     |
| Identified locus          | <i>PCSK9</i>      | <i>GCKR</i>      | <i>LPXN</i>        | <i>FADS1/2/3</i> | <i>RPS6KA4</i>    | <i>ZNF259</i>    | <i>LIPC</i>        | <i>PDXDC1</i>     | <i>SUGP1</i>       | <i>APOE</i>     |
| Top signal                | <i>rs11591147</i> | <i>rs1260326</i> | <i>rs188015703</i> | <i>rs174564</i>  | <i>rs10066207</i> | <i>rs3782101</i> | <i>rs1532085</i>   | <i>rs11644601</i> | <i>rs143988316</i> | <i>rs429358</i> |
| Second independent signal | -                 | -                | -                  | <i>rs174590</i>  | -                 | -                | <i>rs2070895</i>   | -                 | <i>rs182611493</i> | -               |
| Third independent signal  | -                 | -                | -                  | -                | -                 | -                | <i>rs117376818</i> | -                 | -                  | -               |



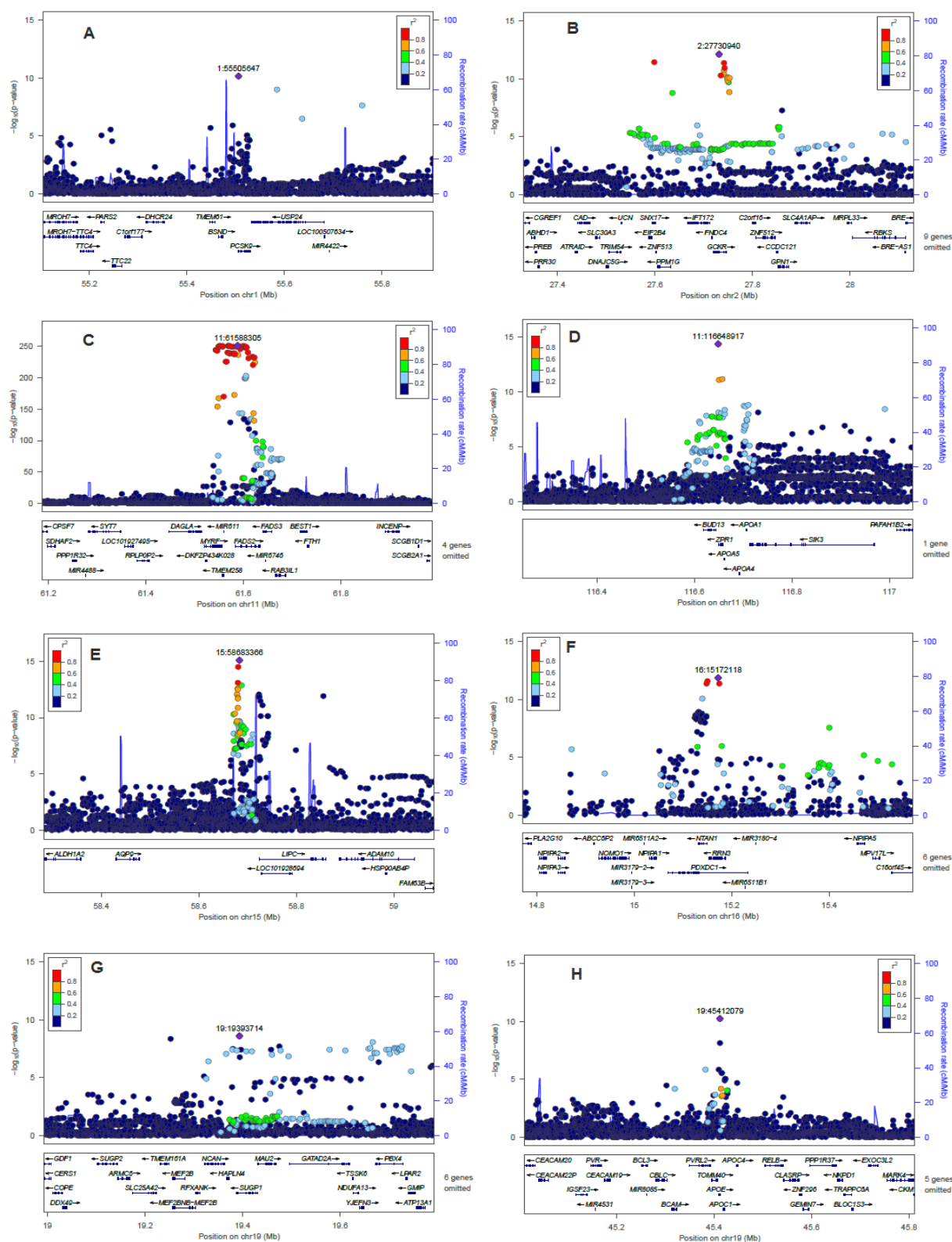


Figure 5.4 Regional plots of previously established lipid loci attaining statistical significance of  $P < 5 \times 10^{-8}$  in our MP-GWAS meta-analysis of FAs. A) PCSK9, B) GSKR, C) FADS1/2, D) ZNF259, E) LIPC, F) PDXDC1, G) SUGP1, and H) APOE.

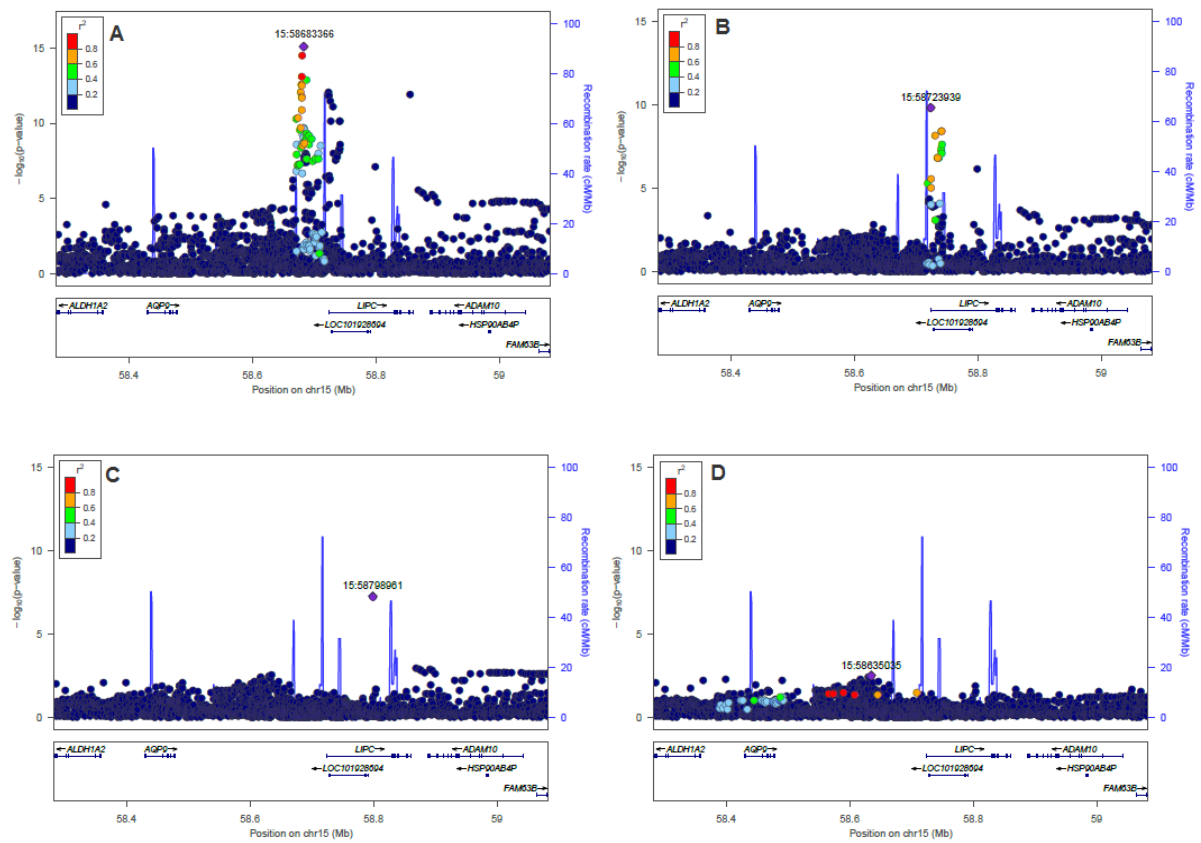


Figure 5.5 Regional plots of the LIPC region before and after conditional MP-GWAS. A) unconditional, B) conditional on rs1532085, C) conditional on rs1532085 and rs2070895, and D) conditional on rs1532085, rs2070895 and rs117376818.

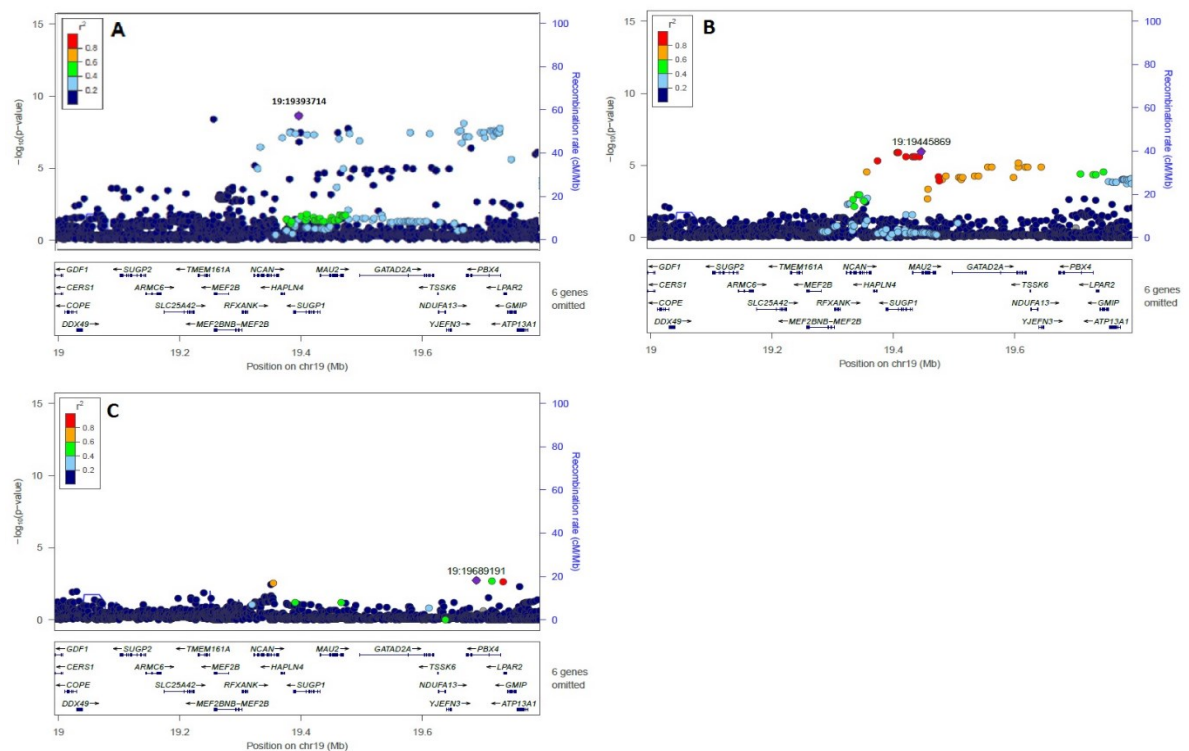


Figure 5.6 Regional plots of the SUGP1 region before and after conditional MP-GWAS. A) unconditional, B) conditional on rs8100204, C) conditional on rs8100204 and rs12463074.

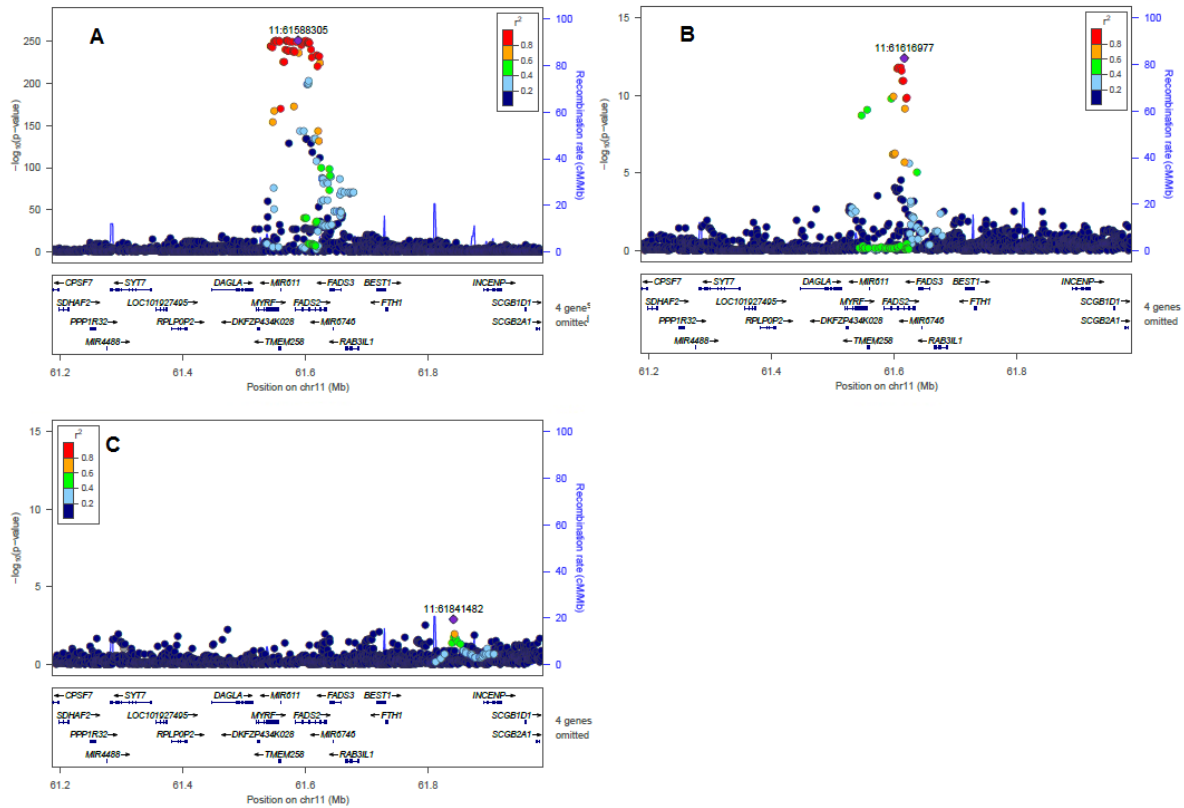


Figure 5.7 Regional plots of the *FADS1/2/3* region before and after conditional MP-GWAS. A) unconditional, B) conditional on rs174564, C) conditional on rs174564 and rs174590.

## 5.4. Discussion and conclusions

We proposed multi-phenotype GWAS (MP-GWAS) to boost the statistical power and performed MP-GWAS of fatty acids in NFBC1966 (N=4955) and replication in NFBC1986 (N=2687) to investigate fatty-acid metabolisms. The meta-analysis conducted by our team detected 10 signals associated with FAs ( $P < 5 \times 10^{-8}$ ) at *PCSK9*, *GCKR*, *LPXN*, *FADS1*, *GPR137*, *ZNF259*, *LIPC*, *PDXDC1*, *PBX4*, and *APOE*. For subsequent analysis, I proposed a new direct conditional analysis method within MP-GWAS, which detected multiple distinct signals within these loci.

I have found suggestive indications of several distinct signals at some of the loci by studying the LD structure of these loci and the LD between most statistically associated variants within each locus.

Meta-analysis findings proposed two different signals in *FADS1*, *PBX4*, and three distinct signals in *LIPC*, while for *PCSK9*, *GCKR*, *LPXN*, *GPR137*, *ZNF259*, *PDXDC1*, and *APOE*, the conditional analysis showed evidence for only one signal affecting FA level for each locus. Common variants from the established lipid loci *FADS1/2/3* and *PDXDC1* showed independent effects on FA levels upon direct conditioning on clinically used blood lipids, strengthening the evidence for their role in FA metabolism.

Our group reported a multi-phenotype GWAS of common variants for FA metabolites. We implemented multi-phenotype imputation using SI and proposed a rigorous approach for phenotype selection for the main MP-GWAS and detected novel FA loci. I performed subsequent conditional MP-GWAS inside our group analysis, and most importantly, detected novel additional/distinct signals at the *FADS1/2/3* locus. The methods used in this analysis show the advantage of MP-GWAS with subsequent conditional MP-GWAS over commonly used SP-GWAS to detect novel additional/distinct signals with multi-phenotype effects in a small to moderate sample size.

The conditional MP-GWAS used in this thesis was the first designed and implemented in the MP-GWAS framework up to date. This method can also be used directly with individual-level data, where the reference sample is the whole cohort. This gives benefit compared to the usual stepwise conditional analysis that usually used in the conditional meta-analysis SP-GWAS using summary statistics, such as that implemented in GCTA-COJO ([Yang et al., 2011](#)). This stepwise forward method is

known to be overly “greedy”; it is prone to eliminating useful predictors that happen to be correlated with selected predictors ([Efron et al., 2004](#)). This implies that, as a result of their LD with observed variants, the stepwise conditional analysis can neglect certain informative variants. More variants could be found where the p-value threshold of discovery in GCTA-COJO is less strict. However, there is a risk of generalizing GCTA-COJO as a fixed-effect model selection technique, particularly with the inclusion of too many predictors in the model when selecting the higher p-value threshold.

There is theoretical and empirical evidence that simultaneous modelling of multiple predictors with penalization provides a better variable selection procedure than stepwise selection ([Bøvelstad et al., 2007](#)).

The conditional analysis proposed in this thesis made it possible to improve the genetic architecture of the identified sites affecting FA metabolism. This novel methodology is applicable to a series of genomic loci that exhibit signals collocated for different features and therefore incorporate stimulating multi-phenotype effects.

Although phenotype imputation using SI has proven adequate to detect novel signals in this study, imputing the data with an advanced imputation method would be suggested for future analysis. This is because SI does not consider the uncertainty of the imputation; therefore, multiple imputation (MI) is usually suggested to improve accuracy. By generating multiply imputed datasets, MI considers the uncertainty of missing data, which is important when imputing multiple traits to avoid bias ([Schafer, 1999](#)). However, our SCOPA software was not designed for analysing multiply imputed datasets; therefore, we need to extend the applicability of SCOPA software or explore alternative imputation approaches from machine learning that may have better performances than MI.



## 6. Implementation of phenotype imputation methods to the MP-GWAS framework

*This chapter describes the simulation study I conducted to compare seven advanced imputation methods to handle missing phenotype data in the MP-GWAS framework. In our previous analyses within the MP-GWAS framework, we used the simple single imputation approach described in the previous chapter for the FA work. I aimed to further improve the power and accuracy of MP-GWAS analyses by improving the quality of input data by better imputation. Here, I present the simulation study setup, the different imputation methods I compared, and the resulting conclusions drawn from the simulation study.*

### 6.1. Introduction

In our FA analysis in Chapter 5, we already applied MP-GWAS using imputed data with SI imputation. However, SI imputation has limitations in overestimating the correlations but underestimates the variances, as mentioned in section 2.4.1.1. In addition, the estimates are all treated as if they were true, which is never the case. Therefore, MI is proposed to solve SI imputation uncertainty in much statistical software (Yuan, 2010). However, MI is computationally intensive to impute high-dimensional data; moreover, our SCOPA software did not implement several imputed datasets to pooled the estimates into one summary statistic.

Besides MI, I evaluated several imputation methods coming from machine learning approaches. Altogether, I compared SI and MI performance using the two most common approaches, the Bayesian approach (MICE package in R) and the maximum-likelihood approach (AMELIAII package in R), and imputation methods from machine-learning, namely QRILC, k-NN, and RF. The original literature has proposed RF to

handle different types of data without assuming any normality (van Buuren, 2012), while kNN (J. S. Shah et al., 2017) was proposed to impute data without discarding the structure and the correlation of the data. In addition, QRILC (Wei, Wang, Su, Jia, Chen, Chen, et al., 2018), as a left-censored missing imputation, was proposed in the metabolomic studies that usually has MNAR assumption due to undetected small values in the measurement. All of the imputation methods have advantages and disadvantages.

To validate the imputed data of the imputation approaches, I finally compared their performance with complete data and CC analyses within the MP-GWAS framework using simulated data under different scenarios. The best imputation method observed was then intended to be applied to the missing data in real data analysis in the MP-GWAS framework.

## 6.2. Methods

### *Simulated data*

To compare the performance of imputation approaches under the MP-GWAS framework, I simulated genetic data using Hapgen2 (Su, Marchini, & Donnelly, 2011) and created datasets with 5,000, 50,000, and 500,000 individuals, as seen in Figure 6.1.

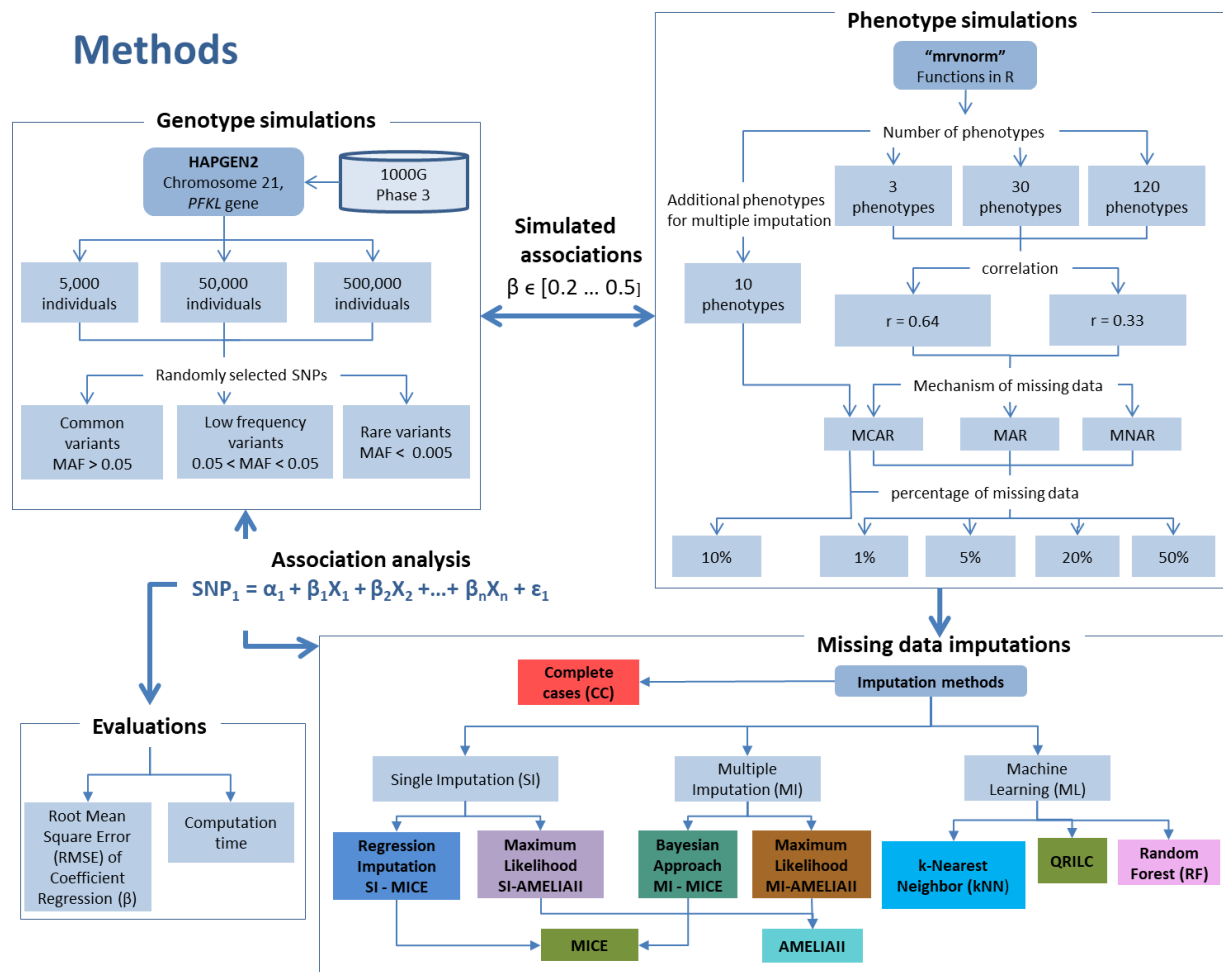


Figure 6.1 Workflow diagram of genotype and phenotype data simulations and missing data imputation methods

These simulated datasets represent the usual numbers of individuals in our sample datasets we used in NFBC1966 (approximately 5,000 individuals) and the numbers of individuals in UK Biobank (approximately 500,000 individuals), while the 50,000 individual data were used in between these number of individuals.

### 6.2.1. Simulation of genetic data

For simplicity, I created the data to a representative gene on chromosome 21: the *PFKL* gene at position 45,719,900 and 45,747,300 base pairs. As all the imputation methods were intended to be implemented in MP-GWAS setting to observe a

complete picture of their relative power, the *PFKL* gene was selected as it represents an average gene with 22 exons and introns, including 143 polymorphic SNPs, of which 39 SNPs are common, and 17 SNPs are rare (Tachmazidou, Morris, & Zeggini, 2013). The simulation of genetic data was performed to create control data of 5000/50,000/500,000 individuals. The resulting .gen file was transformed in R software into a dosage file, derived from the imputation probabilities.

We randomly selected the SNPs with MAF under the categories of common variants ( $MAF > 0.05$ ), low frequency variants ( $0.005 < MAF < 0.05$ ), and rare variants ( $MAF < 0.005$ ).

### 6.2.2. Simulation of phenotype data

The phenotype simulations were done using the “mvnrm” function in R to simulate 3/30/120 of highly ( $r=0.64$ ) and moderately ( $r=0.33$ ) correlated phenotypes. Another 10 phenotypes were created as additional variables in the MI analysis as utilizing these auxiliary variables can improve the MI model to predict the missing data. Conceptually, the variables containing missing data are better predicted when they were regressed to more auxiliary variables. However, taking too many variables can cause biased regression coefficients (Tang & Ishwaran, 2017).

The missing data of the phenotypes used in this analysis were created under the assumptions of MCAR MAR and MNAR, with the proportion of missing data for each mechanism being 1/5/20/50%.

### 6.2.3. Imputation methods

The SI and MI methods were applied using package MICE and AMELIAII in R since they both have different benefits and limitations.

## **MICE**

MICE (Multiple Imputation by Chained Equations) is a library for R to perform MI imputation using the Bayesian approach ([Van Buuren & Groothuis-Oudshoorn, 2011](#)). MICE imputes data variable by variable, which can handle different types of variables, such as continuous, binary, categorical, etc. However, this kind of imputation tends to have a limitation of being more computationally intensive. It supports many imputation models, such as forms of predictive mean matching (pmm) for continuous data, logistic regression for binary data, and polytomous regression for unordered categorical data ([Van Buuren & Groothuis-Oudshoorn, 2011](#)). In this project, we used pmm as we simulated continuous data to mimic metabolomics data.

## **AMELIAII**

AmeliaII uses a joint modelling approach based on a multivariate normal distribution, which utilizes means and covariances to summarize data. First, it takes  $m$  bootstrap samples and applies the expectation-maximisation bootstrapping (EMB) algorithm to each sample. Then it applies the regression of these samples to obtain the estimates. Although data used in AMELIAII needs to be normally distributed or transformed to approximate normality, AMELIAII performs computations faster compared to MICE due to the joint modelling approach ([Honaker et al., 2011](#)).

## **Impute.LCMD (QRILC)**

The 'impute.LCMD' ([Package & Lazar, 2015](#)) method performs data imputation using various methods, including the left-censored missing data method that can handle the MNAR assumption in omics data. The function `impute.QRILC` in the package `impute.LCMD` implements the left-censored method ([Wei, Wang, Su, Jia, Chen, & Chen,](#)

2018) that accounts for truncation at the minimum value. This function imputes left-censored missing data with arbitrary randomizations from a condensed distribution using quantile regression estimated parameters. A scalar is employed to adjust the standard deviation when the entire data distribution is non-Gaussian. We have used the standard value `tune.sigma = 1`, which is equivalent if the total data distribution is Gaussian.

### **impute.knn**

The 'impute.knn' implements a function to impute missing data, using nearest neighbour averaging (Hastie, Tibshirani, & Narasimhan, 2018). When missing values are present, the  $k$ -nearest observations are used to impute them using a weighted mean of the neighbouring values. This method computes a distance measure between each pair of observed values. We used the default setting of  $k = 15$  in this function.

### **missForest**

We used the R package missForest (Package & Lazar, 2015) to implement imputation by RF. This package has several advantages, including its ability to improve the performance with higher correlation, even in moderate to a high percentage of missing data, and it also can tackle MNAR data (Package & Lazar, 2015). While most prevalent imputation methods depend on tuning parameters or specification of a parametric model, this package implements cross-validation to obtain a suitable  $k$ , the number of clusters the algorithm is trying to identify or learn from the data.

#### **6.2.4. Association analysis within the MP-GWAS framework**

The association analysis was done within MP-GWAS framework using the reverse regression approach to test whether the imputed datasets work well to detect association signals compared to the full data. It was done after the imputation step where the phenotypes were regressed against genotype using the following equation:

$$SNP_i = \alpha_i + \beta_{i1}X_1 + \beta_{i2}X_2 + \dots + \beta_{in}X_n + \epsilon_i$$

$$i = 1, 2, 3; n = 3, 30, 120$$

where n is the number of phenotypes and i is the genotype that consist of randomly selected common variants ( $MAF_{n=5,000} = 0.157$  ,  $MAF_{n=50,000} = 0.169$ ,  $MAF_{n=500,000} = 0.34$  ), low-frequency variants ( $MAF_{n=5,000} = 0.022$  ,  $MAF_{n=50,000} = 0.021$ ,  $MAF_{n=500,000} = 0.022$  ), and rare variants ( $MAF_{n=5,000} = 0.002$  ,  $MAF_{n=50,000} = 0.002$ ,  $MAF_{n=500,000} = 0.001$  ).

The Root Mean Square Error (RMSE) was used to measure the least biased estimates of the imputation methods based on the deviation of the predicted values to the true values. It is usually calculated from the square root of the squared error mean. In our study, we applied the RMSE to calculate estimated regression coefficients ( $\beta$ s) from imputation vs those from the full data analysis:

$$RMSE = \sqrt{\sum \frac{(\beta_{imp[1..N]} - \beta_{full[1..N]})^2}{N}}$$

Where  $\beta_{imp}$  is the regression coefficient from the model using the imputed dataset,  $\beta_{full}$  is the regression coefficient from the model of the full dataset, and N denotes the number of phenotypes used in the analysis.

### 6.3.Results

This study evaluated the effect of imputation methods of SI/MI with MICE and AMELIA II, QRILC, k-NN, and RF in the MP-GWAS framework under different scenarios. I used genetic data simulation that represents highly ( $r=0.64$ ) and moderately correlated ( $r=0.33$ ) phenotypes (three/30/120) in three different sizes of datasets with 5,000/50,000/500,000 individuals. I generated the phenotypes with different percentages of missing data (1/5/20/50%) under three missing data mechanisms: MCAR, MAR, and MNAR. I simulated the phenotypes to have significant associations ( $P<5\times 10^{-8}$ ) with random selections of common, low-frequency, and rare variants.

The accuracy of the seven imputation methods in 5,000 individuals is shown in [Figure 6.2](#) under the MAR scenario with different MAF, correlations between the phenotypes and proportions of missing data.

Within all scenarios, I observed that the CC analysis performance was the worst, especially when the proportion of missing data was increased. Moreover, when the number of phenotypes was increased, all the imputation performances were not influenced by that, except CC analysis. However, all of the methods' performances worsened with an increasing proportion of missingness. As the phenotypes are correlated, the proportion of missing data may have less effect with the smaller MAF when performing the regression. I observed similar patterns under MCAR and MNAR assumptions with the same scenarios (see [Supplementary Table 2](#) and [Supplementary Table 3](#)). The performance of the seven imputation methods in 50,000 individuals under MAR, MCAR, and MNAR assumptions with the same scenarios as in 5,000 individuals also showed similar patterns (see [Supplementary Table 2](#) and [Supplementary Table 3](#)).

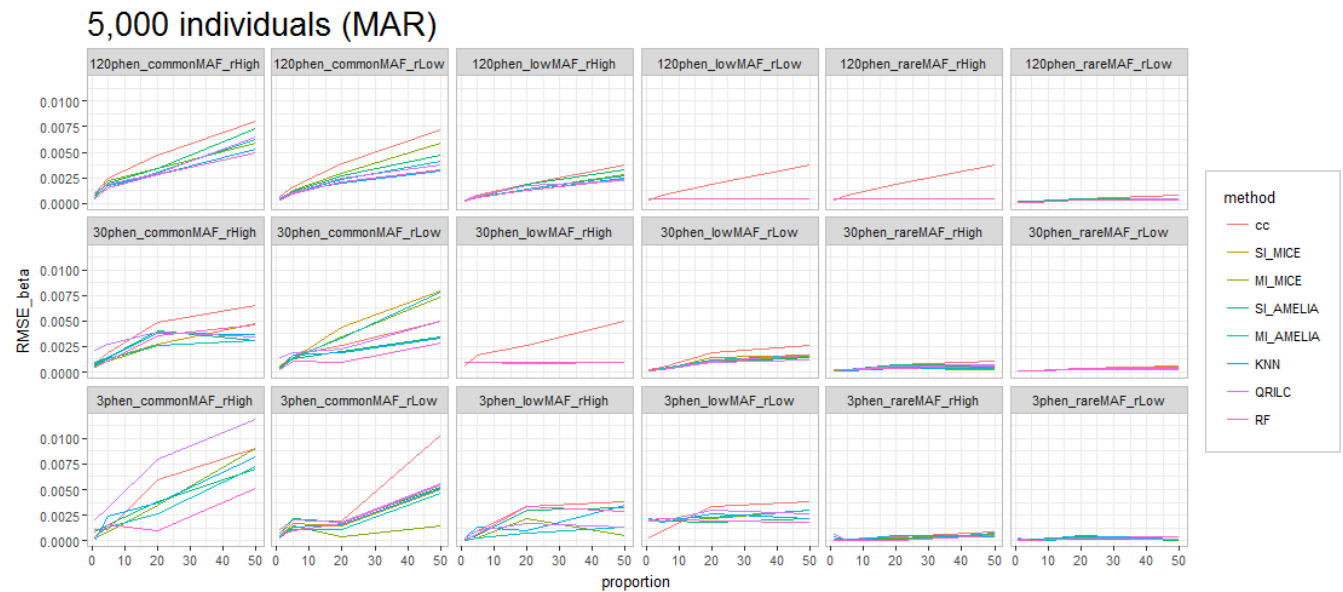


Figure 6.2 The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis in MP-GWAS analysis.

The top panel illustrates RMSEs of 120 phenotypes, the middle panel for 30 and the lower panel for 3 phenotypes. The two pairs of figures are for high and low phenotype correlations, with the first two on the left for common variants, next two for low-frequency variants and the last two for rare variants.

In 500,000 individuals, RF and MI-MICE could not be implemented due to their estimated very long computational times when analysing 3, 30, and 120 phenotypes. After three days, it always stopped due to run out of memory usage. I tested the ability to run RF and MI-MICE in 100,000/250,000/500,000 individuals. Both RF and MI-MICE worked well in 100,000 individuals and stopped working in 250,000 individuals due to insufficient computational power as larger memory is needed. In this scenario, I used the Imperial College HPC Cluster with compute nodes assembled with Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz.

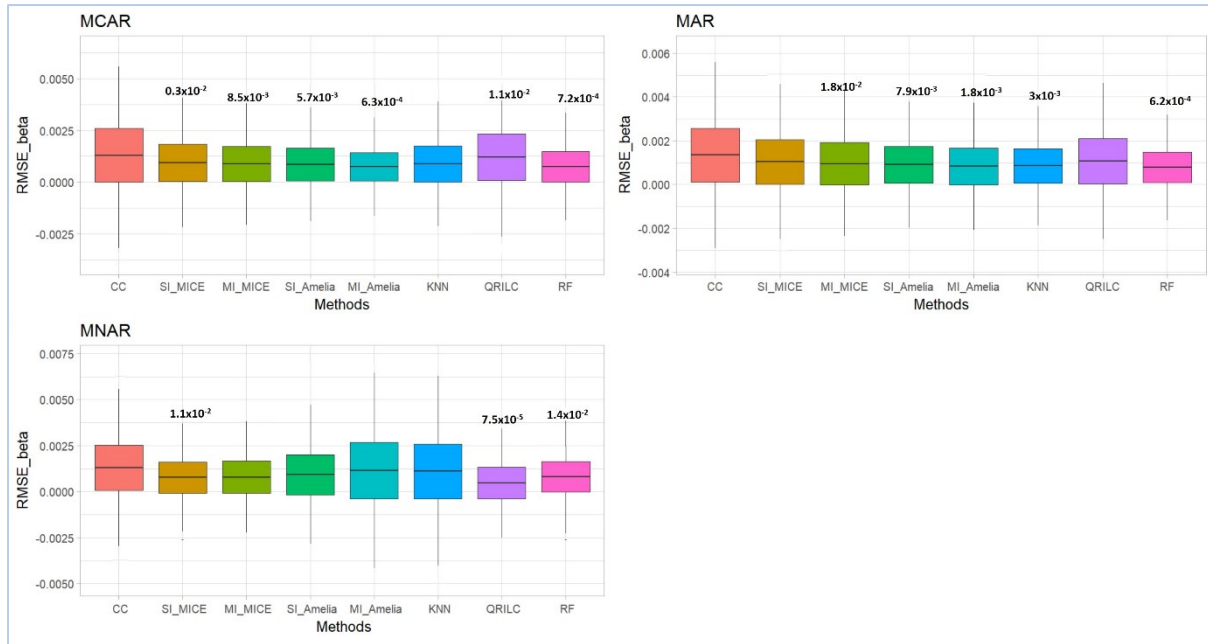


Figure 6.3 The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis in MP-GWAS analysis in 5,000 and 50,000 combined.

The smallest deviations closest to zeros from the reference CC were observed in RF and MI-EMB, i.e., 0.0014 and 0.0015, and their largest RMSE values were under MCAR ( $P_{RF}=7.18 \times 10^{-4}$ ,  $P_{MI-EMB}=6.27 \times 10^{-4}$ ), showing that the predicted regression coefficient ( $\beta_{imp}$ ) is almost similar with the true values of  $\beta_{full}$ . Under the MAR assumption, our simulation studies have revealed that RF remains the best performing imputation method ( $P=6.22 \times 10^{-4}$ ). However, under MNAR, RF was outperformed by QRILC ( $P=7.45 \times 10^{-5}$ ). The performance of RF and SI-MICE remained accurate, showed by relatively small *P-values* even under the scenario of MNAR, although both assume at least MAR. However, MI-AMELIAII drastically outperformed SI-MICE/MI-MICE regarding computational time ( $t_{AMELIAII} \sim 30$  minutes,  $t_{MICE} \sim 72$  hours for 120 phenotypes and 50,000 individuals).

A summary of all the imputation methods used in this project is provided in [Table 6.1](#).

Table 6.1 A comparison of the seven imputation approaches applied to the simulated datasets in the present study.

| Methods                                      |                 | Advantages                                                                                                                                                                                                                                                                                          | Drawbacks                                                                                                                                                                                                                           | Mechanics of Missingness |
|----------------------------------------------|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| <b>Single imputation (SI)</b>                |                 | -computationally fast                                                                                                                                                                                                                                                                               | <ul style="list-style-type: none"> <li>– It does not take into account the uncertainty of the estimates</li> <li>– Underestimates the correlation</li> </ul>                                                                        | MCAR, MAR                |
| <b>Multiple Imputation (MI)</b>              |                 | + Considers the uncertainty of the estimates                                                                                                                                                                                                                                                        | – analysis software needs to accommodate multiple input files and analyses                                                                                                                                                          | MCAR, MAR                |
|                                              | <b>MICE</b>     | + Handles a mixture of phenotype distributions                                                                                                                                                                                                                                                      | – Slow and computationally intensive                                                                                                                                                                                                |                          |
|                                              | <b>AMELIAII</b> | + Computationally fast                                                                                                                                                                                                                                                                              | – Assumes the data to be normally distributed                                                                                                                                                                                       |                          |
| <b>K-Nearest Neighbour (kNN)</b>             |                 | + Computationally fast                                                                                                                                                                                                                                                                              | – It needs to be tuned                                                                                                                                                                                                              | MCAR, MAR, MNAR          |
| <b>Left-censored imputation method QRILC</b> |                 | + Computationally fast                                                                                                                                                                                                                                                                              | – It needs to be tuned                                                                                                                                                                                                              | MNAR                     |
| <b>Random Forests (RF)</b>                   |                 | <ul style="list-style-type: none"> <li>+ Handles a mixture of phenotype distributions and non-linearity</li> <li>+ Robust to outliers</li> <li>+ Robust to correlated and large feature space (<math>n &lt; p</math>)</li> <li>+ It can be run on multiple cores</li> <li>+ Easy to tune</li> </ul> | <ul style="list-style-type: none"> <li>– Fast but computational time dependent on the size of the data</li> <li>– Bootstrapping used to grow the trees worsens the prediction when only a handful of features are useful</li> </ul> | MCAR, MAR, [MNAR]        |

## 6.4. Discussion and conclusions

Throughout this study, in terms of imputation precision, sample distribution, and statistical analysis, I systematically tested a total of seven imputation methods on various forms of missing values. The findings showed that the RF imputation for MCAR/MAR gives the least biased parameter estimates. I observed that QRILC was the least biased imputation method than others for MNAR assumption; however, QRILC did the worse under the missing mechanisms of MCAR/MNAR. Therefore, for the MNAR assumption, I only suggest QRILC when certain that the data only contain MNAR before the imputation.

*Table 6.2 The performance comparison between missing data imputation methods which indicates that RF is the best imputation in term its performance, the ability to handle mix type of data, and the easy setting that require no parameter tuning*

| Methods                | Performance | Missing data mechanisms | Fast | Mix type of data | No parameter tuning |
|------------------------|-------------|-------------------------|------|------------------|---------------------|
| Complete Case Analysis |             | MCAR                    | ✗    |                  | ✗                   |
| AmeliaII*              |             | MNAR, MAR               | ✗    |                  | ✗                   |
| QRILC                  | ✗           | MNAR                    | ✗    |                  |                     |
| kNN                    |             | MCAR, MAR, MNAR         | ✗    |                  |                     |
| MICE*                  |             | MCAR, MAR               |      | ✗                | ✗                   |
| Random Forest          | ✗           | MCAR, MAR, MNAR         |      | ✗                | ✗                   |

\*SI and MI

I investigated the efficiency of the imputation approach for various scenarios in the MP-GWAS framework. I tailored parameter settings or data pre-processing to achieve maximum performance to prevent potential bias. For instance, kNN was typically applied to a matrix of  $p \times n$  gene expression (genes in rows and samples in columns) (J. S. Shah et al., 2017), while I found that kNN worked better on a matrix of  $n \times p$  metabolomics data (samples in rows and metabolites in columns). The underlying

distinction is that I attempted to find k nearest samples to reflect the missing ones rather than discovering k nearest variables, which are genes in their original case (Wei, Wang, Su, Jia, Chen, & Chen, 2018). This is rational since data on gene expression typically includes a huge number of genes (more than 10,000) where co-expression happens frequently; therefore, nearest genes are usually good for missing genes (Wei, Wang, Su, Jia, Chen, & Chen, 2018).

As suggested by the original paper (Van Buuren *et al.*, 2006), the iteration used in MI was  $m=5$ , which does not create a computational burden while also keeping the imputation perform efficiently without sacrificing the accuracy.

Log-transformation was previously considered useful to improve imputation precision and guarantee a positive value in the original scale for QRILC imputation (Wei, Wang, Su, Jia, Chen, & Chen, 2018). However, another study argued that pre-transformation before the imputation would distort the distribution and the correlation of the data (Hong & Lynn, 2020).

I compared the difference in RMSE of  $\beta$ s from the imputation models and CC analysis with those from the full model using linear regression. It shows that under MCAR, RF and MI using AMELIAII were the best performing methods indicated by their smallest P-values, whereas under MAR, RF was the best one. QRILC was the best method under MNAR, as it was designed for MNAR only, therefore under MCAR and MAR, QRILC is considerably worse than the other imputation methods.

The imputation process will not always make a major difference in the predictive ability for datasets with fairly tiny percentages (< 5%) of missing values. Yet, as the proportion of missingness increases, however, the predictive ability of all six strategies is substantially different. In all cases, there was not a clear approach that worked the

best. It is recommended that several different methods of imputation should be attempted, and the most precise method chosen. In certain cases, this might not be possible. Therefore, the proposed best approach to use is under different scenarios is RF, especially if we do not know the type of missing data, as is usually the case.

With the increasing number of missing values, especially those in the simulated study with 50% missing, imputation estimates would decrease significantly.

Therefore, I suggested a systematic approach for dealing with missing values in the MP-GWAS setting, consider eliminating missing variables that have significant proportions of missing values beforehand.

For targeted and non-targeted metabolomics data, I propose (1) testing raw data for both targeted and profiling/non-targeted metabolomics data and, if appropriate, changing parameter settings or manually matching peaks to fill back those missing values correctly; (2) filtering missing variables, e.g., "modified 80% rule" to exclude certain inaccurate variables with a significant proportion of missing values; (3) for missing values that cannot be recovered, the potential explanations for generating missing values could be determined before the imputation and choose a suitable one for missing value imputation; (4) QRILC imputation is recommended for left-censored MNAR circumstances (e.g., in targeted metabolomics data); and RF imputation is recommended for MCAR/MAR situations (i.e., in profiling/non-targeted metabolomics data). Although missing metabolomic data tend to be MNAR because small values typically cannot be detected due to the device limitation, the results of RF, a method not tailored for this scenario specifically, are in line with those from the study conducted by Kokla et al. ([Kokla, Virtanen, Kolehmainen, Paananen, & Hanhineva, 2019](#)). They also found that RF performs the best with the smallest errors using simulated missing

data with the mixing MAR-MNAR and MAR-MCAR in metabolomic data, although they only used small sample datasets with 200 variables (Kokla et al., 2019). I did not explore the mixing of different scenarios but showed that even under MNAR, the RF performed reasonably well and even better under MAR and MCAR. In real life, it is more likely that the mechanisms of missing data are a mixture of MAR, MCAR, and MNAR, suggesting that RF would work extremely well in such a scenario.

Both SI and MI assume the data are at least missing at random (MAR). Based on this assumption, the estimates ( $\beta$ s and SEs) are unbiased and without performance loss. However, applying both methods to unobserved data that only depend on the missing data itself (MNAR) requires additional testing to observe whether SI and MI still work well under MNAR assumption.

A study mentioned that MI works well for up to 30–40 variables; however, these methods are especially poorly suited to data sets with many more variables (Yuan, 2010). However, after testing the MI method with more than 40 variables to confirm this assumption, we found that in 120 variables, MI still works well.

In terms of the missing data mixture, MI also works well for this condition; however, MI relies on distributional assumptions and cannot accommodate nonlinear relations and interactions (Burgette & Reiter, 2010). In contrast, RF was developed based on regression trees and classifications, which do not rely on distributional assumptions. Although when regression trees are used, this can produce an ‘overfitting’ that tends to develop a complex model with low predictive power, RF was developed to avoid this condition by using bootstrap aggregation of multiple regression trees by building different models (T. H. Lee, Ullah, & Wang, 2020).

Nearly half of the molecular values do not obey the normal distribution in most metabolomics datasets because their values are much left-skewed. The molecular variable normally log-transforms in the experimental setting to prevent the imputation of very small values since their values are close to zero. Before imputing, this transformation follows the normality assumptions; nevertheless, it not only altered the distribution of molecular values but also disrupted the relationships with the rest of the data between those variables, and thus this method will lead to imputing outliers or/and generating more bias rather than imputing the skewed variable.

When complex interactions and nonlinearity are present, RF was claimed in its original study as an imputation algorithm developed for mixed continuous and/or categorical data without the need to specify the variables' distributions ([Stekhoven & Bühlmann, 2012b](#)). However, a study discovered that RF, particularly for highly skewed variables in nonlinear models, may generate highly biased regression coefficient estimates and lower-biased confidence interval coverages. They conclude that RF-based imputation, specifically for highly skewed data and/or outcome-dependent MAR, should not be indiscriminately recommended for missing data imputation as a universal solution. Proper analysis requires careful examination of the mechanism of missing data and the relationships between data variables as a pre-imputation step ([Hong & Lynn, 2020](#)). However, this conclusion is also valid for other methods of imputation, such as SI and MI, since these methods assume that the data should be normally distributed. Shah et al. ([A. D. Shah, Bartlett, Carpenter, Nicholas, & Hemingway, 2014](#)). compared the performance of MI with RF and found that RF also has less biased parameter estimates on simulated data with interactions between variables ([A. D. Shah et al., 2014](#)).

The missing mechanism and proportion of missingness affect the performance and suitability of imputation methods within the MP-GWAS framework. In conclusion, the RF-based imputation method performs best in most of the tested scenarios. It works in all three different missing mechanisms and different proportions of missingness. Besides, it can handle mixed types of data, unlike QRILC. Therefore, I recommend using RF-based imputation for imputing missing metabolomics data and other missing data in general as it has been already applied in many studies since, typically, the origin of missingness is not known in advance. When the missingness is known in advance as MNAR, left-censored imputation, such as QRILC, should be chosen. In addition, my approach to evaluating the performance of imputation methods on metabolomics datasets is also applicable to other high-dimensional data that contain missing values.

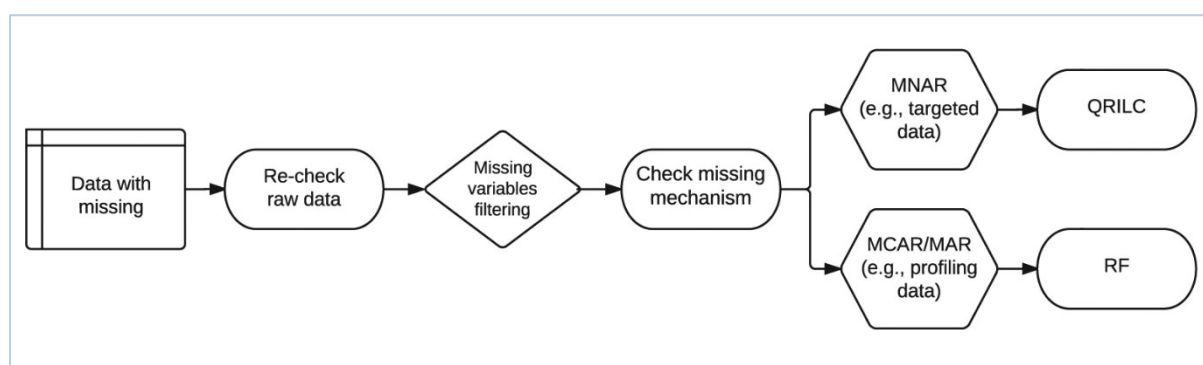


*Figure 6.4 The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis in MP-GWAS analysis in 5,000 and 50,000 combined.*

## 7. Software development for phenotype imputation in the MP-GWAS framework

*While chapter 6 focussed on comparing several advanced imputation approaches to search for the best imputation method in the MP-GWAS setting, this chapter focusses on describing the implementation of the best imputation method I observed, i.e., RF, into a fast, publicly available software. While the RF software available up to date is only available for imputation of small datasets, I implemented an RF imputation software that can impute high-dimensional data, such as UK Biobank. Finally, I compared the performance of my newly developed software imputeSCOPA with the available RF software in terms of the ability to handle high-dimensional data and computational time.*

### 7.1. Introduction

Based on the simulated study in Chapter 6, I observed that RF has the most advantages compared to the other methods in terms of its performance in the imputation that provided regression coefficients closest to those from the true values in the MP-GWAS framework, especially under the missing mechanisms of MAR and MCAR. Moreover, RF can also handle mixed types of data. However, RF, as implemented in the current software, has slow computational time as its drawback. It was the second slowest method after MI using MICE when compared to the other methods.

In addition, the RF implementation should be able to perform imputation both on continuous and binary phenotypes. RF is widely used in classification applications, such as gene expression analysis, protein-protein interactions, biological sequences identification, genome-wide association studies, credit scoring, or image processing (Bosch, Zisserman, & Muñoz, 2007; Kruppa et al., 2014; Qi, 2012). RF works well with a mixture of numerical and categorical features, similarly to MI.

Therefore, I proposed a fast RF implementation in the MP-GWAS framework, enabling the imputation of large-scale high-dimensional phenotypic data. It extends the faster available RF software to perform imputation instead of just performing classification, regression, and survival analysis.

I developed new software, imputeSCOPA, based on the modification of C++ application of RF software, Ranger, to do imputation using the RF approach as the original Ranger software implemented the imputation. My main goal was to provide a cross-platform, modular framework for imputing high dimensional data with RF faster than the original Ranger implementation in C++ and other available RF imputation. This chapter describes the new implementation of RF software in C++ and validates the imputation within MP-GWAS and compares runtime and memory usage with existing software.

I tested the performance of imputeSCOPA with simulated data, varying the number of individuals and the proportion of missing data.

## **7.2.Methods**

### **7.2.1. Implementation of the random forest imputation in C++**

The core of imputeSCOPA was implemented in C++, and we used standard libraries only. Therefore, we do not need to install R software to integrate with it as it is a standalone C++ software with easy installation. We implemented the imputation part using Armadillo, a high-quality linear algebra library (matrix maths) for the C++ language, aiming towards a good balance between speed and accuracy.

A standalone C++ version can easily be installed on any up-to-date platform. In the implementation, we made extensive use of C++11 features. With this, parallel processing is available on all platforms with the thread library, and the random library is used for random number generation.

Data can be a mix of continuous and categorical data, such as in omics data. ImputeSCOPA was developed to analyse this omics data by applying memory profiling with a simple data structure instead of copying the original data. We optimized imputeSCOPA for high-dimensional data by identifying bottlenecks and optimized the relevant algorithms for different input data types with a high number of phenotypes. The fittest forest was used to predict the missing data by regressing each variable against all variables. The fittest forest is chosen by grouping variables by calculating each variable until convergence to avoid time-consuming.

Also, node splitting was used for the values of all mtry parameters (candidate variables, which indicates the number of input variables randomly chosen at each node). A major improvement was implemented as the same as in Ranger to keep the fast performance compared to the other machine learning approaches by using the algorithm for sampling without replacement ([Knuth, 1985](#)).

I modified the Ranger software to enable certain methods to override, extended the tree regression and data classes to re-implement. The actual implementation is very

similar to the original Ranger; however, imputeSCOPA allows for moving the repetitive iteration upfront to make it faster. For example, I only sort each column when it changes, rather than sorting all the data every time (on every imputation).

I also made a small change in the criteria of Mean Percentage Error (MPE) of Ranger. This MPE change significantly reduced the imputation time by 5 hours for running high dimensional data, such as 500,000 individuals with 120 variables. As the prediction error drop for the last two iterations was very small ( $< 1e-5$ ), I made the criteria become  $MPE - \text{last MPE} < -1e-5$ .

I modified the R version of missForest to represent the same algorithm of imputeSCOPA in the R version to compare the performance between the stand-alone C++ version and the R version. I modified the package to receive input with a specific seed, as imputeSCOPA uses a random number generator to create a seed. Thus, I needed to modify the package to accept a specific seed instead of random; thus, it can be compared with the imputeSCOPA version in C++ using the same seed as input. With this modification, I could compare and conclude that both the C++ and R version have the same results, yet faster computational time in C++.

Therefore, I concluded that the actual imputeSCOPA that needs to be implemented is the C++ version.

### **7.2.1. Data input and output**

Since imputeSCOPA is written in C++, it has a command-line user interface. A single run of imputeSCOPA consists of just one step and the required input files, commands, and the resulting output files as described in [Figure 7.1](#).

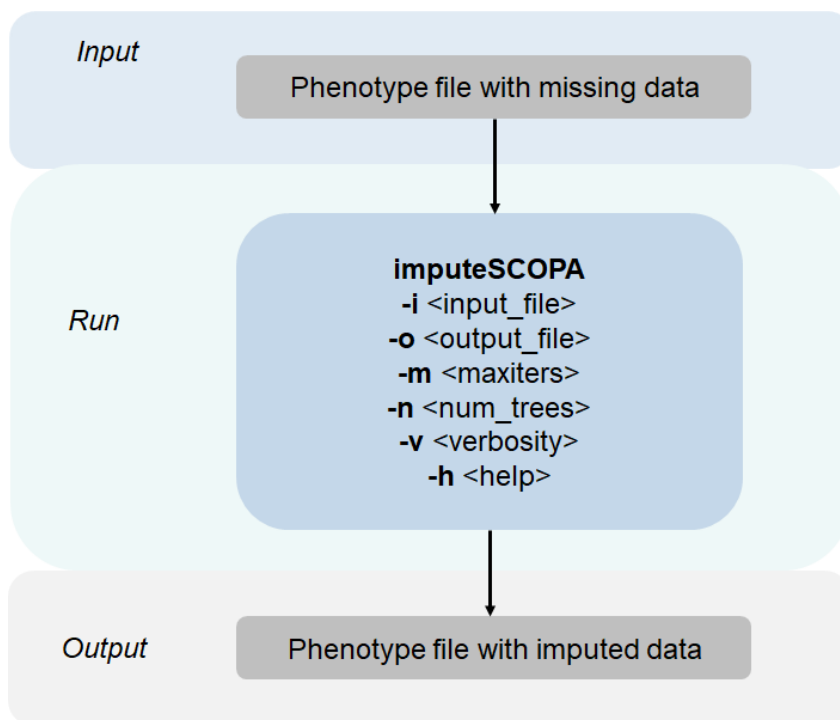


Figure 7.1 Workflow of an *imputeSCOPA*, including input file, commands, and output file.

```

~/imputeSCOPA/build$ ./imputeSCOPA
Missing arguments
./imputeSCOPA: Missing value imputation by chained tree ensembles
Usage: ./imputeSCOPA -i <input_file> [-o <output_file>] [-m <maxiters>] [-n <num_trees>] [-v <verbosity>] [-h]
-i input_file: File with missing data, required
-o output_file: File for the output data, default: out.txt
-I Don't print ID column (for comparing the imputeSCOPA.R output)
-m maxiters: Maximum number of iterations, default: 5
-n num_trees: Number of trees, default: 10
-s separator: separator character, for tab don't pass this argument
-S seed: seed for RNG (default is time)
-v verbosity: verbosity level, default 0
-h: show this usage
~/imputeSCOPA/build$ ./imputeSCOPA -i ../data/input_with_missing_data.txt -S -v 1 -m 10 -o ../data/output.txt
~/imputeSCOPA/build$
  
```

Figure 7.2 *imputeSCOPA*, command run on terminal

ImputeSCOPA requires a phenotype file for a successful run that consists of a header in the first row and the variable values for the rest. However, we intend to extend it to accept the .sample file in the SNPTEST format for the next implementation.

The formatted output data is the same as the input data, except the missing values are filled with the imputed values. The first column of the data is the ID, followed by the variables; the first row is the columns' header (see [Figure 7.3](#) and [Figure 7.4](#)).

```
~/imputeSCOPA/data$ head input_with_missing_data.txt
```

| ID | y1               | y2               | y3                                 |
|----|------------------|------------------|------------------------------------|
| 1  | 4.54077811533716 |                  | 4.18194104089821 3.5271406149932   |
| 2  | NA               | 4.79643210750428 | 3.2488507051803                    |
| 3  | 3.93927302821211 | NA               | 3.77929609261064                   |
| 4  | NA               | 4.52854012608869 | 4.75878926014701                   |
| 5  | NA               | 4.31453859452195 | 3.12988788443059                   |
| 6  | 2.86036279379074 |                  | 2.09748861478793 0.623633834424999 |
| 7  | 4.26202116052791 |                  | 4.87736846006347 3.54644055978994  |
| 8  | 2.84040060879582 | NA               | 4.83836404192318                   |
| 9  | 3.20871164763785 |                  | 2.41735993332475 NA                |

Figure 7.3 imputeSCOPA input file containing missing data before imputation.

```
~/imputeSCOPA/data$ head output.txt
```

| ID | y1               | y2 | y3                                 |
|----|------------------|----|------------------------------------|
| 1  | 4.54077811533716 |    | 4.18194104089821 3.5271406149932   |
| 2  | 5.31613871950922 |    | 4.79643210750428 3.2488507051803   |
| 3  | 3.93927302821211 |    | 4.02932723618174 3.77929609261064  |
| 4  | 3.83974523099613 |    | 4.52854012608869 4.75878926014701  |
| 5  | 4.10774787345464 |    | 4.31453859452195 3.12988788443059  |
| 6  | 2.86036279379074 |    | 2.09748861478793 0.623633834424999 |
| 7  | 4.26202116052791 |    | 4.87736846006347 3.54644055978994  |
| 8  | 2.84040060879582 |    | 3.20456631689846 4.83836404192318  |
| 9  | 3.20871164763785 |    | 2.41735993332475 2.74470317136585  |

Figure 7.4 imputeSCOPA output file after imputation.

## 7.3.Comparison with existing RF implementations

### 7.3.1. Running time and memory

To assess the performance of imputeSCOPA in a high-dimensional setting, I compared the runtime of imputeSCOPA with missForest using simulated data. First, the R packages missForest and the C++ version of the new implementation imputeSCOPA were run with small, simulated datasets with subsequent testing to high-dimensional data by varying the sample size. The number of phenotypes used in this simulation is 120 phenotypes. In each case, the other three parameters were kept fixed to 500 trees. All imputations were run, and their performance was compared

using the Imperial College HPC Cluster. Compute nodes were equipped with Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz machine.

*Table 7.1 Computational time comparison between missForest and imputeSCOPA in a different number of individuals between 5,000 – 500,000 individuals.*

| Number of Individuals | missForest           | imputeSCOPA          |
|-----------------------|----------------------|----------------------|
| 50k                   | 42 hours, 12 minutes | 5 hours, 38 mintues  |
| 100k                  | 70 hours, 46 minutes | 8 hours, 20 minutes  |
| 250k                  | -                    | 14 hours, 22 minutes |
| 500k                  | -                    | 26 hours, 16 minutes |

For 5,000 individuals and 120 phenotypes, it takes about 2.25 minutes per imputation, and most datasets have converged around three iterations with resident memory around 4.2G. It took about 10.5 minutes to perform the imputation with imputeSCOPA in this dataset, which was more efficient than about 35 minutes using the R version of missRanger. This R version is even slower than the original ranger in C++, assuming that we did not make any change to speed up the imputation time. The 500,000 individuals with three variables file only took about 3+ minutes. In 500,000 individuals with 120 variables, it took 801 minutes with five iterations. In total, to complete the imputation in 500,000 individuals takes about 26 hours with 10 iterations ([Table 7.1](#)), while in R, it takes more than three days, and we were not able to finish the imputation using the 72-hour time limitation on our server.

## 7.4. Discussion and conclusions

The currently available RF implementations have weaknesses and advantages of their own.

The original RF implementation ([Breiman, 2001](#)) is in the programming language Fortran 77 and requires recompilation each time the data or altering parameters. Liaw and Wiener developed randomForest in R ([Liaw & Wiener, 2002](#)), which had is feature-rich and widely used. It is not, however, best suited for use with large high-dimensional data ([D. F. Schwarz, König, & Ziegler, 2010](#)). Commencing in 2006, Hothorn, Hornik, & Zeileis provides a recursive partitioning framework and features an RF implementation, but shares the shortcomings of randomForest ([Hothorn, Hornik, & Zeileis, 2006](#)). Ishwaran et al. ([Ishwaran, Kogalur, Blackstone, & Lauer, 2008](#)) implemented randomForestSRC as an R package, which provides regression, classification, and survival, but no imputation application is available ([Ishwaran et al., 2008](#)).

Random Jungle is an RF implementation designed to analyse high-dimensional data ([Kruppa et al., 2014; D. F. Schwarz et al., 2010](#)). Available as a library dependent C++ application alone, this package cannot be ported to R or any other statistical programming language and does not support imputation. The recent fastest implementation of RF to deal with high-dimensional data has been implemented in Ranger ([Wright & Ziegler, 2017](#)), available in a C++ application; however, this software does not support imputation.

There is only one RF application for high-dimensional data imputation, which was implemented in R package missForest ([Stekhoven & Bühlmann, 2012a](#)). In contrast,

other RF implementations are more focused on classification, regression, and survival analyses. However, our evaluations showed that this software struggles to handle high-dimensional data imputation as it is computationally inefficient to handle more than a hundred thousand individuals.

I developed the fastest C++ application of RF imputation named imputeSCOPA with its corresponding R code. I demonstrated the computational efficiency of imputeSCOPA compared with the available R package missForest (Stekhoven & Bühlmann, 2012a) using simulated data. This method's runtime was scaled for several parameters, including the number of trees and the mtry values.

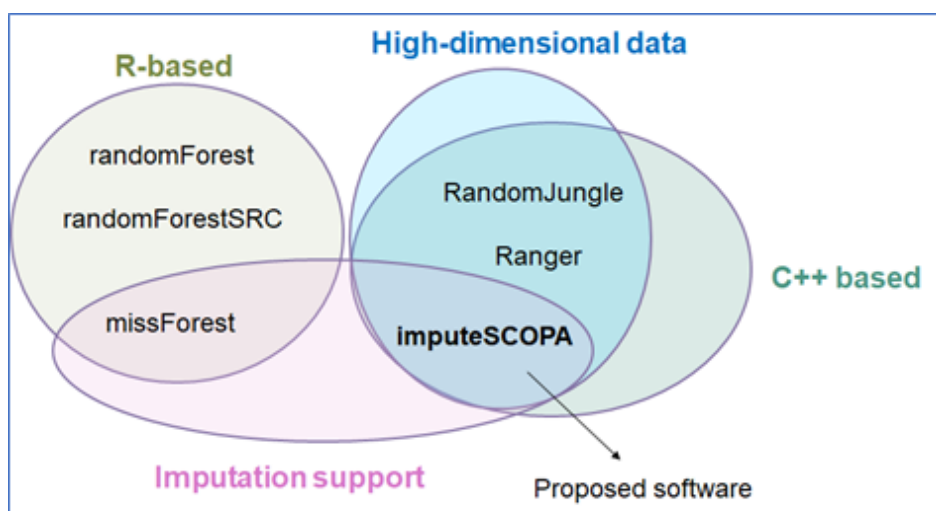


Figure 7.5 Classification of random forest (RF) software, based on their abilities to handle high-dimensional data, imputation support, and whether their codes were written in C++ or R.

The size of high-dimensional data may influence the choice of the number of trees used for the imputation. It should depend on memory usage and hardware specifications that might also affect the tuning parameters and the number of trees to be considered. While growing fewer trees worsen the imputation accuracy, using the same number of trees in any setting is always recommended. With the hardware specification on Imperial HPC, I determined the default of the number of trees to be

100 as it is widely used in standard RF software and still can be run in an efficient time for data with 500k individuals ([Kruppa et al., 2014](#)). With faster implementations available, I can now tackle missing data in high-dimensional omics data, especially within multi-phenotype approaches, as demonstrated by imputing different sizes of simulated datasets using my new software tool imputeSCOPA.

However, although I developed the least biased imputation method in the MP-GWAS framework, this method only focuses on handling bias due to missing data mechanisms. Whereas confounding, which usually affects the observational research and missing values in covariates, may need to be addressed.

Bias produces an invalid association because of missing mechanisms, whereas confounding defines a true association yet is potentially misleading. Methods for the propensity score, first proposed by Rosenbaum and Rubin ([Amos et al., 1986](#)), are increasingly being used to deal with confounding in analyses.

The missing graphs could help to clarify the missing structure. When no effect modification existed, CC analysis yielded valid causal effects even when data were MNAR ([Amos et al., 1986](#)). In some situations, CC analysis was also able to partially correct for unmeasured confounding ([Amos et al., 1986](#)).

In contrast, some questions arise in accordance with propensity scores in MI: Do we include the result in the imputation model? Should we use the methods of imputation introduced in regular software? How can we integrate the outcomes of each imputed dataset of the various propensity scores estimated?

A study revealed that in covariates of a propensity score model that could work optimally in all cases, there is no single approach to handle missing values ([J. Choi,](#)

[Dekkers, & le Cessie, 2019](#)). The best solution relies on the missing data structure and whether impact alteration or unmeasured confounding occurs ([J. Choi et al., 2019](#)).

In order to deal with missing values in covariates, CC analysis may also be a good option. While statistical efficiency is missed, estimated effects still have a causal interpretation if no unmeasured confounding occurs. In such cases, it would be the responsibility of the researcher to assess how generalizable these results are to the general population of interest. Generalizability cannot be taken for granted in the case of considerable heterogeneity in treatment impact.

Where unmeasured confounding is involved, all conventional missing data strategies suffer from precise estimates. Full case review may reduce the bias, however, by managing the unmeasured confounding by constraint. The bias may even be minimized using an indicator variable (even without several imputations) since it acts as a proxy for unmeasured confounding.

In observational comparative effectiveness analyses with a likelihood, score approaches, a new systematic review of how missing data is treated showed that only 86 (51%) of 167 studies conducted from 2010 to 2017 investigated missing data issues, and only 12 (7%) offered reasons for missing data ([J. Choi et al., 2019](#)).

Assumptions regarding the predicted association between the covariates that are ignored and observed are important to establish. This allows us to evaluate the expected incomplete data structure more cautiously and to handle missing values. We encourage researchers to use m-graphs to draw their hypothesis precisely between the covariates and their missing predictor. It is significant to draw predictions about the predicted association between the covariates missed and observed. It helps to prepare

for the anticipated structure of the missing data more carefully and to handle missing values. It is also advised that scientists use m-graphs to precisely map their conclusions between the covariates and the missing predictor for the pre-imputation step. In brief, in the context of propensity score analysis, it is recommended for researchers to deliberately choose missing data methods when taking into consideration the missing data pathways, heterogeneity of treatment effects, and possible unmeasured confounding, including initial exploration of results and the discovery of major covariates and potential confounding factors.

RF with genetic structure between populations and relatedness within populations can be confounded in GWAS approaches (Stephan et al. 2015; Zhao et al. 2012). When two populations differ phenotypically and genetically, studies may find associations representing this stratification rather than true genotype-phenotype associations. Response imbalances and explanatory variables (e.g., sample sizes or variances), increased LD values between study loci, and other considerations, such as repeated phenotypic measures for individuals or studies from different years or different age ranges, may also be confounded in studies. Before performing RF, confounding variables should be established, since there are different approaches to account for their impact when applying the RF algorithm.

In association studies, different methods have been proposed to address for correlations between individuals (e.g., population structure and relatedness) and minimize false-positive findings (Holliday, Wang, & Aitken, 2012; Y. Zhao et al., 2012) presented one method originally designed for RF analyses and expanded by Zhao et al. (Y. Zhao et al., 2012).

First of all, this two-step method includes regressing (one at a time) the phenotypes and genotypes at each locus against a population or family membership scale, including individuals' coordinates on PCA axes, employing linear or generalized linear models. The models' residuals, which reflect the "adjusted" phenotypes and genotypes, are then analysed by RF. As it relies on common regression models, this approach is simple to implement and can be expanded to account for other potential confounding variables, such as age and sex ([Y. Zhao et al., 2012](#)). Moreover, Zhao et al. have confirmed its efficacy in accounting for population stratification ([Y. Zhao et al., 2012](#))

However, correcting the demographic composition could also limit association statistical power and other genomic assessments ([Lotterhos & Whitlock, 2015](#)). That said, modern, more scalable approaches are being developed, including correction by best linear unbiased estimation or eigenvectors of covariance ([Azevedo et al., 2017](#)), and further research is worthwhile. Therefore, as an imputation approach that works for general conditions, RF is still recommended to handle missingness in high dimensional data.

## 8. Phenotype imputation in the MP-GWAS framework: real data application on blood metabolite levels

*I tested the imputeSCOPA software to impute the missing data on three projects with high-dimensional data having 1,010, 125, and 31 variables. The last-mentioned project was an MPA on blood metabolite levels, previously performed by our team using complete case (CC) data since imputeSCOPA had not been developed at that time, while the two other projects were already utilising data imputed with imputeSCOPA. Therefore, in this chapter, I present the last-mentioned project by comparing the MP-GWAS results between the data analysed as complete cases and after imputed by imputeSCOPA.*

### 8.1.Introduction

Complex diseases are associated with multiple metabolic processes characterised by changes in metabolite concentrations (Illig et al., 2010; Suhre & Gieger, 2012). Therefore, metabolite levels are usually used in clinical diagnosis since they represent the regulation of metabolic pathways for cellular processes (Zimmet et al., 2001). For example, a high level of FG is associated with diabetes (Inzucchi et al., 2013; Zimmet et al., 2001)

Many GWASs which focus on studying complex diseases have initially been performed using observed traits such as BMI or other anthropometric measurements (Emdin et al., 2017; Karaderi et al., 2015) as they are the easier risk factors to be measured. However, many researchers have recently shown interest in analysing genetic variations in metabolisms to intervene in dysregulations in metabolic disorders.

In this study, I would like to identify metabolic health biomarkers from the MPA of human blood metabolite levels representing biological pathways of metabolic health. We used the imputed phenotype data of NFBC1966 (N = 4,949 individuals) and NFBC1986 (N = 3,055) using imputeSCOPA and SI resulting from the use of MICE software (Van Buuren & Groothuis-Oudshoorn, 2011). I compared the MP-GWAS results from the two imputed data to observe whether using the advanced phenotype imputation method would increase the power of associations.

## **8.2.Methods**

### **8.2.1. Phenotype imputation**

We used all the variables in our datasets as predictors in the imputation processes even if not all of them will be used in the analysis; this is because imputeSCOPA based on the RF method in this analysis has better prediction more variables. However, the imputation was performed only on continuous phenotypes as metabolite levels are quantitative values.

Oral contraceptives and pregnancy status were predictors for female data only; thus, different-sex imputations were performed separately. Additional predictors such as PCs in NFBC1966 and MDS in NFBC1986 were also included to predict the missing data as they are widely used to avoid bias due to population structure in GWAS.

The lipoprotein concentrations were previously log-transformed before imputation to avoid imputation with low quality as they are small values close to zero. These concentrations were back-transformed for the analysis after the imputation.

### 8.2.2. Quality control (QC) and multi-phenotype analysis (MPA)

QC for the previous data was performed by our previous group members and included 31 selected metabolites. The selection of 31 phenotypes (see Table 8.1) was performed by restricting multi-collinearity (inspections on correlation matrix and variance inflation factor (VIF)) and by selecting representative metabolites from each molecular window. The correlation between any two metabolites was set to a maximum of |0.8|. This also provided reasonable VIF estimates (<10).

The exclusions were applied afterwards for female and male data separately as we need to exclude the pregnant status for female data only. People with T1D and T2D confirmed status were also excluded, and also those on lipid-lowering medication or diabetes medication and with serum glucose levels > 7mmol/L.

After the QC, the NFBC1966 data contained 4,955 individuals, and the NFBC1986 contained 2,687 individuals ready for further analysis.

Table 8.1 A summary of multicollinearity of the 31 metabolites used in the multi-phenotype analysis (MPA).

*The darker the colour in the correlation column, the higher positive or negative correlation between the metabolites. The darker the colour in VIF column, the higher the variance inflation index contributed by the corresponding variable.*

| Metabolite variable              | VIF   | Strong correlation within the group | ...with metabolite variable |
|----------------------------------|-------|-------------------------------------|-----------------------------|
| <b>Lipoproteins (N = 12)</b>     |       |                                     |                             |
| Extremely large VLDL conc.       | 3.399 | 0.737                               | Very large VLDL conc.       |
| Very large VLDL conc.            | 3.610 | 0.785                               | Large VLDL conc.            |
| Large VLDL conc.                 | 4.016 | 0.785                               | Very large VLDL conc.       |
| Very small VLDL conc.            | 7.628 | 0.793                               | Medium LDL conc.            |
| Medium LDL conc.                 | 9.914 | 0.793                               | Very small VLDL conc.       |
| Very large HDL conc.             | 5.216 | 0.614                               | Large HDL conc.             |
| Large HDL conc.                  | 6.938 | 0.614                               | Very large HDL conc.        |
| Small HDL conc.                  | 5.216 | 0.466                               | 18:2, linoleic Acid         |
| Triglycerides in LDL             | 6.363 | 0.693                               | Triglycerides in HDL        |
| Triglycerides in HDL             | 5.872 | 0.693                               | Triglycerides in LDL        |
| Mean diameter for VLDL particles | 6.152 | 0.692                               | Large VLDL conc.            |

|                                                  |       |        |                                  |
|--------------------------------------------------|-------|--------|----------------------------------|
| Mean diameter for LDL particles                  | 4.237 | -0.594 | Medium LDL conc.                 |
| <b>Fatty Acids (N = 7)</b>                       |       |        |                                  |
| Estimated degree of unsaturation                 | 7.873 | 0.716  | Omega-3 FAs to total FAs         |
| 22:6, docosahexaenoic acid                       | 8.748 | 0.627  | Omega-3 FAs to total FAs         |
| 18:2, linoleic acid                              | 8.185 | 0.755  | Medium LDL conc.                 |
| Omega-3 FAs to total FAs                         | 6.451 | 0.716  | Estimated degree of unsaturation |
| Omega-6 FAs to total FAs                         | 9.939 | -0.714 | Monounsaturated FAs to total FAs |
| Monounsaturated FAs to total FAs                 | 9.495 | -0.714 | Omega-6 to total FAs             |
| Saturated FAs to total FAs                       | 5.135 | -0.447 | Estimated degree of unsaturation |
| <b>Low-molecular-weight metabolites (N = 12)</b> |       |        |                                  |
| Glucose                                          | 2.034 | 0.483  | Glutamine                        |
| Lactate                                          | 1.519 | 0.432  | Alanine                          |
| Pyruvate                                         | 1.851 | 0.588  | Alanine                          |
| Citrate                                          | 1.285 | 0.190  | Glycerol                         |
| Glycerol                                         | 1.810 | 0.408  | Small HDL conc.                  |
| Alanine                                          | 2.182 | 0.588  | Pyruvate                         |
| Glutamine                                        | 1.944 | 0.483  | Glycine                          |
| Glycine                                          | 1.335 | 0.399  | Glutamine                        |
| Histidine                                        | 1.696 | 0.492  | Tyrosine                         |
| Isoleucine                                       | 4.196 | 0.704  | Valine                           |
| Valine                                           | 2.707 | 0.704  | Isoleucine                       |
| Tyrosine                                         | 1.721 | 0.501  | Valine                           |

I applied exclusions on genotype data by removing SNPs that deviate from the binomial distributions with a p-value of  $HWE < 0.00001$  and with inadequate imputation quality and having an info score below 0.4.

I run only the full model analysis to the selected NFBC1966 phenotypes on SCOPA. The Bonferroni correction of  $0.05/10^6$  was used to correct the multiple testing on genome-wide of one million SNPs; thus genome-wide threshold becomes  $5 \times 10^{-8}$ , whereas it is  $5 \times 10^{-6}$  for a suggestive meaningful association threshold. The suggestive significant threshold has been proposed to be a suggestive association that is expected to result in 1 false-positive association per GWAS. In follow-up research, the

suggestive threshold is recommended to classify variants for consideration, whereas the variants reaching the significant threshold will be considered regions with higher association probability (Vergara et al., 2018).

### 8.3.Results

I ran two meta-analysis MP-GWAS of NFBC1966 and NFBC1986 with complete cases (CC) and data imputed by imputeSCOPA. I identified 12 signals that have been previously established for metabolic traits (Figure 8.1), and with imputeSCOPA imputed data, I identified 16 signals, including two novel signals (Figure 8.2) covering 1,274 SNPs reaching the genome-wide threshold of  $5 \times 10^{-8}$ . The top SNPs per chromosome, reaching genome-wide significance for each imputed dataset are detailed in Table 8.2.

I identified the established signals in the *FADS1-2-3* region, *GLS2*, and *LIPC*. We also detected apolipoprotein E (APOE) loci that previously have been known to affect the regulation of total and LDL cholesterol levels (Belloy, Napolioni, & Greicius, 2019). The most significant variant reaching genome-wide significance of  $6.92 \times 10^{-118}$  (using the complete case data) and  $3.81 \times 10^{-127}$  (using imputeSCOPA imputed data) for the FA metabolite window is the SNP rs174564 within *FADS1*. This is a well-established locus for lipids (G. L. Russo, 2009; Wolters et al., 2017).

Furthermore, I detected two additional variants attaining the genome-wide significant threshold in the imputed data using imputeSCOPA (Figure 8.2): rs181847072 within *FCGR3B* gene and rs61803025 within the *ADAMTS3* gene. The variant rs61803025 on chromosome 4 has MAF in NFBC1966 of 1.01% and 1.30% in NFBC1986.

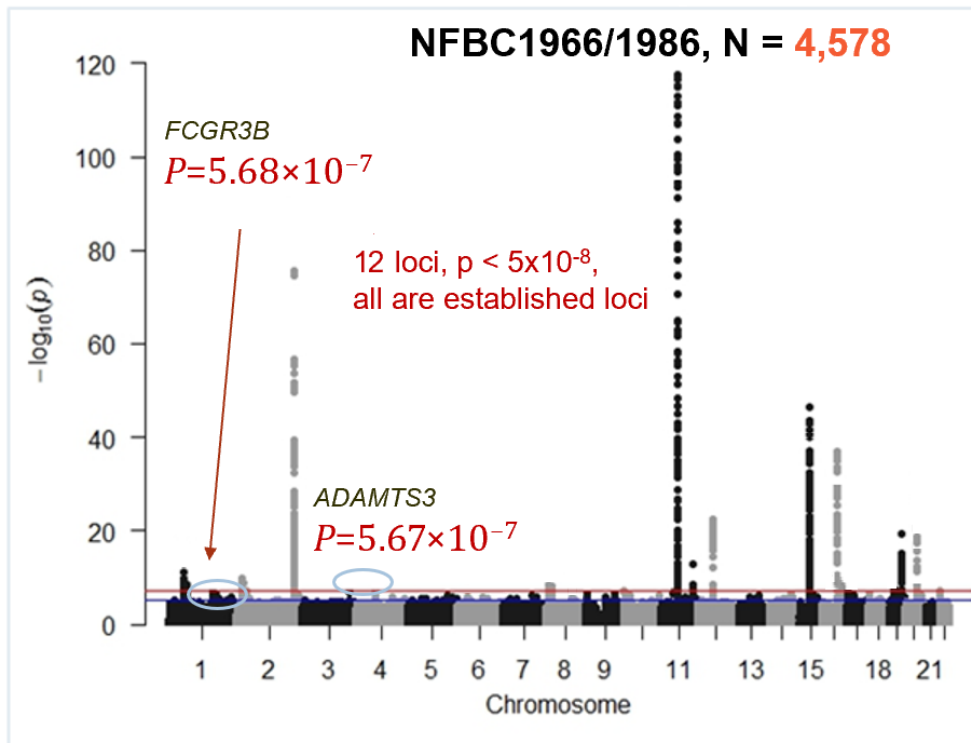


Figure 8.1 – Manhattan plot of MP-GWAS results using complete case data. The X-axis indicates the genomic position with the association P-value for each tested SNP on the Y-axis. The genome-wide significance threshold of  $5 \times 10^{-8}$  and the suggested significance threshold of  $5 \times 10^{-6}$  are represented with the red line and the blue line.

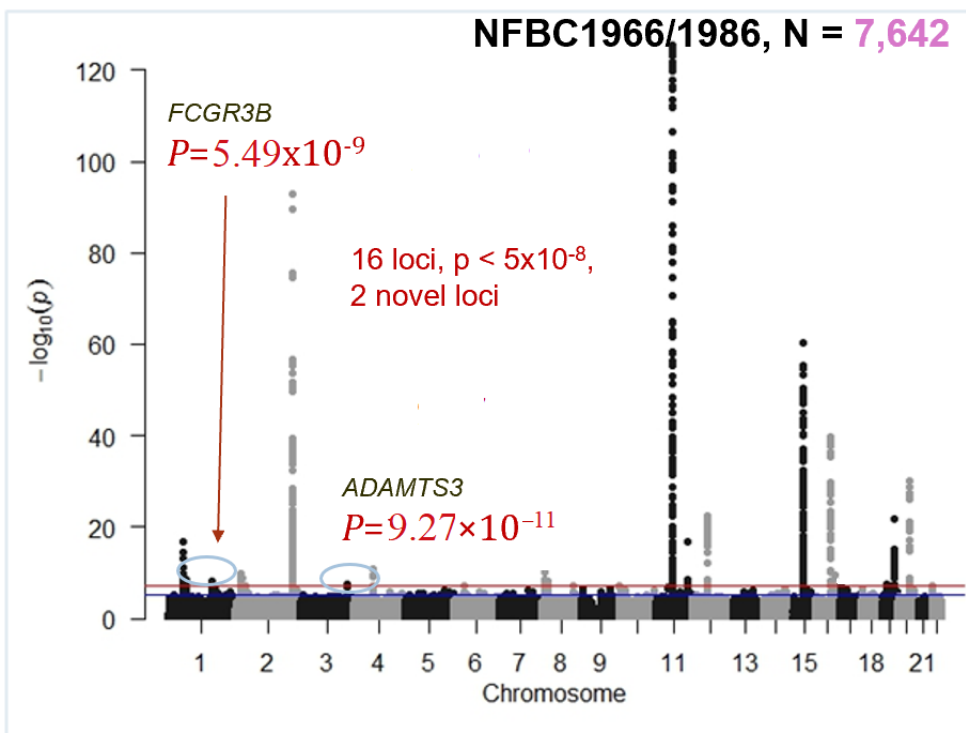


Figure 8.2 – Manhattan plot of MP-GWAS results using imputeSCOPA imputed data. The X-axis indicates the genomic position with the association P-value for each tested SNP on the Y-axis. The genome-wide significance threshold of  $5 \times 10^{-8}$  and the suggested significance threshold of  $5 \times 10^{-6}$  are represented with the red line and the blue line.

Table 8.2 – Signals detected in the full-model of meta-analysis MP-GWAS, consist of NFBC1966 and NFBC1986 cohorts using complete case analysis and imputeSCOPA imputed data.

| Chr | Top SNP     | Chr: position | Effect/ another allele | MA F (%) | P <sub>imp</sub>                                 | P <sub>cc</sub>         | Reported gene                                       | Previously associated trait                                                                                                                                                                                | Reference                                                                                                                      |
|-----|-------------|---------------|------------------------|----------|--------------------------------------------------|-------------------------|-----------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| 1   | rs11591147  | 1:55505647    | G/T                    | 3.5      | 1.11x10 <sup>-16</sup>                           | 2.15x10 <sup>-11</sup>  | PCSK9;<br>N/A                                       | Total cholesterol, LDL cholesterol, metabolite levels                                                                                                                                                      | Nagy et al., 2017; Kettunen et al., 2016; Surakka et al., 2015                                                                 |
| 1   | rs61803025  | 1: 161600591  | T/C                    | 12.9     | 5.49x10 <sup>-9</sup>                            | 5.68x10 <sup>-7</sup>   | FCGR3B                                              | Gamma immunoglobulin, total cholesterol, LDL cholesterol, metabolite levels                                                                                                                                | Kettunen et al., 2012                                                                                                          |
| 2   | rs1260326   | 2:27730940    | C/T                    | 28.5     | 2.97x10 <sup>-9</sup>                            | 3.65x10 <sup>-6</sup>   | GCKR                                                | Triglyceride levels, Chrohn's disease, inflammatory bowel disease, gout, renal overload gout                                                                                                               | Spracklen et al., 2017; de Lange et al., 2017; Nakayama et al., 2017                                                           |
| 2   | rs1047891   | 2:211540507   | C/A                    | 35.6     | 1.87x10 <sup>-95</sup>                           | 2.12x10 <sup>-78</sup>  | CPS1                                                | Glomerular filtration, metabolite levels, plasma, homocysteine levels, HDL cholesterol                                                                                                                     | Gorski et al., 2017; Kettunen et al., 2012; Williams et al., 2014; Willer et al., 2013                                         |
| 4   | rs181847072 | 4: 73134560   | C/G                    | 1.01     | 9.27x10 <sup>-11</sup>                           | 5.67x10 <sup>-7</sup>   | ADAMTS3                                             | Albumin levels, total cholesterol, LDL cholesterol, metabolite levels                                                                                                                                      | Kettunen et al., 2012; Inouye et al., 2012                                                                                     |
| 8   | rs4240624   | 8:9184231     | A/G                    | 13.7     | 3.36x10 <sup>-13</sup><br>3.92x10 <sup>-13</sup> | 8.48x10 <sup>-9</sup>   | PPP1R3B<br>; N/A                                    | HDL cholesterol, Schizophrenia                                                                                                                                                                             | Spracklen et al., 2017; Goes et al. 2015                                                                                       |
| 8   | rs60223219  | 8:19881801    | A/G                    | 9.7      | 2.37x10 <sup>-9</sup>                            | 5.47x10 <sup>-8</sup>   | MET<br>LPL                                          | Triglyceride levels;<br>Hypertriglyceridemia                                                                                                                                                               | Spracklen et al., 2017; Johansen et al., 2010                                                                                  |
| 11  | rs174564    | 11:61588305   | A/G                    | 4.3      | 3.81x10 <sup>-127</sup>                          | 6.92x10 <sup>-118</sup> | FADS1,<br>FADS2,<br>FADS3                           | Total cholesterol, LDL cholesterol, inflammatory bowel disease, glycerophospholipid levels, C-reactive protein, glycerophospholipid levels, -trans-fatty acid levels, triglyceride levels, HDL cholesterol | Spracklen et al., 2017; Liu et al., 2015; Draisma et al., 2015; Ligthart et al., 2016; Mozaffarian et al.; Willer et al., 2013 |
| 11  | rs964184    | 11:116648917  | C/G                    | 14.8     | 2.19x10 <sup>-17</sup>                           | 1.66x10 <sup>-14</sup>  | APOA1,<br>APOC3,AP<br>OA4,<br>APOA5;<br>N/A;<br>N/A | Total cholesterol, LDL cholesterol, age-related diseases hypertriglyceridemia                                                                                                                              | Spracklen et al., 2017; He et al., 2016; Ram et al., 2017                                                                      |

|    |            |             |        |      |                        |                        |                           |                                                                                     |                                                                   |
|----|------------|-------------|--------|------|------------------------|------------------------|---------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------|
| 12 | rs2638314  | 12:56866334 | T/A    | 16.2 | $4.16 \times 10^{-23}$ | $2.65 \times 10^{-20}$ | <i>GLS2</i>               | Metabolite levels, amino acid levels                                                | Kettunen et al., 2016; Draisma et al., 2015                       |
| 15 | rs1532085  | 15:58683366 | G/A    | 44.5 | $7.68 \times 10^{-61}$ | $1.19 \times 10^{-47}$ | <i>LIPC</i>               | HDL cholesterol metabolite levels, triglycerides total cholesterol, HDL cholesterol | Spracklen et al., 2017; Kettunen et al. 2016; Willer et al., 2013 |
| 16 | rs3764261  | 16:56993324 | C/A    | 16.7 | $2.29 \times 10^{-40}$ | $7.56 \times 10^{-37}$ | <i>CETP</i>               | HDL cholesterol, triglycerides, total cholesterol, HDL cholesterol                  | Nagy et al., 2017; Spracklen et al., 2017                         |
| 16 | rs77303550 | 16:72079657 | C/T    | 16.7 | $8.33 \times 10^{-8}$  | $4.43 \times 10^{-6}$  | <i>HP</i>                 | Metabolite levels                                                                   | Kettunen et al., 2016                                             |
| 19 | rs3841260  | 19:19756073 | AGCC/A | 5.5  | $6.57 \times 10^{-8}$  | $6.13 \times 10^{-7}$  | <i>ATP13A1</i>            | Response to fenofibrate                                                             | Irvin et al., 2016                                                |
| 19 | rs7412     | 19:45412079 | C/T    | 4.2  | $2.32 \times 10^{22}$  | $9.68 \times 10^{-20}$ | <i>APOE; APOC1, APOC2</i> | LDL cholesterol, HDL cholesterol, total cholesterol, LDL cholesterol                | Zhu et al., 2017a Nagy et al., 2017; Spracklen et al., 2017       |
| 20 | rs6073958  | 20:44551855 | T/C    | 19.5 | $5.74 \times 10^{-31}$ | $3.97 \times 10^{-19}$ | <i>PLTP</i>               | Metabolite levels, Triglycerides                                                    | Kettunen et al. 2016; Surakka et al., 2015                        |

## 8.4. Discussions and conclusions

### 8.4.1 Established signals in MP-GWAS

In this study, I run MP-GWAS analysis of 31 metabolites from the CC analysis and imputeSCOPA imputed data of NFBC1966 and NFBC1986 cohorts. I aimed to look for the associations of these 31 metabolites with common and low-frequency variants (MAF >1%) by employing MP-GWAS software, SCOPA. I detected 12 established signals with the CC data and 16 signals with the imputeSCOPA imputed data.

The well-established *FADS1/2/3* locus was detected in our MP-GWAS to be associated with lipids as previously has been identified from several studies ([Kettunen et al., 2012](#); [Malerba et al., 2008](#)). The top SNP identified in this locus in our study is rs174564 with p-value =  $6.92 \times 10^{-118}$  using the CC data and p-value =  $3.81 \times 10^{-127}$  using those with imputeSCOPA. This locus is known to regulate the unsaturation of FAs by encoding FA desaturase enzyme that removes two hydrogen atoms to insert a double bond at FA chains.

I identified a top SNP rs153208 in the *LIPC* gene on chromosome 15, as expected. This is also an established locus that is known to encode hepatic lipase (HL) or hepatic triglyceride lipase (HTGL), a lipolytic enzyme contributing to the regulation of triglycerides (TG) and phospholipids. The deficiency of this enzyme will disturb the body's ability to decompose fats (lipids).

In chromosome 12, we detected an intronic SNP (rs2638314) as the top variant in this chromosome. Another SNP (rs2657879) in the same locus was previously reported in another study ([Kettunen et al., 2012](#)), which is incomplete LD with rs2638314. These SNPs are in the *GLS2* gene, which has a function to encode glutaminase, an enzyme that plays a vital role in decomposing glutamine as energy and biosynthetic precursor.

In addition, the association of the *GLS2* gene with atherosclerosis was also identified (Kettunen et al., 2012).

In chromosome 19, I detected an SNP rs3841260 in *ATP13A1* (ATPase 13A1) gene that has high LD ( $r^2 = 0.84$ ) with rs17216525 in *PBX4* (pre-B-cell leukaemia homeobox 4) gene, a variant which has previously been reported to have an association with triglyceride levels (Kathiresan et al., 2008). Another study also pointed out this variant in association with lipids and response to fenofibrate treatment, commonly used to lower triglyceride and LDL levels, yet it raises HDL cholesterol levels (Irvin et al., 2016).

I also identified several established loci previously found to be associated with cardiometabolic traits, including SNPs located in the *APOE*, *APOA1*, *APOC3*, *APOA4*, *APOA5*, *CETP*, *HP*, *CPS1*, *GCKR*, *MET*, *PPP1R3B*, *HP*, *LPL*, *PCSK9*, and *PLTP* genes.

#### 8.4.2 Additional signals detected in the imputeSCOPA imputed data

I detected two additional signals within our imputeSCOPA imputed data, which have interesting associations with metabolic outcomes. The first novel signal in this study is rs61803025 lying in the exon of chromosome 1 within the *FCGR3B* (Fc fragment of IgG receptor IIIb) gene. The imputation quality score (IQS) for this variant is 1, indicating no uncertainty in the imputed genotypes. It is a common variant with MAF = 12.9% in the NFBC1966 and MAF = 12.6% in the NFBC1986, having C/T as a major/minor allele.

The *FCGR3B* gene, which is also known as CD16b (Cluster of Differentiation 16b), is a protein-encoding gene related to the Fc region of gamma immunoglobulins (IgG). Although there is no literature on the established SNPs in this gene, another marker

(rs67418890), which is in LD with the SNP ( $r^2$  of 0.60), has been reported to have an association with lipid measurement ([Kettunen et al., 2016](#)).

The second additional signal detected in this analysis is rs181847072 in an intergenic region (IGR) closest to *ADAMTS3* (ADAM metalloproteinase with thrombospondin type 1 motif 3) gene in chromosome 4. This SNP is a low-frequency variant with MAF = 1.01% in NFBC1966 and MAF = 1.30% in NFBC1986. However, the SNP has high IQS ( $r^2 = 0.96$ ), suggesting that this is not a spurious association due to genotype imputation. Furthermore, based on literature, several variants lying in this *ADAMTS3* gene, such as downstream SNP (rs115136538) of the variant, are associated with albumin levels ([Kettunen et al., 2012](#)). An upstream SNP (rs1851025) of the variant also indicates an association with albumin and lipid traits ([Inouye et al., 2012](#)). However, these SNPs are not in LD with the SNP ( $r^2 < 0.01$ ), indicating a possible novel distinct signal in this locus.

## 9. Discussion

### 9.1. General discussion

Over a decade, GWAS have transformed the landscape of glycaemic trait genetics. We collaborated with other MAGIC investigators to combine data from multiple GWAS to pinpoint additional sites that affect glycaemic and metabolic properties. Our team previously contributed to a meta-analysis of FI and FG for European ancestry and trans-ethnic meta-analysis using single-phenotype GWAS for each trait. Therefore, for this thesis, I calculated PRS to assess the extent of variance explained for FG and FI using NFBC1966 and NFBC1986 cohorts. I found that the variance explained in NFBC1966 was slightly higher than in NFBC1986; however, the results were consistently similar to the other cohorts within MAGIC. Finally, it can be concluded that by using a trans-ethnic approach and by increasing the sample size and thus the number of identified loci, we were able to increase the variance explained for FG from 1.5% to 4.3%, as compared to previous studies.

However, another way in the recent development of GWAS methodology is to increase the power of association with small or moderate sample size using MP-GWAS that allows the analysis of several phenotypes at once ([Mägi et al., 2017](#); [O'Reilly et al., 2012](#)). One example of this methodology uses the reverse regression method by treating the phenotypes as the predictors and the genotypes as the regression outcome. Our team utilised this reverse regression methodology by developing SCOPA software ([Mägi et al., 2017](#)). Therefore, the subsequent analysis within this project was conducting MP-GWAS for FA metabolisms using SCOPA.

Many studies have examined the favourable effects of diets rich in omega-3 FAs to attenuate cardiometabolic disease risk, especially CVD (Philip C. Calder, 2015; Mozaffarian & Wu, 2011; Simopoulos, 2002). Before this project, our team also had studied the genetic contribution to FA levels via MP-GWAS of omega-3, 6, 7/9, and  $\alpha$ -PUFA. The meta-analysis from NFBC1966 and NFBC1986 (N=8,004) conducted by our team detected 10 signals associated with FAs ( $P < 5 \times 10^{-8}$ ) at *PCSK9*, *GCKR*, *LPXN*, *FADS1*, *GPR137*, *ZNF259*, *LIPC*, *PDXDC1*, *PBX4*, and *APOE*. For this thesis, I carried out subsequent, direct conditional MP-GWAS to dissect whether these 10 omega FA loci harbour more than one distinct signal affecting FA levels. After direct conditioning, I detected two distinct signals within *FADS1* and *PBX4* loci and three distinct signals in *LIPC*, while for *PCSK9*, *GCKR*, *LPXN*, *GPR137*, *ZNF259*, *PDXDC1*, and *APOE*, the conditional analysis showed evidence for only one signal affecting FA level. I refined the genetic architecture of FA metabolism by detecting additional signals at these loci. This novel methodology could be applied to a number of genomic loci, featuring collocated signals for different traits, thus containing suggestive multi-phenotype effects.

Although novel loci affecting FA levels could be detected using MP-GWAS studies, I found that missing data issues in multi-phenotype studies also influenced the power of locus discovery. As the number of phenotypes increases within MP-GWAS, the chance of missing data within individuals will also increase gradually. Therefore, CC analysis is not suggested to be applied within the MP-GWAS framework, as it will drastically reduce the sample size and attenuate the power for locus discovery. As in the same case as in genotype imputation, an advanced phenotype imputation method should be applied prior to the analysis. Previously in the study of FA level metabolisms, our team employed the SI method for phenotype imputation as at that time, this

imputation method was relatively easy to be implemented in R, while SI imputation assumes that the missing data are at least MAR, which is a better approach compared to CC analysis that assumes MCAR which is basically never a case. However, SI imputation assumes that the imputed values are all treated as if they were true, which is never the case. It also overestimates the correlations but underestimates the variances.

In a real dataset, the most commonly missing data is MAR, while some may contain a mixed type of MCAR-MAR-MNAR assumption or the subset of these three missing mechanisms (R. J. Little & Rubin, 2015). Therefore, I investigated better imputation approaches by comparing the performances of SI imputation with MI imputation, a proposed approach in mainstream statistics to solve the uncertainty of SI imputation. I also compared their performance with imputation approaches from machine learning, i.e., kNN, QRILC, and RF. To assess their performances, I compared these phenotype-imputation approaches with the standard CC analysis within the MP-GWAS framework by conducting an extensive simulation study with different scenarios of the proportion of missingness, different missing data mechanisms, different MAF, and different correlations between the phenotypes. I evaluated all the imputation methods tested by comparing their RMSE of  $\beta$ s from the imputed methods or CC analysis vs RMSE of  $\beta$ s from complete data using linear regression. The best imputation method observed was then applied to the missing data in real data analysis before submitting it into SCOPA software.

I found that in the MCAR scenario, RF and MI using AMELIAII were the best performing methods as indicated by their closest values to the estimates coming from full data analysis, whereas in MAR, RF was the best one. My findings are in line with

a previous study that demonstrated RF was the best under MAR and MCAR scenario for metabolomics data (Kokla et al., 2019), although I found that RF still works well under MNAR. I also found that QRILC was the best method under MNAR, as it was designed for MNAR only; therefore, under MCAR and MAR, QRILC was considerably worse than the other imputation methods. In general, I can say that RF was the best among the others as it works in all three different missing data mechanisms, and it can handle mixed types of data, unlike QRILC. Moreover, RF does not need parameter tuning as the parameter tuning will not affect the accuracy of parameter estimates but only influence the computational time.

However, in my simulated data, when I imputed high-dimensional phenotype data of 500k individuals, such as those from UK Biobank, RF was not the option as it needs to grow the trees, which is computationally slow. Moreover, there was only one implementation of RF imputation to date, missForest in the software package R, as RF software is usually only designed for regression, classification, and survival analysis, such as Ranger, the fastest implementation of RF written in C++.

For this reason, I developed a faster RF imputation for high-dimensional data based on Ranger, written in C++. I named it imputeSCOPA and compared its performance with missForest using simulated data. I found that imputeSCOPA was 30%-75% faster compared to missForest in 5k to 50k data. In addition, missForest struggled to impute high-dimensional data with over 100k individuals, whereas our imputeSCOPA is the only method that can handle the imputation of studies with over 100k individuals.

Finally, I tested imputeSCOPA into real data analysis and found that imputed data with RF can boost the power for locus discovery.

## 9.2.Conclusion

Meta-analysis GWAS within MAGIC provided us with the best prediction for the glucose-related traits that can be made when considering variation in multiple genetic variants by combining the results from multiple trans-ethnic studies. However, the recent development of GWAS methodology proposed another way to increase the power of association with small or moderate sample size using MP-GWAS. By adopting MP-GWAS, we have demonstrated a better approach to improve the power of locus discovery in a relatively small dataset by analysing FA levels in NFBC1966 and NFBC1986. I detected novel loci and independent signals through subsequent, direct conditional analysis. This novel methodology could be applied to a number of complex traits, featuring collocated signals for different traits and thus containing suggestive multi-phenotype effects.

Through extensive simulation studies and novel software development, I introduced imputeSCOPA written in C++ application and its corresponding R package to improve the performance of MP-GWAS by imputing the missing data of the phenotypes used in the analysis rather than using only the complete data such as in CC analysis.

In the simulation studies, I demonstrated that the computational time and memory efficiency of imputeSCOPA were the fastest compared to the other RF applications. Moreover, imputeSCOPA is the only RF application available up to date that can handle high-dimensional data with more than 100k individuals, such as those in the UK Biobank. I also demonstrated that imputeSCOPA could boost the power of association by implementing the imputation of lipid analysis within MP-GWAS.

## 9.3.Strengths and limitations

### 9.3.1. Strengths

We used reverse regression of the MP-GWAS method with the full model that does not require correction for multiple testing as it used the same model (Kaakinen et al., 2017); thus, it will increase the statistical power. SCOPA is analysing each phenotype, and each SNP is regressed in a single test on several phenotypes, like the Multiphen approach. A primary genome-wide study of all phenotypes in the method is implemented, preventing all potential phenotype combinations from being corrected at the reach locus for multiple testing. There is then the option to evaluate signals reaching genome-wide significance for interaction with all combinations of phenotype models. Specifically, this includes a complete model of all phenotypes studied, all possible combinations of various phenotypes, and individual phenotypes as univariate models. This is useful because most MP techniques do not provide univariate comparisons, which usually means that single-phenotype techniques are still preferred (O'Reilly et al., 2012). This methodology can also analyse both binary and qualitative measurements, which does not need the prior assumption of phenotype distribution.

I extended the applicability of MP-GWAS to handle the missing data problem by developing imputeSCOPA, the fastest RF imputation software to date. Moreover, the RF implemented in imputeSCOPA has several advantages, including the least biased parameter estimates, as confirmed via my simulated study, and its ability to handle a mixture of high-dimensional data.

As I imputed the data using imputeSCOPA, I could retain the full dataset to boost the power of statistical association without shrinking the size of our data, such as those

using CC analysis. I lose about 330 values per variable on average due to the missing data removal with CC analysis.

In terms of the data used in this study, I managed to avoid the population effects by using NFBC cohorts, which are rich with phenotype and genotype data of individuals from the same region separated by birth years. The PCs (for NFBC1966) and MDS (for NFBC1986) were calculated prior to the analysis to prevent bias due to population stratification.

The metabolites used in this study were also available from the two cohorts despite a few differences of variables in both datasets. Therefore, we could impute with the same predictors and manage to replicate and meta-analyse the two cohorts using the same statistical models.

### **9.3.2. Limitations**

While my study focused on cardiometabolic traits, which are usually related to the risks of late-onset diseases ([J. Deelen et al., 2014](#); [Johnson, Li, Kuh, & Hardy, 2015](#)), the data provided by NFBC1986 has younger participants at 15/16-year-old compared to the 31-year-old individuals in NFBC1966. Younger participants could be a weakness of this NFBC1986 data as they did not have any report of cardiometabolic diseases; thus, it can decrease the power of these data to detect associations.

In addition, I could not properly control several variables in NFBC1986 data, such as information on lipid medication and pregnancy status, which were missing. Although teenagers in this study are highly unlikely to have such medication or to be pregnant at this age; however, an accurate report should be provided. Therefore, in NFBC1986, it is reasonable to assume that the differences are not detectable to cause a significant

association as they are teenagers compared to the study of the older population in NFBC1966; thus, younger age may truly have an impact on the results. Also, certain differences were found in the data obtained from the questionnaires, complicating direct comparison and data handling between the cohorts.

I also found that our installed imputeSCOPA worked fine on Imperial HPC, but then after updates to the HPC system, triggered an error that may occur due to dependency issues. This dependency affects not only imputeSCOPA but also SCOPA software. I have contacted the IT admin to solve this problem; however, we still cannot figure out which system updates that triggered this issue. However, I have tested imputeSCOPA in my local UNIX machine and other HPC platform, both of them run smoothly without any problem. An investigation is needed to ensure that imputeSCOPA works on all UNIX platforms, especially on the Imperial HPC machine.

Finally, although RF can handle MNAR assumption, I found in my simulated study that QRILC was the best approach to be implemented considering omics data, which has a great chance to have such kind of missing data mechanism as lower values usually cannot be detected due to technical limitations. However, RF may not be the best option to missing data under MNAR assumption; therefore, imputeSCOPA should be extended to apply QRILC in such conditions to facilitate the best possible imputation under all scenarios of missing data, although the real data is rarely to have only one type of missing mechanism.

#### **9.4.Future directions**

This thesis research provided a solid fundamental concept in developing MPA to increase the statistical power in elucidating the genetic architecture of cardiometabolic traits.

All of the works that have been done in this thesis were presented at several conferences. The results of the meta-analysis GWAS of glycaemic traits is currently under review in *Nature genetics*, while the imputeSCOPA software and MP-GWAS of FA metabolism will be submitted shortly.

The first concrete step towards the MP-GWAS is conducting a much broader replication to continue the presented work. In order to increase the sample size of the replication cohort NFBC1986, ExomeChip-specific MDS coordinates should be obtained from the NFBC1986, adding about 700 additional replication samples. In the directly genotyped results, only half of 34 GWAS discovered SNPs had proxies.

Replication into an individual sample of a different demographic origin from a different race will confirm the discovery signals. The detection of genetic effects was encouraged by homogeneous population cohorts, but it was hypothesized that common variants with small effect sizes should be observable throughout the populations ([Mahajan et al., 2014](#)). This would solidify its role in lipids metabolisms by replicating a discovery signal in a large cohort of different ethnic origins.

In addition, conducting a large-scale meta-analysis will enlarge the sample sizes. Since the SCOPA enables associated software for meta-analysis, where summary statistics of multiple studies are aggregated into combined results, increasing the sample size by combining the data with other European cohorts will increase the power of association to detect more novel loci affecting lipids.

In this report, very important aspects of data processing are addressed and how they can be properly analysed to reach their potential. The research opens new opportunities for small samples using NFBC cohorts to discover novel genetic variants and lays the groundwork for further developmental studies in the MP-GWAS setting.

By design, many GWASs' ultimate goals are to provide candidate genes for subsequent experiments that will ultimately lead to a better understanding of how phenotypes are affected by these genes. Thus, practical studies of signals that are identified and successfully replicated will ideally be included in the follow-up to this report. However, a functional experiment where the genes are knocked out, and omega lipid levels are tested may provide clues to the role of the metabolism of lipids in satiety and obesity. Functional experiments relate the calculated associations to actual biological phenomena that fulfil the objective of completely validating the role of genetic variants in lipids' metabolism.

In the future, it would be interesting to explore longitudinal genetic effects on the NFBC1966 cohort. The NFBC1966 cohort collected phenotype and genotype data at different ages of the same set of individuals, which means it is relevant to the study area. The discrepancy between the phenotype measurements in our data taken at the age of 31 and the data obtained at the age of 46 will be separated by the transition in time that will generate a new longitudinal phenotype. The NFBC1986 has already gathered data over the life cycle of the cohort and offers a possible replication opportunity. The longitudinal analysis may offer a glimpse into genetic influences that have time-related trajectories, particularly of concern to us as late-onset diseases that worsen or appear with age are often cardiometabolic diseases ([Hardy, Lawlor, & Kuh, 2015](#); [Warrington et al., 2015](#)). This will offer a possible starting point for prospective drug treatments to address particular genetic effects specific to the population that may raise the future risk of cardiometabolic disease.

The other risk factors that are also strong predictors of T2D and positively associated with CVD risk are BCAAs ([Tobias et al., 2018](#)). It was discovered that total BCAAs

were associated with a 23% greater risk of total CVD, greater than those of FA (Tobias et al., 2018). A previous study found that fasting plasma BCAA levels were associated with higher homeostatic model assessment insulin-resistant HOMA-IR (Mahendran et al., 2017). However, the PRS for circulating BCAA levels was not associated with FI levels or HOMA-IR. By contrast, the insulin-resistance-increasing PRS was significantly associated with increased BCAA levels. This study indicated that HOMA-IR is causally related to higher circulating fasting BCAA levels. They suggested that higher BCAA levels do not have a causal effect on insulin resistance, while increased insulin resistance drives higher circulating fasting BCAA levels. Utilising the MP-GWAS framework would help detect the association between BCAAs, HOMA-IR, and FI levels, suggesting such a multi-phenotype effect.

As an individual PhD project, the range of imputation methods explored in this thesis is only a fraction of the possible options available. For example, I limited the number of imputation methods being compared in the simulated studies, these being the most widely used in the epidemiological and clinical data in the literature. Other ML imputation methods such as support vector machine (SVM) and deep learning could be explored in the future.

The imputeSCOPA developed in this thesis based on the RF approach that already well studied can handle the mixed type of data and non-normality without assuming any distributions of the variables. Therefore, in the simulated data, different scenarios of mixed categorical and continuous data were not investigated as the simulated data was created only for continuous data as in the metabolomic data. The different distributions of the data, including the high-skewed data and outliers, were also not investigated. However, I later found a recent study that argues this claim (Hong &

Lynn, 2020). They concluded that RF-based imputation leads to underperformance with outcome-dependent MAR and highly skewed data (Hong & Lynn, 2020). Thus, it should not be indiscriminately recommended for missing data imputation as a universal recommendation (Hong & Lynn, 2020). Proper analysis requires careful examination of the mechanism of missing data and the relationships between data variables (Hong & Lynn, 2020). However, this conclusion is also valid for other imputation methods, such as SI and MI, since they assume that the data should be normally distributed. Shah et al. (A. D. Shah et al., 2014). Moreover, data with MNAR assumptions are typically left-skewed, and I found that RF still works well on that missing data problem in my simulated study. However, I did not evaluate the distribution of the simulated data.

Moreover, RF has been claimed to "handle missing values successfully, particularly in datasets, including different variable types" (Stekhoven & Bühlmann, 2012a). In addition, Waljee et al. concluded that it 'is a highly accurate imputation method for missing laboratory data and exceeds other traditional imputation techniques' (Waljee et al., 2013). These two studies concentrated on MCAR data. While RF was analysed by other researchers in the presence of MAR (Ramosaj & Pauly, 2019) and MNAR data (Ishwaran et al., 2008), all these studies evaluated RF in terms of its predictive accuracy.

In summary, I developed the least biased imputation method in the MP-GWAS framework, and this method only focusses on handling bias due to missing data mechanisms. However, observational research and missing values in covariates frequently suffer from confounding and may need to be addressed.

Bias produces an invalid association because of missing mechanisms, whereas confounding defines a true association, potentially misleading. Methods for the propensity score, first proposed by Rosenbaum and Rubin ([Amos et al., 1986](#)), need to be used to deal with confounding in the analysis.

Due to time limits, I was unable to implement QRILC in imputeSCOPA. We hope to extend the applicability of imputeSCOPA to handle specific data that has been known initially only contains MNAR assumptions using QRILC. However, most data assume MAR assumption or the mixture of MAR-MCAR-MNAR; therefore, the current version of imputeSCOPA is sufficient to be used for large-scale studies. By improving the accuracy of imputing missing data, a greater number of improved MP-GWAS can be achieved to boost the power for loci discovery, targeting high dimensional data in largescale settings like in the UK Biobank.

## 10. References

- A Century of Mendelism in Human Genetics*. (2000). *A Century of Mendelism in Human Genetics*. <https://doi.org/10.4324/9780203392157>
- Amos, C. I., Wilson, A. F., Rosenbaum, P. A., Srinivasan, S. R., Webber, L. S., Elston, R. C., ... Rao, D. C. (1986). An approach to the multivariate analysis of high-density-lipoprotein cholesterol in a large kindred: The Bogalusa Heart Study. *Genetic Epidemiology*, 3(4), 255–267. <https://doi.org/10.1002/gepi.1370030406>
- Araújo, J., Cai, J., & Stevens, J. (2019). Prevalence of Optimal Metabolic Health in American Adults: National Health and Nutrition Examination Survey 2009-2016. *Metabolic Syndrome and Related Disorders*. <https://doi.org/10.1089/met.2018.0105>
- Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. M. E., ... Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*. <https://doi.org/10.1016/j.cell.2018.12.019>
- Auer, P. L., & Lettre, G. (2015). Rare variant association studies: Considerations, challenges and opportunities. *Genome Medicine*. <https://doi.org/10.1186/s13073-015-0138-2>
- Azevedo, C. F., de Resende, M. D. V., e Silva, F. F., Nascimento, M., Viana, J. M. S., & Valente, M. S. F. (2017). Population structure correction for genomic selection through eigenvector covariates. *Crop Breeding and Applied Biotechnology*. <https://doi.org/10.1590/1984-70332017v17n4a53>
- Bansal, V., Libiger, O., Torkamani, A., & Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg2867>
- Barclay, A. W., Petocz, P., McMillan-Price, J., Flood, V. M., Prvan, T., Mitchell, P., & Brand-Miller, J. C. (2008). Glycemic index, glycemic load, and chronic disease risk - A metaanalysis of observational studies. *American Journal of Clinical Nutrition*.
- Bell, J. A., Kivimaki, M., & Hamer, M. (2014). Metabolically healthy obesity and risk of incident type 2 diabetes: A meta-analysis of prospective cohort studies. *Obesity Reviews*. <https://doi.org/10.1111/obr.12157>
- Belloy, M. E., Napolioni, V., & Greicius, M. D. (2019). A Quarter Century of APOE and Alzheimer's Disease: Progress to Date and the Path Forward. *Neuron*, 101(5), 820–838. <https://doi.org/10.1016/j.neuron.2019.01.056>
- Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16(Suppl 3). <https://doi.org/10.1186/s12911-016-0318-z>
- Bosch, A., Zisserman, A., & Muñoz, X. (2007). Image classification using random forests and ferns. In *Proceedings of the IEEE International Conference on*

- Computer Vision*. <https://doi.org/10.1109/ICCV.2007.4409066>
- Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Frigessi, A., & Lingjærde, O. C. (2007). Predicting survival from microarray data - A comparative study. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btm305>
- Breiman, L. (2001). Random Forests. *Machine Learning*.
- Brunzell, J. D., Davidson, M., Furberg, C. D., Goldberg, R. B., Howard, B. V., Stein, J. H., & Witztum, J. L. (2008). Lipoprotein Management in Patients With Cardiometabolic Risk. *Journal of the American College of Cardiology*. <https://doi.org/10.1016/j.jacc.2008.02.034>
- Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*. <https://doi.org/10.1093/aje/kwq260>
- Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1002822>
- Calder, P. C. (2015). Functional Roles of Fatty Acids and Their Effects on Human Health. *Journal of Parenteral and Enteral Nutrition*, 0148607115595980-. <https://doi.org/10.1177/0148607115595980>
- Calder, Philip C. (2015). Functional Roles of Fatty Acids and Their Effects on Human Health. *Journal of Parenteral and Enteral Nutrition*. <https://doi.org/10.1177/0148607115595980>
- Calder, Philip C. (2015). Marine omega-3 fatty acids and inflammatory processes: Effects, mechanisms and clinical relevance ☆. *BBA - Molecular and Cell Biology of Lipids*, 1851(4), 469–484. <https://doi.org/10.1016/j.bbalip.2014.08.010>
- Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., ... Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*. <https://doi.org/10.1038/nature13907>
- Cho, B. H. S., Dokko, R. C., & Chung, B. H. (2002). Oleic, linoleic and linolenic acids enhance receptor-mediated uptake of low density lipoproteins in Hep-G2 cells. *The Journal of Nutritional Biochemistry*, 13(6), 330–336. [https://doi.org/10.1016/S0955-2863\(02\)00174-2](https://doi.org/10.1016/S0955-2863(02)00174-2)
- Choi, J., Dekkers, O. M., & le Cessie, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*. <https://doi.org/10.1007/s10654-018-0447-z>
- Choi, Y. H., Chowdhury, R., & Swaminathan, B. (2014). Prediction of hypertension based on the genetic analysis of longitudinal phenotypes: A comparison of different modeling approaches for the binary trait of hypertension. In *BMC Proceedings*. <https://doi.org/10.1186/1753-6561-8-S1-S78>
- Craig, J. (2008). Complex Diseases: Research and Applications. Retrieved July 12, 2019, from <https://www.nature.com/scitable/topicpage/complex-diseases->

- Craig, S. J. C., Kenney, A. M., Lin, J., Paul, I. M., Birch, L. L., Savage, J., ... Makova, K. D. (2019). Polygenic risk score based on weight gain trajectories is a strong predictor of childhood obesity. *BioRxiv*. <https://doi.org/10.1101/606277>
- Cui, X., Sha, Q., Zhang, S., & Chen, H.-S. (2009). A combinatorial approach for detecting gene-gene interaction using multiple traits of Genetic Analysis Workshop 16 rheumatoid arthritis data. *BMC Proceedings*. <https://doi.org/10.1186/1753-6561-3-s7-s43>
- Dahl, A., Iotchkova, V., Baud, A., Johansson, Å., Gyllenstein, U., Soranzo, N., ... Marchini, J. (2016). A multiple-phenotype imputation method for genetic studies. *Nature Genetics*, 48(4), 466–472. <https://doi.org/10.1038/ng.3513>
- Das, S., Abecasis, G. R., & Browning, B. L. (2018). Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics*. <https://doi.org/10.1146/annurev-genom-083117-021602>
- Deelen, J., Beekman, M., Uh, H. W., Broer, L., Ayers, K. L., Tan, Q., ... Slagboom, P. E. (2014). Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Human Molecular Genetics*, 23(16), 4420–4432. <https://doi.org/10.1093/hmg/ddu139>
- Deelen, P., Menelaou, A., van Leeuwen, E. M., Kanterakis, A., van Dijk, F., Medina-Gomez, C., ... Swertz, M. A. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the “Genome of The Netherlands”. *European Journal of Human Genetics: EJHG*, 22(11), 1321–1326. <https://doi.org/10.1038/ejhg.2014.19>
- Deigner, H. P., & Kohl, M. (2018). *Precision Medicine: Tools and Quantitative Approaches*. *Precision Medicine: Tools and Quantitative Approaches*.
- Dempfle, A., Scherag, A., Hein, R., Beckmann, L., Chang-Claude, J., & Schäfer, H. (2008). Gene–environment interactions for complex traits: definitions, methodological requirements and challenges. *European Journal of Human Genetics*, 16(10), 1164–1172. <https://doi.org/10.1038/ejhg.2008.106>
- Deng, Y. (2017). Conditional analysis of multiple quantitative traits based on marginal GWAS summary statistics, (November 2016), 427–436. <https://doi.org/10.1002/gepi.22046>
- Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., ... Crawford, D. C. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btq126>
- Dorajoo, R., Sun, Y., Han, Y., Ke, T., Burger, A., Chang, X., ... Friedlander, Y. (2015). A genome-wide association study of n-3 and n-6 plasma fatty acids in a Singaporean Chinese population. *Genes & Nutrition*, 10(6), 53. <https://doi.org/10.1007/s12263-015-0502-2>
- Draisma, H. H. M., Pool, R., Kobl, M., Jansen, R., Petersen, A.-K., Vaarhorst, A. A.

- M., ... Boomsma, D. I. (2015). Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature Communications*, 6, 7208. <https://doi.org/10.1038/ncomms8208>
- Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.1003348>
- Dudley, J. T., Kim, Y., Liu, L., Markov, G. J., Gerold, K., Chen, R., ... Kumar, S. (2012). Human genomic disease variants: A neutral evolutionary explanation. *Genome Research*. <https://doi.org/10.1101/gr.133702.111>
- Duncan, L., Shen, H., Gelaye, B., Meijssen, J., Ressler, K., Feldman, M., ... Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*. <https://doi.org/10.1038/s41467-019-11112-0>
- Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*. <https://doi.org/10.1016/j.jclinepi.2013.09.009>
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Ishwaran, H., Knight, K., ... Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*. <https://doi.org/10.1214/009053604000000067>
- Emdin, C. A., Khera, A. V., Natarajan, P., Klarin, D., Zekavat, S. M., Hsiao, A. J., & Kathiresan, S. (2017). Genetic association of waist-to-hip ratio with cardiometabolic traits, type 2 diabetes, and coronary heart disease. *JAMA - Journal of the American Medical Association*. <https://doi.org/10.1001/jama.2016.21042>
- Emwas, A. H. M. (2015). The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods in Molecular Biology*. [https://doi.org/10.1007/978-1-4939-2377-9\\_13](https://doi.org/10.1007/978-1-4939-2377-9_13)
- Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btu848>
- Faculty of Medicine University of Oulu. (2015a). 31-year follow-up study, clinical examination.
- Faculty of Medicine University of Oulu. (2015b). NFBC 1986 data collection.
- Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The ( in ) famous GWAS P -value threshold revisited and updated for low-frequency variants, (November 2015), 1202–1205. <https://doi.org/10.1038/ejhg.2015.269>
- Fedak, K. M., Bernal, A., Capshaw, Z. A., & Gross, S. (2015). Applying the Bradford Hill criteria in the 21st century: How data integration has changed causal inference in molecular epidemiology. *Emerging Themes in Epidemiology*. <https://doi.org/10.1186/s12982-015-0037-4>
- Galesloot, T. E., Van Steen, K., Kiemeneij, L. a L. M., Janss, L. L., & Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods.

- PLoS ONE*, 9(4), 1–8. <https://doi.org/10.1371/journal.pone.0095923>
- Gelman, A., Hill, J., Gelman, A., & Hill, J. (2010). Missing-data imputation. In *Data Analysis Using Regression and Multilevel/Hierarchical Models*. <https://doi.org/10.1017/cbo9780511790942.031>
- Geneletti, S., Gallo, V., Porta, M., Khoury, M. J., & Vineis, P. (2013). Assessing causal relationships in genomics: From Bradford Hill criteria to complex gene-environment interactions and directed acyclic graphs. In *Epigenetics and Pathology: Exploring Connections between Genetic Mechanisms and Disease Expression*. <https://doi.org/10.1201/b16304>
- Glazier, A. M., Nadeau, J. H., & Aitman, T. J. (2002). Genetics: Finding genes that underline complex traits. *Science*. <https://doi.org/10.1126/science.1076641>
- Grajeda-Iglesias, C., & Aviram, M. (2018). Specific Amino Acids Affect Cardiovascular Diseases and Atherogenesis via Protection against Macrophage Foam Cell Formation: Review Article. *Rambam Maimonides Medical Journal*. <https://doi.org/10.5041/rmmj.10337>
- Groop, L., Mäkiperna, A., Stenman, S., DeFronzo, R. A., & Teppo, A. M. (1990). Urinary excretion of kappa light chains in patients with diabetes mellitus. *Kidney International*. <https://doi.org/10.1038/ki.1990.94>
- Grund, S. M. (2016). Metabolic syndrome update. *Trends in Cardiovascular Medicine*. <https://doi.org/10.1016/j.tcm.2015.10.004>
- Grund, S. M., Brewer, H. B., Cleeman, J. I., Smith, S. C., & Lenfant, C. (2004). Definition of Metabolic Syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association Conference on Scientific Issues Related to Definition. In *Circulation*. <https://doi.org/10.1161/01.CIR.0000111245.75752.C6>
- Grund, S. M., Hansen, B., Smith, S. C., Cleeman, J. I., & Kahn, R. A. (2004). Clinical Management of Metabolic Syndrome: Report of the American Heart Association/National Heart, Lung, and Blood Institute/American Diabetes Association Conference on Scientific Issues Related to Management. In *Circulation*. <https://doi.org/10.1161/01.CIR.0000112379.88385.67>
- Guan, W., Steffen, B. T., Lemaitre, R. N., Wu, J. H. Y., Tanaka, T., Manichaikul, A., ... Steffen, L. M. (2014). Genome-Wide association study of plasma n6 polyunsaturated fatty acids within the cohorts for heart and aging research in genomic epidemiology consortium. *Circulation: Cardiovascular Genetics*, 7(3), 321–331. <https://doi.org/10.1161/CIRCGENETICS.113.000208>
- Haidich, A. B. (2010). Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1), 29–37. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049418/>
- Hardy, R., Lawlor, D. A., & Kuh, D. (2015). A life course approach to cardiovascular aging. *Future Cardiology*. <https://doi.org/10.2217/fca.14.67>
- Hastie, A. T., Tibshirani, R., & Narasimhan, B. (2018). Package ‘impute’, 1–5.
- Hebbring, S. J. (2014). The challenges, advantages and future of phenome-wide association studies. *Immunology*. <https://doi.org/10.1111/imm.12195>

- Hidayat, K., Yang, J., Zhang, Z., Chen, G. C., Qin, L. Q., Eggersdorfer, M., & Zhang, W. (2018). Effect of omega-3 long-chain polyunsaturated fatty acid supplementation on heart rate: A meta-Analysis of randomized controlled trials. *European Journal of Clinical Nutrition*. <https://doi.org/10.1038/s41430-017-0052-3>
- Holliday, J. A., Wang, T., & Aitken, S. (2012). Predicting adaptive phenotypes from multilocus genotypes in sitka spruce (*picea sitchensis*) using random forest. *G3: Genes, Genomes, Genetics*. <https://doi.org/10.1534/g3.112.002733>
- Honaker, J., King, G., & Blackwell, M. (2011). AMELIA II : A Program for Missing Data. *Journal Of Statistical Software*, 45(7), 1–54. <https://doi.org/10.1.1.149.9611>
- Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*. <https://doi.org/10.1186/s12874-020-01080-1>
- Horikoshi, M., Mägi, R., van de Bunt, M., Surakka, I., Sarin, A. P., Mahajan, A., ... Morris, A. P. (2015). Discovery and Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation. *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.1005230>
- Hormozdiari, F., Zhu, A., Kichaev, G., Ju, C. J. T., Segrè, A. V., Joo, J. W. J., ... Eskin, E. (2017). Widespread Allelic Heterogeneity in Complex Traits. *American Journal of Human Genetics*. <https://doi.org/10.1016/j.ajhg.2017.04.005>
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*. <https://doi.org/10.1198/106186006X133933>
- Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., ... Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1005589>
- Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., ... Eichler, E. E. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*. <https://doi.org/10.1101/gr.214007.116>
- Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., ... Suhre, K. (2010). A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics*. <https://doi.org/10.1038/ng.507>
- Inzucchi, S. E., Bergenstal, R. M., Buse, J. B., Diamant, M., Ferrannini, E., Nauck, M., ... Matthews, D. R. (2013). Erratum: Management of hyperglycaemia in type 2 diabetes: A patient-centered approach. Position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD) (*Diabetologia* (2013) 56 (680)). *Diabetologia*. <https://doi.org/10.1007/s00125-012-2785-9>

- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*. <https://doi.org/10.1214/08-AOAS169>
- Jamshidian, M., & Mata, M. (2007). Advances in Analysis of Mean and Covariance Structure when Data are Incomplete. In *Handbook of Latent Variable and Related Models*. <https://doi.org/10.1016/B978-044452044-9/50005-7>
- Järvelin, M R, Elliott, P., Kleinschmidt, I., Martuzzi, M., Grundy, C., Hartikainen, A. L., & Rantakallio, P. (1997). Ecological and individual predictors of birthweight in a northern Finland birth cohort 1986. *Paediatric and Perinatal Epidemiology*, 11(3), 298–312.
- Järvelin, Marjo Riitta, Sovio, U., King, V., Lauren, L., Xu, B., McCarthy, M. I., ... Elliott, P. (2004). Early life factors and blood pressure at age 31 years in the 1966 Northern Finland birth cohort. *Hypertension*. <https://doi.org/10.1161/01.HYP.0000148304.33869.ee>
- Johnson, W., Li, L., Kuh, D., & Hardy, R. (2015). How Has the Age-Related Process of Overweight or Obesity Development Changed over Time? Co-ordinated Analyses of Individual Participant Data from Five United Kingdom Birth Cohorts. *PLOS Medicine*, 12(5), e1001828. <https://doi.org/10.1371/journal.pmed.1001828>
- Kaakinen, M., Mägi, R., Fischer, K., Heikkinen, J., Järvelin, M.-R., Morris, A. P., & Prokopenko, I. (2017). MARV: a tool for genome-wide multi-phenotype analysis of rare variants. *BMC Bioinformatics*, 18(1), 110. <https://doi.org/10.1186/s12859-017-1530-2>
- Kalupahana, N. S., Claycombe, K. J., & Moustaid-Moussa, N. (2011). (n-3) Fatty acids alleviate adipose tissue inflammation and insulin resistance: mechanistic insights. *Advances in Nutrition (Bethesda, Md.)*, 2(4), 304–316. <https://doi.org/10.3945/an.111.000505>
- Kanai, M., Tanaka, T., & Okada, Y. (2016). Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *Journal of Human Genetics*. <https://doi.org/10.1038/jhg.2016.72>
- Karaderi, T., Drong, A. W., & Lindgren, C. M. (2015). Insights into the Genetic Susceptibility to Type 2 Diabetes from Genome-Wide Association Studies of Obesity-Related Traits. *Current Diabetes Reports*. <https://doi.org/10.1007/s11892-015-0648-8>
- Kasner, E., Hunter, C. A., Ph, D., Kariko, K., & Ph, D. (2013). NIH Public Access, 70(4), 646–656. <https://doi.org/10.1002/ana.22528>. Toll-like
- Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burt, N. P., Rieder, M. J., ... Orho-Melander, M. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics*, 40(2), 189–197. <https://doi.org/10.1038/ng.75>
- Kettunen, J., Demirkan, A., Würtz, P., Draisma, H. H. M. M., Haller, T., Rawal, R., ... Ala-Korpela, M. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications*, 7,

11122. <https://doi.org/10.1038/ncomms11122>

- Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.-P., ... Ripatti, S. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics*, 44(3), 269–276. <https://doi.org/10.1038/ng.1073>
- Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., ... Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(September). <https://doi.org/10.1038/s41588-018-0183-z>
- Kingsbury, K. J., & Bondy, G. (2003). Understanding the essentials of blood lipid metabolism. *Progress in Cardiovascular Nursing*. <https://doi.org/10.1111/j.0889-7204.2003.01176.x>
- Knuth, D. E. (1985). Algorithmic Thinking and Mathematical Thinking. *The American Mathematical Monthly*. <https://doi.org/10.2307/2322871>
- Kohl, S. M., Klein, M. S., Hochrein, J., Oefner, P. J., Spang, R., & Gronwald, W. (2012). State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, 8, 146–160. <https://doi.org/10.1007/s11306-011-0350-z>
- Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., & Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study. *BMC Bioinformatics*. <https://doi.org/10.1186/s12859-019-3110-0>
- Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., ... Li, Y. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.1008500>
- Kramer, C. K., Zinman, B., & Retnakaran, R. (2013). Are metabolically healthy overweight and obesity benign conditions?: A systematic review and meta-analysis. *Annals of Internal Medicine*. <https://doi.org/10.7326/0003-4819-159-11-201312030-00008>
- Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D., & Ziegler, A. (2014). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*. <https://doi.org/10.1002/bimj.201300068>
- Läärä, E., & Rantakallio, P. (1996). Body size and mortality in women: A 29 year follow up of 12,000 pregnant women in northern Finland. *Journal of Epidemiology and Community Health*. <https://doi.org/10.1136/jech.50.4.408>
- LaFramboise, T. (2009). Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkp552>

- Lahti-Koski, M., & Gill, T. (2004). Defining Childhood Obesity. In *Obesity in Childhood and Adolescence*. <https://doi.org/10.1159/000078278>
- Laitinen, J., Pietiläinen, K., Wadsworth, M., Sovio, U., & Järvelin, M. R. (2004). Predictors of abdominal obesity among 31-y-old men and women born in Northern Finland in 1966. *European Journal of Clinical Nutrition*. <https://doi.org/10.1038/sj.ejcn.1601765>
- Lambert, S. A., Abraham, G., & Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*. <https://doi.org/10.1093/hmg/ddz187>
- Lamy, P., Grove, J., & Wiuf, C. (2011). A review of software for microarray genotyping. *Human Genomics*. <https://doi.org/10.1186/1479-7364-5-4-304>
- Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, 95, 5–23. <https://doi.org/10.1016/j.ajhg.2014.06.009>
- Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., ... Wray, N. R. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*. <https://doi.org/10.1038/ng.2711>
- Lee, T. H., Ullah, A., & Wang, R. (2020). Bootstrap Aggregating and Random Forest. In *Advanced Studies in Theoretical and Applied Econometrics*. [https://doi.org/10.1007/978-3-030-31150-6\\_13](https://doi.org/10.1007/978-3-030-31150-6_13)
- Lemaitre, R. N., Tanaka, T., Tang, W., Manichaikul, A., Foy, M., Kabagambe, E. K., ... Steffen, L. M. (2011). Genetic loci associated with plasma phospholipid N-3 fatty acids: A Meta-Analysis of Genome-Wide association studies from the charge consortium. *PLoS Genetics*, 7(7). <https://doi.org/10.1371/journal.pgen.1002193>
- Li, Y. R., & Keating, B. J. (2014). Trans-ethnic genome-wide association studies: Advantages and challenges of mapping in diverse populations. *Genome Medicine*, 6(10), 1–14. <https://doi.org/10.1186/s13073-014-0091-5>
- Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*. <https://doi.org/10.1177/154405910408300516>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. *Wiley Series in Probability and Statistics*.
- Little, R. J., & Rubin, D. B. (2015). Missing Data. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. <https://doi.org/10.1016/B978-0-08-097086-8.42082-9>
- Liu, C., Raghavan, S., Maruthur, N., Kabagambe, E. K., Hong, J., Ng, M. C. Y., ... Li, X. (2016). Trans-ethnic Meta-analysis and Functional Annotation Illuminates the Genetic Architecture of Fasting Glucose and Insulin, 56–75. <https://doi.org/10.1016/j.ajhg.2016.05.006>
- Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to

- detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*. <https://doi.org/10.1111/mec.13100>
- Maciellag, M. J. (2012). Chemical properties of antimicrobials and their uniqueness. In *Antibiotic Discovery and Development*. [https://doi.org/10.1007/978-1-4614-1400-1\\_24](https://doi.org/10.1007/978-1-4614-1400-1_24)
- Mägi, R., Suleimanov, Y. V., Clarke, G. M., Kaakinen, M., Fischer, K., Prokopenko, I., & Morris, A. P. (2017). SCOPA and META-SCOPA: software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes. *BMC Bioinformatics*, 18(1), 25. <https://doi.org/10.1186/s12859-016-1437-3>
- Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., ... Morris, A. D. P. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3), 234–244. <https://doi.org/10.1038/ng.2897>
- Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., ... McCarthy, M. I. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*. <https://doi.org/10.1038/s41588-018-0241-6>
- Mahendran, Y., Jonsson, A., Have, C. T., Allin, K. H., Witte, D. R., Jørgensen, M. E., ... Hansen, T. (2017). Genetic evidence of a causal effect of insulin resistance on branched-chain amino acid levels. *Diabetologia*. <https://doi.org/10.1007/s00125-017-4222-6>
- Malerba, G., Schaeffer, L., Xumerle, L., Klopp, N., Trabetti, E., Biscuola, M., ... Illig, T. (2008). SNPs of the FADS gene cluster are associated with polyunsaturated fatty acids in a cohort of patients with cardiovascular disease. *Lipids*. <https://doi.org/10.1007/s11745-008-3158-5>
- Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews. Genetics*, 11(7), 499–511. <https://doi.org/10.1038/nrg2796>
- Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. New York: Academic Press.
- Marian, A. J. (2009). Nature's genetic gradients and the clinical phenotype. *Circulation: Cardiovascular Genetics*. <https://doi.org/10.1161/CIRCGENETICS.109.921940>
- Marian, A. J. (2014). Causality in genetics : The gradient of genetic effects and back to Koch's postulates of causality. *Circulation Research*. <https://doi.org/10.1161/CIRCRESAHA.114.302904>
- Marigorta, U. M., & Navarro, A. (2013). High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. *PLoS Genetics*, 9(6). <https://doi.org/10.1371/journal.pgen.1003566>
- Marullo, L., El-Sayed Moustafa, J. S., & Prokopenko, I. (2014, November 25).

- Insights into the Genetic Susceptibility to Type 2 Diabetes from Genome-Wide Association Studies of Glycaemic Traits. *Current Diabetes Reports*.  
<https://doi.org/10.1007/s11892-014-0551-8>
- Molarius, A., Seidell, J. C., Sans, S., Tuomilehto, J., & Kuulasmaa, K. (1999). Waist and hip circumferences, and waist-hip ratio in 19 populations of the WHO MONICA project. *International Journal of Obesity*.  
<https://doi.org/10.1038/sj.ijo.0800772>
- Morris, A. P. (2011). Transethnic meta-analysis of genomewide association studies. *Genetic Epidemiology*. <https://doi.org/10.1002/gepi.20630>
- Mozaffarian, D., & Wu, J. H. Y. (2011). Omega-3 fatty acids and cardiovascular disease: Effects on risk factors, molecular pathways, and clinical events. *Journal of the American College of Cardiology*, 58(20), 2047–2067.  
<https://doi.org/10.1016/j.jacc.2011.06.063>
- Nature Education. (2014). Nature Education. Retrieved July 12, 2019, from  
<https://www.nature.com/scitable/definition/phenotype-phenotypes-35>
- Newgard, C. B. (2012). Interplay between lipids and branched-chain amino acids in development of insulin resistance. *Cell Metabolism*.  
<https://doi.org/10.1016/j.cmet.2012.01.024>
- Ng, M. C. Y., Shriner, D., Chen, B. H., Li, J., Chen, W. M., Guo, X., ... Surdulescu, G. (2014). Meta-Analysis of Genome-Wide Association Studies in African Americans Provides Insights into the Genetic Architecture of Type 2 Diabetes. *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.1004517>
- Ning, Z., Lee, Y., Joshi, P. K., Wilson, J. F., Pawitan, Y., & Shen, X. (2017). A Selection Operator for Summary Association Statistics Reveals Allelic Heterogeneity of Complex Traits. *American Journal of Human Genetics*.  
<https://doi.org/10.1016/j.ajhg.2017.09.027>
- Nolte, I. M., Van Der Most, P. J., Alizadeh, B. Z., De Bakker, P. I. W., Boezen, H. M., Bruinenberg, M., ... Snieder, H. (2017). Missing heritability: Is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study. *European Journal of Human Genetics*. <https://doi.org/10.1038/ejhg.2017.50>
- O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M. R., & Coin, L. J. M. (2012). MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE*, 7(5).  
<https://doi.org/10.1371/journal.pone.0034861>
- Ogedengbe, O. S., & Ezeani, I. U. (2014). Metabolic syndrome: Performance of five different diagnostic criterias. *Indian Journal of Endocrinology and Metabolism*.  
<https://doi.org/10.4103/2230-8210.137494>
- Oresic, M. (2012). Obesity and psychotic disorders: uncovering common mechanisms through metabolomics. *Disease Models & Mechanisms*, 5, 614–620. <https://doi.org/10.1242/dmm.009845>
- Package, T., & Lazar, A. C. (2015). Package ‘imputeLCMD’.

- Panagiotara, A., Skoufas, E., Chalikiopoulou, C., Tsermpini, E.-E., Bartsakoulia, M., Katsila, T., & Patrinos, G. P. (2019). Implementation of Genomic Medicine. In *Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics* (pp. 369–380). Elsevier. <https://doi.org/10.1016/B978-0-12-812536-6.00015-8>
- Pandey, S., Singh, B., & Verma, R. (2013). Current understanding of Parkinson's disease. *International Journal of Pharmacotherapy*.
- Park, S. H., Lee, J. Y., & Kim, S. (2011). A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes. *BMC Systems Biology*. <https://doi.org/10.1186/1752-0509-5-S2-S13>
- Piccolo, S. R., Abo, R. P., Allen-Brady, K., Camp, N. J., Knight, S., Anderson, J. L., & Horne, B. D. (2009). Evaluation of genetic risk scores for lipid levels using genome-wide markers in the Framingham Heart Study. *BMC Proceedings*. <https://doi.org/10.1186/1753-6561-3-s7-s46>
- Poudyal, H., Panchal, S. K., Diwan, V., & Brown, L. (2011). Omega-3 fatty acids and metabolic syndrome: Effects and emerging mechanisms of action. *Progress in Lipid Research*, 50(4), 372–387. <https://doi.org/10.1016/j.plipres.2011.06.003>
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. a, & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904–909. <https://doi.org/10.1038/ng1847>
- Prokopenko, I., Langenberg, C., Florez, J. C., Saxena, R., Soranzo, N., Thorleifsson, G., ... Abecasis, G. R. (2009). Variants in MTNR1B influence fasting glucose levels. *Nature Genetics*. <https://doi.org/10.1038/ng.290>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. <https://doi.org/10.1086/519795>
- Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble Machine Learning: Methods and Applications*. [https://doi.org/10.1007/9781441993267\\_10](https://doi.org/10.1007/9781441993267_10)
- Ramosaj, B., & Pauly, M. (2019). Predicting missing values: a comparative study on non-parametric approaches for imputation. *Computational Statistics*. <https://doi.org/10.1007/s00180-019-00900-3>
- Rantakallio P. (1969). Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr Scand.*, 193(Suppl 193), 1+.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3), 581–592.
- Rubin, D. B. (1988). An Overview of Multiple Imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 79–84.
- Russo, G. L. (2009). Dietary n - 6 and n - 3 polyunsaturated fatty acids: From biochemistry to clinical implications in cardiovascular prevention. *Biochemical Pharmacology*. <https://doi.org/10.1016/j.bcp.2008.10.020>

- Russo, R., Capasso, M., Paolucci, P., & Iolascon, A. (2011). Pediatric pharmacogenetic and pharmacogenomic studies: The current state and future perspectives. *European Journal of Clinical Pharmacology*, 67(SUPPL. 1). <https://doi.org/10.1007/s00228-010-0931-1>
- Saari, K. M., Lindeman, S. M., Viilo, K. M., Isohanni, M. K., Järvelin, M. R., Laurén, L. H., ... Koponen, H. J. (2005). A 4-fold risk of metabolic syndrome in patients with schizophrenia: The Northern Finland 1966 birth cohort study. *Journal of Clinical Psychiatry*. <https://doi.org/10.4088/JCP.v66n0503>
- Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., ... Peltonen, L. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, 41(1), 35–46. <https://doi.org/10.1038/ng.271>
- Saklayen, M. G. (2018). The Global Epidemic of the Metabolic Syndrome. *Current Hypertension Reports*. <https://doi.org/10.1007/s11906-018-0812-z>
- Sandler, I., & Sandler, L. (1986). On the origin of mendelian genetics. *Integrative and Comparative Biology*. <https://doi.org/10.1093/icb/26.3.753>
- Sarwar, N., Aspelund, T., Eiriksdottir, G., Gobin, R., Seshasai, S. R. K., Forouhi, N. G., ... Gudnason, V. (2010). Markers of dysglycaemia and risk of coronary heart disease in people without diabetes: Reykjavik prospective study and systematic review. *PLoS Medicine*. <https://doi.org/10.1371/journal.pmed.1000278>
- Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1), 3–15. <https://doi.org/10.1191/096228099671525676>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*. <https://doi.org/10.1038/s41576-018-0016-z>
- Schifano, E. D., Li, L., Christiani, D. C., & Lin, X. (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. *American Journal of Human Genetics*, 92(5), 744–759. <https://doi.org/10.1016/j.ajhg.2013.04.004>
- Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics and Development*. <https://doi.org/10.1016/j.gde.2009.04.010>
- Schwarz, D. F., König, I. R., & Ziegler, A. (2010). On safari to random Jungle: A fast implementation of random forests for high-dimensional data. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btq257>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Schwertner, A., & Mosser, L. (1993). Comparison of Lipid Fatty Acids on a Concentration Basis in Patients with and without Coronary Artery Basis vs

- Weight Percentage Disease or Diabetes, 39(4), 659–663.
- Scott, R. A., Lagou, V., Welch, R. P., Wheeler, E., Montasser, M. E., Luan, J., ... Barroso, I. (2012). Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature Genetics*. <https://doi.org/10.1038/ng.2385>
- Scutari, M., Howell, P., Balding, D. J., & Mackay, I. (2014). Multiple Quantitative Trait Analysis Using Bayesian Networks. *Genetics*. <https://doi.org/10.1534/genetics.114.165704>
- Servin, B., & Stephens, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genetics*, 3(7), 1296–1308. <https://doi.org/10.1371/journal.pgen.0030114>
- Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, 179(6), 764–774. <https://doi.org/10.1093/aje/kwt312>
- Shah, J. S., Rai, S. N., DeFilippis, A. P., Hill, B. G., Bhatnagar, A., & Brock, G. N. (2017). Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC Bioinformatics*, 18(1), 114. <https://doi.org/10.1186/s12859-017-1547-6>
- Shin, S.-Y., Petersen, A.-K., Wahl, S., Zhai, G., Römisch-Margl, W., Small, K. S., ... Soranzo, N. (2014). Interrogating causal pathways linking genetic variants, small molecule metabolites, and circulating lipids. *Genome Medicine*, 6(3), 25. <https://doi.org/10.1186/gm542>
- Shriner, D. (2012). Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Frontiers in Genetics*, 3(JAN), 1–7. <https://doi.org/10.3389/fgene.2012.00001>
- Simopoulos, A. P. (2002). The importance of the ratio of omega-6/omega-3 essential fatty acids. *Biomedicine and Pharmacotherapy*. [https://doi.org/10.1016/S0753-3322\(02\)00253-6](https://doi.org/10.1016/S0753-3322(02)00253-6)
- Simopoulos, A. P. (2016). An increase in the Omega-6/Omega-3 fatty acid ratio increases the risk for obesity. *Nutrients*. <https://doi.org/10.3390/nu8030128>
- So, H. C., & Sham, P. C. (2017). Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. *Scientific Reports*. <https://doi.org/10.1038/srep41262>
- Soininen, P., Kangas, A. J., Würtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circulation: Cardiovascular Genetics*. <https://doi.org/10.1161/CIRCGENETICS.114.000216>
- Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews*.

- Genetics*, 14(7), 483–495. <https://doi.org/10.1038/nrg3461>
- Spanakis, E. K., & Golden, S. H. (2013). Race/ethnic difference in diabetes and diabetic complications. *Current Diabetes Reports*. <https://doi.org/10.1007/s11892-013-0421-9>
- Stasny, E. A. (2015). Nonsampling Errors. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. <https://doi.org/10.1016/B978-0-08-097086-8.42064-7>
- Stefan, N., Häring, H. U., Hu, F. B., & Schulze, M. B. (2013). Metabolically healthy obesity: Epidemiology, mechanisms, and clinical implications. *The Lancet Diabetes and Endocrinology*. [https://doi.org/10.1016/S2213-8587\(13\)70062-7](https://doi.org/10.1016/S2213-8587(13)70062-7)
- Stekhoven, D. J., & Bühlmann, P. (2012a). Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118. <https://doi.org/10.1093/bioinformatics/btr597>
- Stekhoven, D. J., & Bühlmann, P. (2012b). Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btr597>
- Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*. <https://doi.org/10.1534/genetics.110.120907>
- Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics*, 27(16), 2304–2305. <https://doi.org/10.1093/bioinformatics/btr341>
- Succop, P. A., Clark, S., Chen, M., & Galke, W. (2004). Imputation of data values that are less than a detection limit. *Journal of Occupational and Environmental Hygiene*. <https://doi.org/10.1080/15459620490462797>
- Suhre, K., & Gieger, C. (2012). Genetic variation in metabolic phenotypes: Study designs and applications. *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg3314>
- Suhre, K., Shin, S.-Y., Petersen, A.-K., Mohny, R. P., Meredith, D., Wägele, B., ... Gieger, C. (2012). Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362), 54–60. <https://doi.org/10.1038/nature10354>
- Tachmazidou, I., Morris, A., & Zeggini, E. (2013). Rare variant association testing for next-generation sequencing data via hierarchical clustering. *Human Heredity*, 74, 165–171. <https://doi.org/10.1159/000346022>
- Tada, H., Won, H. H., Melander, O., Yang, J., Peloso, G. M., & Kathiresan, S. (2014). Multiple associated variants increase the heritability explained for plasma lipids and coronary artery disease. *Circulation: Cardiovascular Genetics*, 7(5), 583–587. <https://doi.org/10.1161/CIRCGENETICS.113.000420>
- Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*.

<https://doi.org/10.1038/s41576-019-0127-1>

- Tan, C. H., Bonham, L. W., Fan, C. C., Mormino, E. C., Sugrue, L. P., Broce, I. J., ... Desikan, R. S. (2019). Polygenic hazard score, amyloid deposition and Alzheimer's neurodegeneration. *Brain*. <https://doi.org/10.1093/brain/awy327>
- Tanaka, T., Shen, J., Abecasis, G. R., Kisialiou, A., Ordovas, J. M., Guralnik, J. M., ... Ferrucci, L. (2009). Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI study. *PLoS Genetics*, 5(1), 1–8. <https://doi.org/10.1371/journal.pgen.1000338>
- Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining*. <https://doi.org/10.1002/sam.11348>
- Teslovich, T. Musunuru, K. Smith, a. E. Al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307), 707–713. <https://doi.org/10.1038/nature09270>.Biological
- Tobias, D. K., Lawler, P. R., Harada, P. H., Demler, O. V., Ridker, P. M., Manson, J. A. E., ... Mora, S. (2018). Circulating Branched-Chain Amino Acids and Incident Cardiovascular Disease in a Prospective Cohort of US Women. *Circulation. Genomic and Precision Medicine*. <https://doi.org/10.1161/CIRCGEN.118.002157>
- Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., ... Benjamin, D. J. (2017). MTAG: Multi-Trait Analysis of GWAS. *BioRxiv*. <https://doi.org/10.1101/118810>
- Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., ... Pitts, S. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*. <https://doi.org/10.1038/s41588-017-0009-4>
- Udler, M. S. (2019). Type 2 Diabetes: Multiple Genes, Multiple Diseases. *Current Diabetes Reports*. <https://doi.org/10.1007/s11892-019-1169-7>
- van Buuren, S. (2012). *Flexible Imputation of Missing Data*. *Flexible Imputation of Missing Data*. <https://doi.org/10.1201/b11826>
- van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition*. *Flexible Imputation of Missing Data, Second Edition*. <https://doi.org/10.1201/9780429492259>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Multivariate Imputation by Chained Equations. *Journal Of Statistical Software*, 45(3), 1–67. <https://doi.org/10.1177/0962280206074463>
- van Heyningen, V., & Yeyati, P. L. (2004). Mechanisms of non-Mendelian inheritance in genetic disease. *Human Molecular Genetics*. <https://doi.org/10.1093/hmg/ddh254>
- Vega, F. M. D. La, & Bustamante, C. D. (2018). Polygenic risk scores : a biased

prediction ?, 95–97.

- Vergara, C., Parker, M. M., Franco, L., Cho, M. H., Valencia-Duarte, A. V., Beaty, T. H., & Duggal, P. (2018). Genotype imputation performance of three reference panels using African ancestry individuals. *Human Genetics*. <https://doi.org/10.1007/s00439-018-1881-4>
- Vilhjálmsdóttir, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., ... Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics*. <https://doi.org/10.1016/j.ajhg.2015.09.001>
- Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., ... Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8), e002847. <https://doi.org/10.1136/bmjopen-2013-002847>
- Wang, Z., Sha, Q., & Zhang, S. (2016). Joint Analysis of Multiple Traits Using “Optimal” Maximum Heritability Test. *PloS One*, 11(3), e0150975. <https://doi.org/10.1371/journal.pone.0150975>
- Warrington, N. M., Howe, L. D., Paternoster, L., Kaakinen, M., Herrala, S., Huikari, V., ... Palmer, L. J. (2015). A genome-wide association study of body mass index across early life and childhood. *International Journal of Epidemiology*. <https://doi.org/10.1093/ije/dyv077>
- Wei, R., Wang, J., Su, M., Jia, E., Chen, S., & Chen, T. (2018). Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*, 1–10. <https://doi.org/10.1038/s41598-017-19120-0>
- Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., & Ni, Y. (2018). Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*. <https://doi.org/10.1038/s41598-017-19120-0>
- Wheeler, E. (2017). Genetic aetiology of glycaemic traits : approaches and insights, 26(July). <https://doi.org/10.1093/hmg/ddx293>
- Wheeler, E., Leong, A., Liu, C. T., Hivert, M. F., Strawbridge, R. J., Podmore, C., ... Meigs, J. B. (2017). Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Medicine*. <https://doi.org/10.1371/journal.pmed.1002383>
- Wheeler, E., Marenne, G., & Barroso, I. (2017). Genetic aetiology of glycaemic traits: Approaches and insights. *Human Molecular Genetics*. <https://doi.org/10.1093/hmg/ddx293>
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377–399. <https://doi.org/10.1002/sim.4067>
- Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*.

<https://doi.org/10.1093/bioinformatics/btq340>

- Wolters, M., Dering, C., Siani, A., Russo, P., Kaprio, J., Risé, P., ... Börnhorst, C. (2017). The role of a FADS1 polymorphism in the association of fatty acid blood levels, BMI and blood pressure in young children—Analyses based on path models. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0181485>
- Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*. <https://doi.org/10.18637/jss.v077.i01>
- WTCCC. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678. <https://doi.org/10.1038/nature05911>
- Würtz, P., Wang, Q., Kangas, A. J., Richmond, R. C., Skarp, J., Tiainen, M., ... Ala-Korpela, M. (2014). Metabolic Signatures of Adiposity in Young Adults: Mendelian Randomization Analysis and Effects of Weight Change. *PLoS Medicine*. <https://doi.org/10.1371/journal.pmed.1001765>
- Würtz, P., Wang, Q., Soininen, P., Kangas, A. J., Fatemifar, G., Tynkkynen, T., ... Ala-Korpela, M. (2016). Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase. *Journal of the American College of Cardiology*. <https://doi.org/10.1016/j.jacc.2015.12.060>
- Yamakado, M., Nagao, K., Imaizumi, A., Tani, M., Toda, A., Tanaka, T., ... Ishizaka, Y. (2015). Plasma Free Amino Acid Profiles Predict Four-Year Risk of Developing Diabetes, Metabolic Syndrome, Dyslipidemia and Hypertension in Japanese Population. *Scientific Reports*, 5(1), 11918. <https://doi.org/10.1038/srep11918>
- Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*. <https://doi.org/10.1016/j.ajhg.2010.11.011>
- Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., ... Visscher, P. M. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. *Human Molecular Genetics*. <https://doi.org/10.1093/hmg/ddy271>
- Yuan, Y. (2010). Multiple imputation for missing data: Concepts and new development. *SAS Institute Inc, Rockville, MD*, 1–13. Retrieved from <http://www.math.montana.edu/~jimrc/classes/stat506/notes/multipleimputation-SAS.pdf>
- Zhao, B., & Zou, F. (2018). Is Polygenic Risk Scores Prediction Good ?
- Zhao, Y., Chen, F., Zhai, R., Lin, X., Wang, Z., Su, L., & Christiani, D. C. (2012). Correction for population stratification in random forest analysis. *International Journal of Epidemiology*. <https://doi.org/10.1093/ije/dys183>
- Zimmet, P., Alberti, K. G. M. M., & Shaw, J. (2001). Global and societal implications of the diabetes epidemic. *Nature*. <https://doi.org/10.1038/414782a>

## 11. Supplementary

*Supplementary Table 1 The variables of NFBC cohorts.*

Cohort column specifies whether the variable was present in NFBC1966 (66 only), NFBC1986 (86 only), or both. Imp=imputed (yes or no), Pred=predictor (yes or no), Log trans=log transformed (yes or no), Quant norm=quantile normalised (yes or no)

| Variable name | Description                                            | Cohort  | Imp | Pred   | Log trans | Quant norm |
|---------------|--------------------------------------------------------|---------|-----|--------|-----------|------------|
| ID            | Participant identifier                                 | both    | ×   | ×      | ×         | ×          |
| sex           | Gender                                                 | both    | ×   | ✓      | ×         | ×          |
| t1d           | Type 1 diabetes diagnosis                              | 66 only | ×   | ✓      | ×         | ×          |
| t2d           | Type 2 diabetes diagnosis                              | 66 only | ×   | ✓      | ×         | ×          |
| diab          | Diabetes diagnosis                                     | 86 only | ×   | ✓      | ×         | ×          |
| height        | Height                                                 | both    | ✓   | ✓      | ×         | ×          |
| weight        | Weight                                                 | both    | ✓   | ✓      | ×         | ×          |
| hip           | Hip circumference                                      | both    | ✓   | ✓      | ×         | ×          |
| waist         | Waist circumference                                    | both    | ✓   | ✓      | ×         | ×          |
| pills31       | Contraceptive pill use                                 | both    | ×   | ♀ only | ×         | ×          |
| bmi31         | Body mass index                                        | both    | ✓   | ✓      | ×         | ×          |
| pc1           | Principle component 1                                  | 66 only | ×   | ✓      | ×         | ×          |
| pc2           | Principle component 2                                  | 66 only | ×   | ✓      | ×         | ×          |
| pc3           | Principle component 3                                  | 66 only | ×   | ✓      | ×         | ×          |
| MDS1          | Multidimensional Scaling component 1                   | 86 only | ×   | ✓      | ×         | ×          |
| MDS2          | Multidimensional Scaling component 2                   | 86 only | ×   | ✓      | ×         | ×          |
| MDS3          | Multidimensional Scaling component 3                   | 86 only | ×   | ✓      | ×         | ×          |
| FS_KOL        | Fasting cholesterol level                              | both    | ✓   | ✓      | ×         | ×          |
| FS_KOL_L      | Fasting LDL cholesterol level                          | both    | ✓   | ✓      | ×         | ×          |
| FS_KOL_H      | Fasting HDL cholesterol level                          | both    | ✓   | ✓      | ×         | ×          |
| FS_TRIGL      | Fasting triglycerides level                            | both    | ✓   | ✓      | ×         | ×          |
| FB_GLUK       | Fasting glucose level                                  | both    | ✓   | ✓      | ×         | ×          |
| FS_INS        | Fasting insulin level                                  | both    | ✓   | ✓      | ×         | ×          |
| crp3dec       | C-reactive protein level                               | 66 only | ✓   | ✓      | ×         | ×          |
| HOMA_IR       | Homeostasis model assessment of insulin resistance     | 66 only | ✓   | ✓      | ×         | ×          |
| homab         | Homeostasis model assessment of $\beta$ -cell function | 66 only | ✓   | ✓      | ×         | ×          |
| LMED          | Lipid-lowering medication use                          | 66 only | ×   | ✓      | ×         | ×          |
| EXCLUDE       | Non-fasting status                                     | both    | ×   | ✓      | ×         | ×          |
| a10atc        | Diabetes medication use                                | 66 only | ×   | ✓      | ×         | ×          |
| DIABMED       | Diabetes medication use                                | 86 only | ×   | ✓      | ×         | ×          |
| pregnant      | Pregnancy status                                       | both    | ×   | ♀ only | ×         | ×          |
| systa         | Systolic blood pressure                                | both    | ✓   | ✓      | ×         | ×          |
| diasta        | Diastolic blood pressure                               | both    | ✓   | ✓      | ×         | ×          |
| BPMED         | Blood pressure medication use                          | both    | ×   | ✓      | ×         | ×          |
| XXL_VLDL_P    | Concentration of extra-extra-large VLDL particles      | both    | ✓   | ✓      | ✓         | ✓          |
| XXL_VLDL_L    | Total lipids in extra-extra-large VLDL                 | both    | ✓   | ✓      | ×         | ✓          |
| XXL_VLDL_PL   | Phospholipids in extra-extra-large VLDL                | both    | ✓   | ✓      | ×         | ✓          |

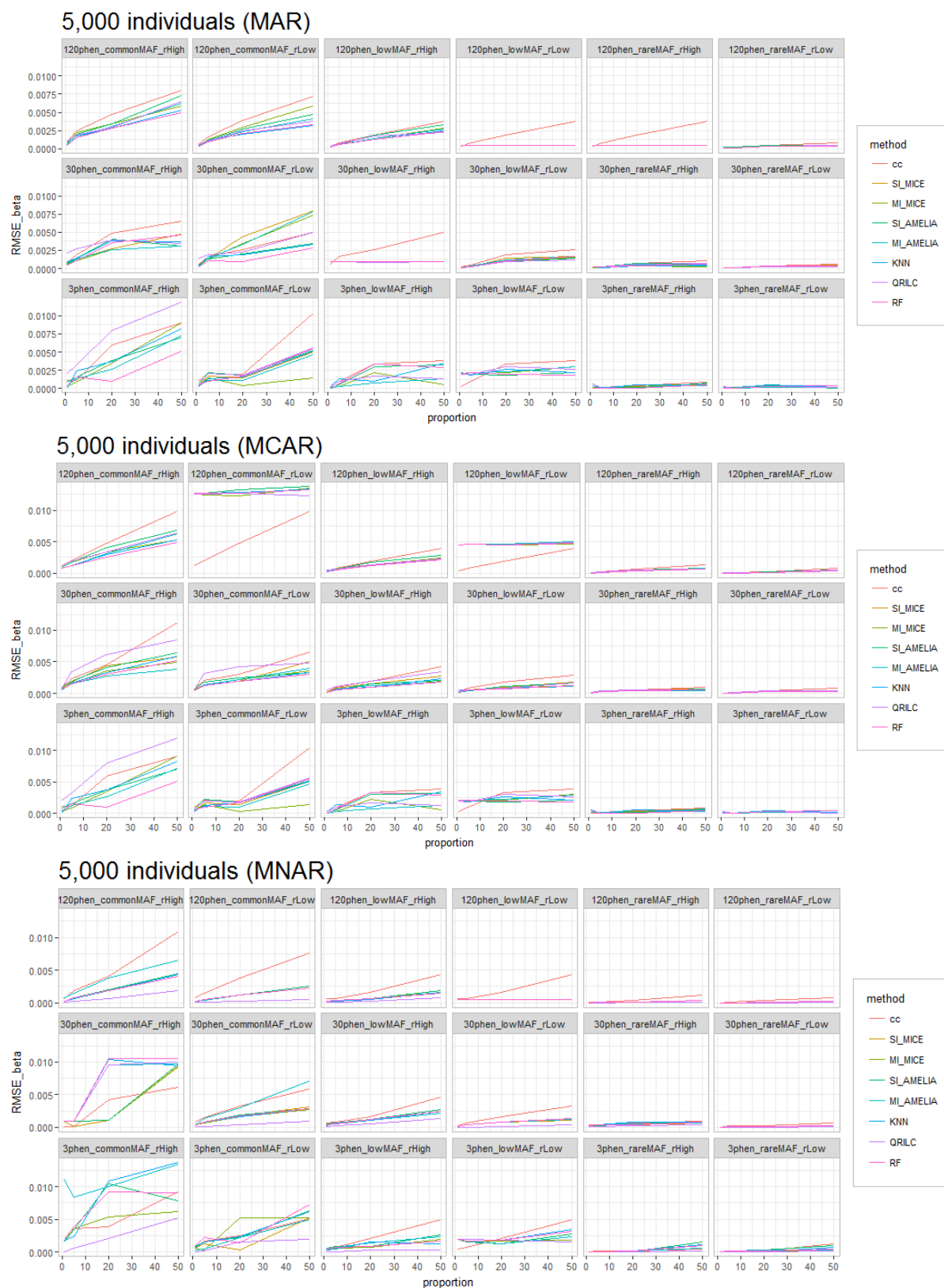
|                    |                                             |      |   |   |   |   |
|--------------------|---------------------------------------------|------|---|---|---|---|
| <b>XXL_VLDL_TG</b> | Triglycerides in extra-extra-large VLDL     | both | ✓ | ✓ | ✗ | ✓ |
| <b>XL_VLDL_P</b>   | Concentration of extra-large VLDL particles | both | ✓ | ✓ | ✓ | ✓ |
| <b>XL_VLDL_L</b>   | Total lipids in extra-large VLDL            | both | ✓ | ✓ | ✗ | ✓ |
| <b>XL_VLDL_PL</b>  | Phospholipids in extra-large VLDL           | both | ✓ | ✓ | ✗ | ✓ |
| <b>XL_VLDL_TG</b>  | Triglycerides in extra-large VLDL           | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_VLDL_P</b>    | Concentration of large VLDL particles       | both | ✓ | ✓ | ✓ | ✓ |
| <b>L_VLDL_L</b>    | Total lipids in large VLDL                  | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_VLDL_PL</b>   | Phospholipids in large VLDL                 | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_VLDL_C</b>    | Total cholesterol in large VLDL             | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_VLDL_CE</b>   | Cholesteresters in large VLDL               | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_VLDL_FC</b>   | Free cholesterol in large VLDL              | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_VLDL_TG</b>   | Triglycerides in large VLDL                 | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_VLDL_P</b>    | Concentration of medium VLDL particles      | both | ✓ | ✓ | ✓ | ✓ |
| <b>M_VLDL_L</b>    | Total lipids in medium VLDL                 | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_VLDL_PL</b>   | Phospholipids in medium VLDL                | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_VLDL_C</b>    | Total cholesterol in medium VLDL            | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_VLDL_CE</b>   | Cholesterol esters in medium VLDL           | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_VLDL_FC</b>   | Free cholesterol in medium VLDL             | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_VLDL_TG</b>   | Triglycerides in medium VLDL                | both | ✓ | ✓ | ✗ | ✓ |
| <b>S_VLDL_P</b>    | Concentration of small VLDL particles       | both | ✓ | ✓ | ✓ | ✓ |
| <b>S_VLDL_L</b>    | Total lipids in small VLDL                  | both | ✓ | ✓ | ✗ | ✓ |
| <b>S_VLDL_PL</b>   | Phospholipids in small VLDL                 | both | ✓ | ✓ | ✗ | ✓ |
| <b>S_VLDL_C</b>    | Total cholesterol in small VLDL             | both | ✓ | ✓ | ✗ | ✓ |
| <b>S_VLDL_FC</b>   | Free cholesterol in small VLDL              | both | ✓ | ✓ | ✗ | ✓ |
| <b>S_VLDL_TG</b>   | Triglycerides in small VLDL                 | both | ✓ | ✓ | ✗ | ✓ |
| <b>XS_VLDL_P</b>   | Concentration of extra small VLDL particles | both | ✓ | ✓ | ✓ | ✓ |
| <b>XS_VLDL_L</b>   | Total lipids in extra small VLDL            | both | ✓ | ✓ | ✗ | ✓ |
| <b>XS_VLDL_PL</b>  | Phospholipids in extra small VLDL           | both | ✓ | ✓ | ✗ | ✓ |
| <b>XS_VLDL_TG</b>  | Triglycerides in extra small VLDL           | both | ✓ | ✓ | ✗ | ✓ |
| <b>IDL_P</b>       | Concentration of IDL particles              | both | ✓ | ✓ | ✓ | ✓ |
| <b>IDL_L</b>       | Total lipids in IDL                         | both | ✓ | ✓ | ✗ | ✓ |
| <b>IDL_PL</b>      | Phospholipids in IDL                        | both | ✓ | ✓ | ✗ | ✓ |
| <b>IDL_C</b>       | Total cholesterol in IDL                    | both | ✓ | ✓ | ✗ | ✓ |
| <b>IDL_FC</b>      | Free cholesterol in IDL                     | both | ✓ | ✓ | ✗ | ✓ |
| <b>IDL_TG</b>      | Triglycerides in IDL                        | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_LDL_P</b>     | Concentration of large LDL particles        | both | ✓ | ✓ | ✓ | ✓ |
| <b>L_LDL_L</b>     | Total lipids in large LDL                   | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_LDL_PL</b>    | Phospholipids in large LDL                  | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_LDL_C</b>     | Total cholesterol in large LDL              | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_LDL_CE</b>    | Cholesterol esters in large LDL             | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_LDL_FC</b>    | Free cholesterol in large LDL               | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_LDL_P</b>     | Concentration of medium LDL particles       | both | ✓ | ✓ | ✓ | ✓ |
| <b>M_LDL_L</b>     | Total lipids in medium LDL                  | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_LDL_PL</b>    | Phospholipids in medium LDL                 | both | ✓ | ✓ | ✗ | ✓ |

|                  |                                         |      |   |   |   |   |
|------------------|-----------------------------------------|------|---|---|---|---|
| <b>M_LDL_C</b>   | Total cholesterol in medium LDL         | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_LDL_CE</b>  | Cholesterol esters in medium LDL        | both | ✓ | ✓ | ✗ | ✓ |
| <b>S_LDL_P</b>   | Concentration of small LDL particles    | both | ✓ | ✓ | ✓ | ✓ |
| <b>S_LDL_L</b>   | Total lipids in small LDL               | both | ✓ | ✓ | ✗ | ✓ |
| <b>S_LDL_C</b>   | Total cholesterol in small LDL          | both | ✓ | ✓ | ✗ | ✓ |
| <b>XL_HDL_P</b>  | Concentration extra-large HDL particles | both | ✓ | ✓ | ✓ | ✓ |
| <b>XL_HDL_L</b>  | Total lipids in extra-large HDL         | both | ✓ | ✓ | ✗ | ✓ |
| <b>XL_HDL_PL</b> | Phospholipids in extra-large HDL        | both | ✓ | ✓ | ✗ | ✓ |
| <b>XL_HDL_C</b>  | Total cholesterol in extra-large HDL    | both | ✓ | ✓ | ✗ | ✓ |
| <b>XL_HDL_CE</b> | Cholesterol esters in extra-large HDL   | both | ✓ | ✓ | ✗ | ✓ |
| <b>XL_HDL_FC</b> | Free cholesterol in extra-large HDL     | both | ✓ | ✓ | ✗ | ✓ |
| <b>XL_HDL_TG</b> | Triglycerides in extra-large HDL        | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_HDL_P</b>   | Concentration of large HDL particles    | both | ✓ | ✓ | ✓ | ✓ |
| <b>L_HDL_L</b>   | Total lipids in large HDL               | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_HDL_PL</b>  | Phospholipids in large HDL              | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_HDL_C</b>   | Total cholesterol in large HDL          | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_HDL_CE</b>  | Cholesterol esters in large HDL         | both | ✓ | ✓ | ✗ | ✓ |
| <b>L_HDL_FC</b>  | Free cholesterol in large HDL           | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_HDL_P</b>   | Concentration medium HDL particles      | both | ✓ | ✓ | ✓ | ✓ |
| <b>M_HDL_L</b>   | Total lipids in medium HDL              | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_HDL_PL</b>  | Phospholipids in medium HDL             | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_HDL_C</b>   | Total cholesterol in medium HDL         | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_HDL_CE</b>  | Cholesterol esters in medium HDL        | both | ✓ | ✓ | ✗ | ✓ |
| <b>M_HDL_FC</b>  | Free cholesterol in medium HDL          | both | ✓ | ✓ | ✗ | ✓ |
| <b>S_HDL_P</b>   | Concentration of small HDL particles    | both | ✓ | ✓ | ✓ | ✓ |
| <b>S_HDL_L</b>   | Total lipids in small HDL               | both | ✓ | ✓ | ✗ | ✓ |
| <b>S_HDL_TG</b>  | Triglycerides in small HDL              | both | ✓ | ✓ | ✗ | ✓ |
| <b>VLDL_TG</b>   | Triglycerides in VLDL                   | both | ✓ | ✓ | ✗ | ✓ |
| <b>LDL_C</b>     | Total cholesterol in LDL                | both | ✓ | ✓ | ✗ | ✓ |
| <b>HDL_C</b>     | Total cholesterol in HDL                | both | ✓ | ✓ | ✗ | ✓ |
| <b>Serum_TG</b>  | Serum total triglycerides               | both | ✓ | ✓ | ✗ | ✓ |
| <b>Serum_C</b>   | Serum total cholesterol                 | both | ✓ | ✓ | ✗ | ✓ |
| <b>EstC</b>      | Total esterified cholesterol            | both | ✓ | ✓ | ✗ | ✓ |
| <b>FreeC</b>     | Total free cholesterol                  | both | ✓ | ✓ | ✗ | ✓ |
| <b>FAw3</b>      | Omega-3 fatty acids                     | both | ✓ | ✓ | ✗ | ✓ |
| <b>FAw6</b>      | Omega-6 fatty acids                     | both | ✓ | ✓ | ✗ | ✓ |
| <b>FAw79S</b>    | Omega-7, -9, and saturated fatty acids  | both | ✓ | ✓ | ✗ | ✓ |
| <b>TotFA</b>     | Total fatty acids                       | both | ✓ | ✓ | ✗ | ✓ |
| <b>LA</b>        | 18:2, linoleic acid                     | both | ✓ | ✓ | ✗ | ✓ |
| <b>otPUFA</b>    | Other polyunsaturated fatty acids       | both | ✓ | ✓ | ✗ | ✓ |
| <b>DHA</b>       | 22:6, docosahexaenoic acid              | both | ✓ | ✓ | ✗ | ✓ |
| <b>TotTG</b>     | Total triglycerides                     | both | ✓ | ✓ | ✗ | ✓ |
| <b>PC</b>        | Phosphatidylcholine and other cholines  | both | ✓ | ✓ | ✗ | ✓ |
| <b>SM</b>        | Sphingomyelins                          | both | ✓ | ✓ | ✗ | ✓ |

|                    |                                                  |      |   |   |   |   |
|--------------------|--------------------------------------------------|------|---|---|---|---|
| <b>TotCH</b>       | Total cholines (and other N-trimethyl compounds) | both | ✓ | ✓ | ✗ | ✓ |
| <b>boHBut</b>      | 3-hydroxybutyrate                                | both | ✓ | ✓ | ✗ | ✓ |
| <b>Ace</b>         | Acetate                                          | both | ✓ | ✓ | ✗ | ✓ |
| <b>AcAce</b>       | Acetoacetate                                     | both | ✓ | ✓ | ✗ | ✓ |
| <b>Ala</b>         | Alanine                                          | both | ✓ | ✓ | ✗ | ✓ |
| <b>Alb</b>         | Albumin                                          | both | ✓ | ✓ | ✗ | ✓ |
| <b>MobCH</b>       | Double bond protons of mobile lipids             | both | ✓ | ✓ | ✗ | ✓ |
| <b>MobCH2</b>      | CH2(methylene) groups of mobile lipids           | both | ✓ | ✓ | ✗ | ✓ |
| <b>MobCH3</b>      | CH3(methyl) groups of mobile lipids              | both | ✓ | ✓ | ✗ | ✓ |
| <b>Cit</b>         | Citrate                                          | both | ✓ | ✓ | ✗ | ✓ |
| <b>Crea</b>        | creatinine                                       | both | ✓ | ✓ | ✗ | ✓ |
| <b>Glc</b>         | Glucose                                          | both | ✓ | ✓ | ✗ | ✓ |
| <b>Gln</b>         | Glutamine                                        | both | ✓ | ✓ | ✗ | ✓ |
| <b>GloI</b>        | Glycerol                                         | both | ✓ | ✓ | ✗ | ✓ |
| <b>Gly</b>         | glycine                                          | both | ✓ | ✓ | ✗ | ✓ |
| <b>Gp</b>          | Glycoprotein acetyls                             | both | ✓ | ✓ | ✗ | ✓ |
| <b>His</b>         | Histidine                                        | both | ✓ | ✓ | ✗ | ✓ |
| <b>Ile</b>         | Isoleucine                                       | both | ✓ | ✓ | ✗ | ✓ |
| <b>Lac</b>         | lactate                                          | both | ✓ | ✓ | ✗ | ✓ |
| <b>Leu</b>         | Leucine                                          | both | ✓ | ✓ | ✗ | ✓ |
| <b>Phe</b>         | Phenylalanine                                    | both | ✓ | ✓ | ✗ | ✓ |
| <b>Pyr</b>         | pyruvate                                         | both | ✓ | ✓ | ✗ | ✓ |
| <b>Tyr</b>         | Tyrosine                                         | both | ✓ | ✓ | ✗ | ✓ |
| <b>Urea</b>        | Urea                                             | both | ✓ | ✓ | ✗ | ✓ |
| <b>Val</b>         | Valine                                           | both | ✓ | ✓ | ✗ | ✓ |
| <b>VLDL_D</b>      | Mean diameter for VLDL particles                 | both | ✓ | ✓ | ✗ | ✓ |
| <b>LDL_D</b>       | Mean diameter for LDL particles                  | both | ✓ | ✓ | ✗ | ✓ |
| <b>HDL_D</b>       | Mean diameter for HDL particles                  | both | ✓ | ✓ | ✗ | ✓ |
| <b>VLDL_TG_eFR</b> | Triglycerides in VLDL                            | both | ✓ | ✓ | ✗ | ✓ |
| <b>IDL_C_eFR</b>   | Total cholesterol in IDL                         | both | ✓ | ✓ | ✗ | ✓ |
| <b>LDL_C_eFR</b>   | Total cholesterol in LDL                         | both | ✓ | ✓ | ✗ | ✓ |
| <b>HDL2_C</b>      | Total cholesterol in HDL2                        | both | ✓ | ✓ | ✗ | ✓ |
| <b>ApoA1</b>       | Apolipoprotein A-I                               | both | ✓ | ✓ | ✗ | ✓ |
| <b>ApoB</b>        | Apolipoprotein B                                 | both | ✓ | ✓ | ✗ | ✓ |

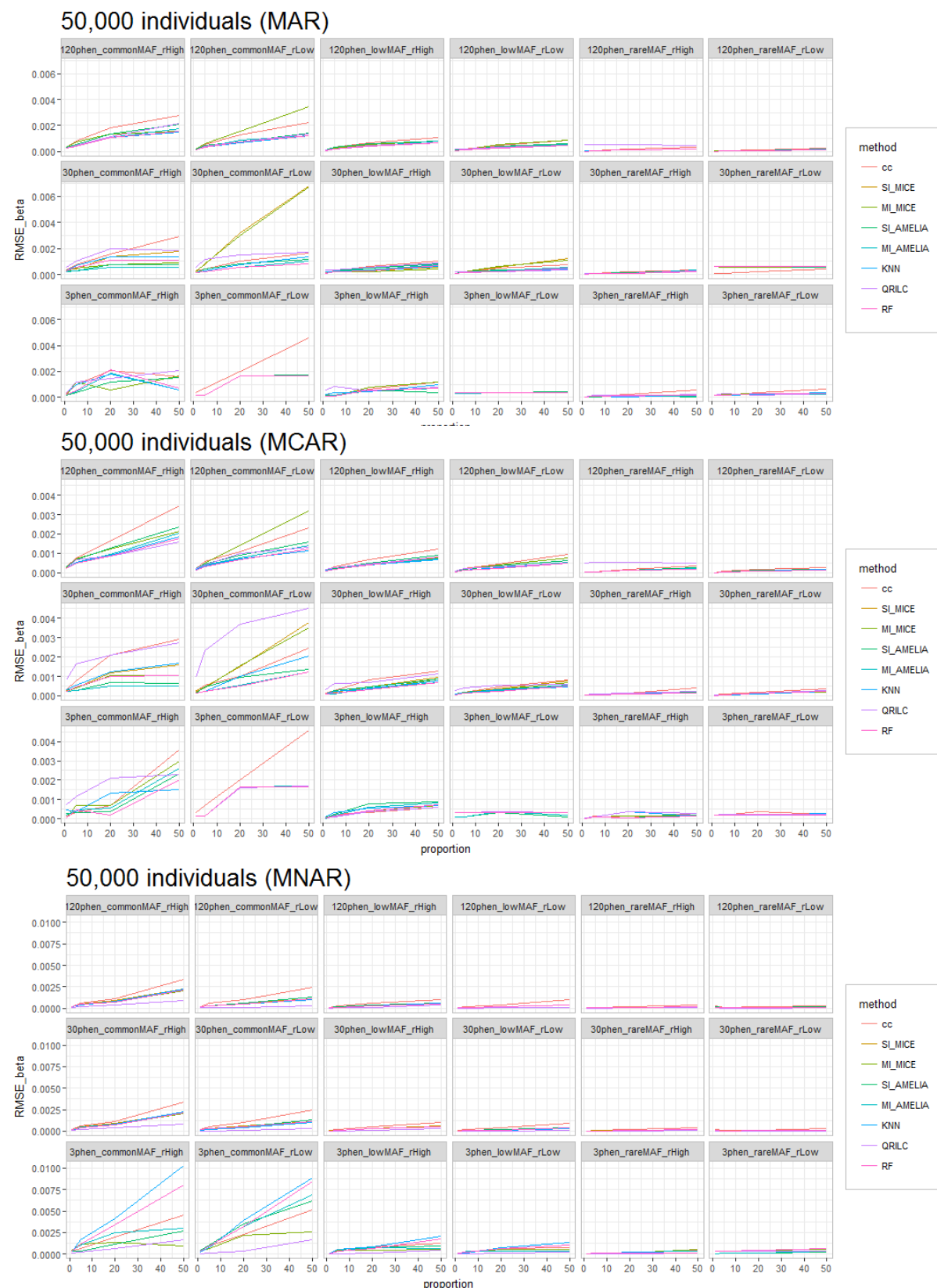
Supplementary Table 2 The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis of 5,000 individuals in MP-GWAS analysis.

The top panel illustrates RMSEs of 120 phenotypes, the middle panel for 30 and the lower panel for 3 phenotypes. The two pairs of figures are for high and low phenotype correlations, with the first two on the left for common variants, next two for low-frequency variants and the last two for rare variants.



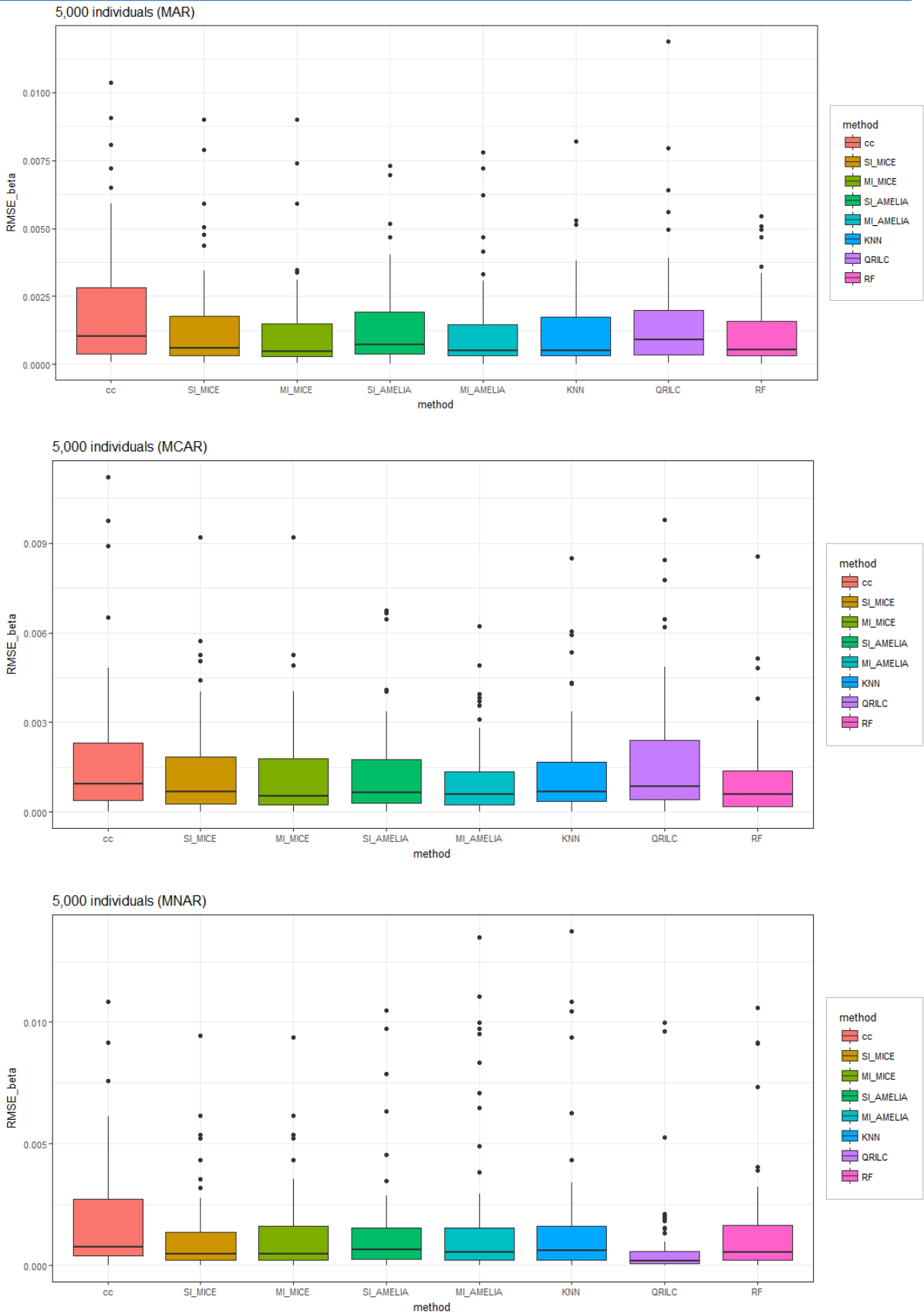
**Supplementary Table 3** The Difference between the root mean square errors (RMSEs) obtained from the regression coefficient of seven imputation methods and those from full data analysis of 50,000 individuals in MP-GWAS analysis.

The top panel illustrates RMSEs of 120 phenotypes, the middle panel for 30 and the lower panel for 3 phenotypes. The two pairs of figures are for high and low phenotype correlations, with the first two on the left for common variants, next two for low-frequency variants and the last two for rare variants.



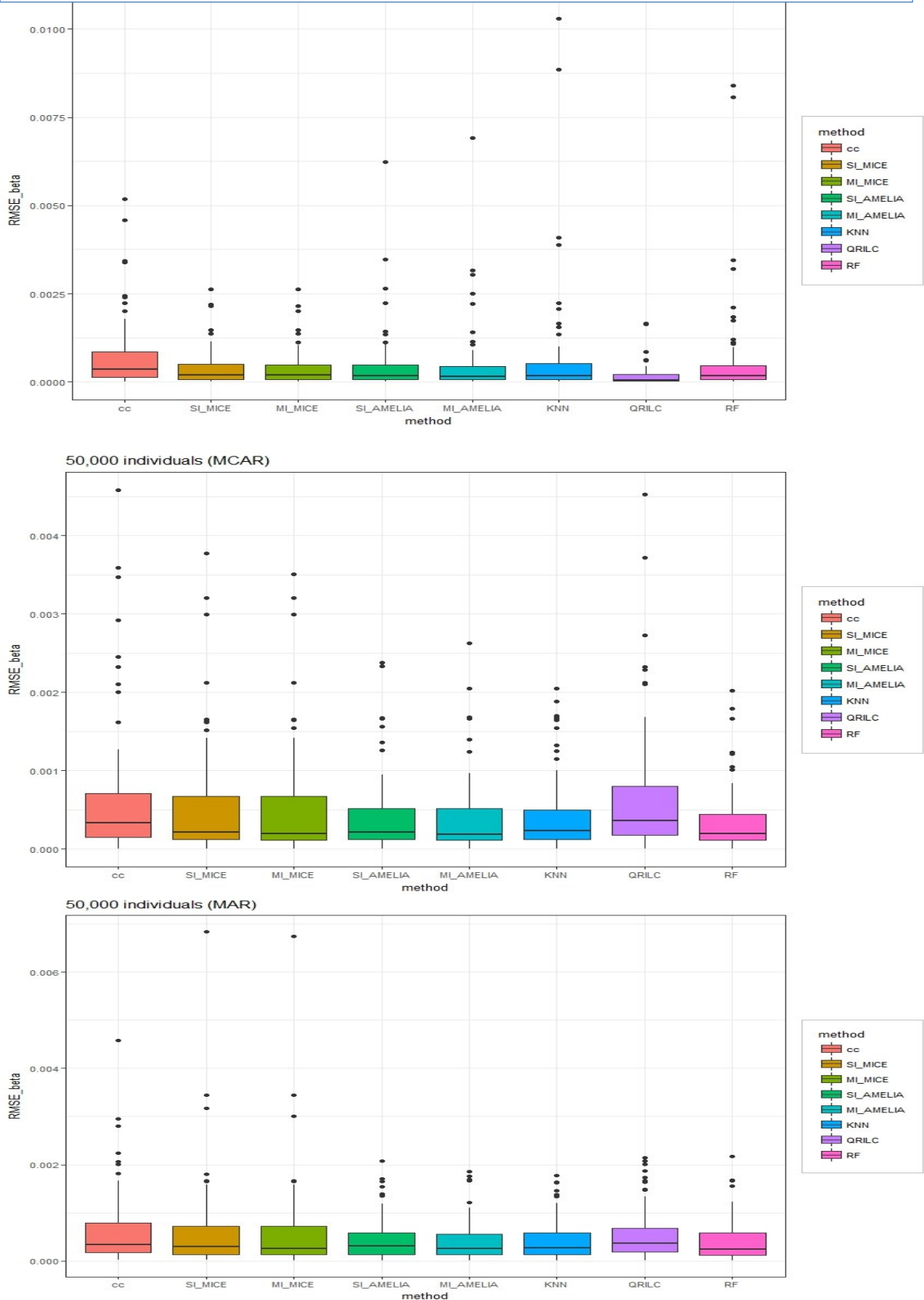
Supplementary Table 4 Root mean squared error (RMSE) boxplot obtained from the regression coefficient of seven imputation methods and those from full data analysis of 5,000 individuals in MP-GWAS analysis.

The boxes range from the 25th to 75th quantiles. The black line within the box are the median RMSE for each method.



Supplementary Table 5 Root mean squared error (RMSE) boxplot obtained from the regression coefficient of seven imputation methods and those from full data analysis of 50,000 individuals in MP-GWAS analysis.

The boxes range from the 25th to 75th quantiles. The black line within the box are the median RMSE for each method.



*Supplementary Table 6 The SNP list contains 394 SNPs that represent signals across any of the phenotypes analysed in the meta-analysis, and a few of them did not pass imputation quality and did not present either in NFBC1966 or NFBC1986.*

| Chr | Bp       | SNP | rsid        | Present in NFBC1966 | Present in NFBC1986 |
|-----|----------|-----|-------------|---------------------|---------------------|
| 1   | 3107287  | SNP | rs12142172  | Yes                 | Yes                 |
| 1   | 43436051 | SNP | rs841572    | Yes                 | Yes                 |
| 1   | 1.01E+08 | SNP | rs6662924   | Yes                 | Yes                 |
| 1   | 1.51E+08 | SNP | rs267738    | Yes                 | Yes                 |
| 1   | 2.14E+08 | SNP | rs340874    | Yes                 | Yes                 |
| 1   | 2.3E+08  | SNP | rs348330    | Yes                 | Yes                 |
| 1   | 2.2E+08  | SNP | rs6674544   | Yes                 | Yes                 |
| 1   | 3691727  | SNP | rs1175549   | Yes                 | Yes                 |
| 1   | 25703156 | SNP | rs644592    | Yes                 | Yes                 |
| 1   | 1.55E+08 | SNP | rs10796941  | Yes                 | Yes                 |
| 1   | 1.56E+08 | SNP | rs12127403  | Yes                 | No                  |
| 1   | 1.59E+08 | SNP | rs857725    | Yes                 | Yes                 |
| 1   | 2.04E+08 | SNP | rs10900585  | Yes                 | Yes                 |
| 1   | 2.14E+08 | SNP | rs340882    | Yes                 | Yes                 |
| 1   | 25529038 | SNP | rs2375278   | Yes                 | Yes                 |
| 1   | 1.55E+08 | SNP | rs7534795   | Yes                 | Yes                 |
| 1   | 2.04E+08 | SNP | rs7547793   | Yes                 | Yes                 |
| 1   | 1.51E+08 | SNP | rs78132593  | Yes                 | Yes                 |
| 1   | 2.14E+08 | SNP | rs2075423   | Yes                 | Yes                 |
| 2   | 27730940 | SNP | rs1260326   | Yes                 | Yes                 |
| 2   | 1.66E+08 | SNP | rs12692738  | Yes                 | Yes                 |
| 2   | 27152874 | SNP | rs1371614   | Yes                 | Yes                 |
| 2   | 28268742 | SNP | rs115640879 | Yes                 | Yes                 |
| 2   | 40444081 | SNP | rs141107738 | Yes                 | Yes                 |
| 2   | 43775309 | SNP | rs183381538 | Yes                 | Yes                 |
| 2   | 45192080 | SNP | rs12712928  | Yes                 | Yes                 |
| 2   | 54941112 | SNP | rs189548    | Yes                 | Yes                 |
| 2   | 1.69E+08 | SNP | rs78996387  | Yes                 | Yes                 |
| 2   | 1.7E+08  | SNP | rs150171632 | Yes                 | Yes                 |
| 2   | 1.7E+08  | SNP | rs560887    | Yes                 | Yes                 |
| 2   | 1.7E+08  | SNP | rs145353824 | Yes                 | Yes                 |
| 2   | 1.7E+08  | SNP | rs56100844  | Yes                 | Yes                 |
| 2   | 1.7E+08  | SNP | rs3755158   | Yes                 | Yes                 |
| 2   | 1.74E+08 | SNP | rs6731931   | Yes                 | Yes                 |
| 2   | 2.22E+08 | SNP | rs7572235   | Yes                 | Yes                 |
| 2   | 630902   | SNP | rs77935490  | Yes                 | Yes                 |
| 2   | 27177428 | SNP | rs61007968  | Yes                 | Yes                 |
| 2   | 65296280 | SNP | rs2252867   | Yes                 | Yes                 |
| 2   | 1.66E+08 | SNP | rs1128249   | Yes                 | Yes                 |
| 2   | 2.27E+08 | SNP | rs2972145   | Yes                 | Yes                 |
| 2   | 2.28E+08 | SNP | rs17508368  | Yes                 | Yes                 |
| 2   | 24093756 | SNP | rs12612492  | Yes                 | Yes                 |

|   |          |       |             |     |     |
|---|----------|-------|-------------|-----|-----|
| 2 | 24235704 | SNP   | rs6545222   | Yes | Yes |
| 2 | 26281459 | SNP   | rs1560868   | Yes | Yes |
| 2 | 43449385 | SNP   | rs1367173   | Yes | Yes |
| 2 | 48123915 | SNP   | rs6723441   | Yes | Yes |
| 2 | 68939499 | SNP   | rs4854444   | Yes | Yes |
| 2 | 1.75E+08 | SNP   | rs17256082  | Yes | Yes |
| 2 | 2.19E+08 | SNP   | rs13427681  | Yes | Yes |
| 2 | 2.27E+08 | SNP   | rs2138157   | Yes | Yes |
| 2 | 48587198 | SNP   | rs17037289  | Yes | Yes |
| 2 | 2.19E+08 | SNP   | rs4674280   | Yes | Yes |
| 2 | 27140022 | SNP   | rs877273    | Yes | Yes |
| 2 | 1.7E+08  | SNP   | rs72883292  | Yes | Yes |
| 2 | 1.7E+08  | SNP   | rs573225    | Yes | Yes |
| 2 | 1.66E+08 | SNP   | rs13389219  | Yes | Yes |
| 2 | 2.27E+08 | SNP   | rs2943646   | Yes | Yes |
| 3 | 1.23E+08 | SNP   | rs11708067  | Yes | Yes |
| 3 | 1.86E+08 | SNP   | rs7630554   | Yes | Yes |
| 3 | 49642430 | SNP   | rs13064576  | Yes | Yes |
| 3 | 1E+08    | SNP   | rs4132537   | Yes | Yes |
| 3 | 1.41E+08 | SNP   | rs75098673  | Yes | Yes |
| 3 | 1.52E+08 | INDEL | rs148134482 | Yes | Yes |
| 3 | 1.71E+08 | SNP   | rs1604038   | Yes | Yes |
| 3 | 1.86E+08 | SNP   | rs7631557   | Yes | Yes |
| 3 | 1.88E+08 | SNP   | rs6808574   | Yes | Yes |
| 3 | 1.92E+08 | SNP   | rs189651013 | Yes | Yes |
| 3 | 12116620 | SNP   | rs308971    | Yes | Yes |
| 3 | 12344730 | SNP   | rs11712037  | Yes | Yes |
| 3 | 49884261 | SNP   | rs9819511   | Yes | Yes |
| 3 | 52844534 | SNP   | rs17331151  | Yes | Yes |
| 3 | 1.5E+08  | SNP   | rs62271373  | Yes | Yes |
| 3 | 12266804 | SNP   | rs1822534   | Yes | Yes |
| 3 | 49382925 | SNP   | rs9818758   | Yes | Yes |
| 3 | 52887861 | INDEL | rs66499923  | Yes | Yes |
| 3 | 1.23E+08 | SNP   | rs11719201  | Yes | Yes |
| 3 | 1.72E+08 | SNP   | rs4894769   | Yes | Yes |
| 3 | 1.72E+08 | SNP   | rs7632281   | Yes | Yes |
| 3 | 1.86E+08 | INDEL | rs35188816  | Yes | Yes |
| 3 | 1.88E+08 | SNP   | rs6777684   | Yes | Yes |
| 3 | 1.96E+08 | SNP   | rs9846651   | Yes | Yes |
| 3 | 12268244 | SNP   | rs12491937  | Yes | Yes |
| 3 | 1.72E+08 | SNP   | rs13089972  | Yes | Yes |
| 3 | 1.41E+08 | SNP   | rs16851397  | Yes | Yes |
| 3 | 1.52E+08 | SNP   | rs17437560  | Yes | Yes |
| 3 | 1.86E+08 | INDEL | rs144346724 | Yes | Yes |
| 3 | 12351521 | SNP   | rs35000407  | Yes | Yes |
| 4 | 1297720  | SNP   | rs190355720 | Yes | Yes |

|   |          |       |             |     |     |
|---|----------|-------|-------------|-----|-----|
| 4 | 1.86E+08 | SNP   | rs4862423   | Yes | Yes |
| 4 | 89726283 | SNP   | rs2276936   | Yes | Yes |
| 4 | 1.06E+08 | SNP   | rs10010325  | Yes | Yes |
| 4 | 1.45E+08 | SNP   | rs7654571   | Yes | Yes |
| 4 | 1.46E+08 | SNP   | rs11727676  | Yes | Yes |
| 4 | 1.58E+08 | SNP   | rs6855363   | Yes | Yes |
| 4 | 1.11E+08 | SNP   | rs144392753 | Yes | Yes |
| 4 | 1.45E+08 | SNP   | rs13129993  | Yes | Yes |
| 4 | 1.45E+08 | SNP   | rs77909720  | Yes | Yes |
| 4 | 1.45E+08 | SNP   | rs13134327  | Yes | Yes |
| 4 | 1.45E+08 | SNP   | rs112578089 | Yes | Yes |
| 4 | 89739808 | SNP   | rs3775380   | Yes | Yes |
| 4 | 1.06E+08 | SNP   | rs9884482   | Yes | Yes |
| 5 | 55809127 | SNP   | rs157512    | Yes | Yes |
| 5 | 76427311 | SNP   | rs6878122   | Yes | Yes |
| 5 | 95703329 | SNP   | rs6876986   | Yes | Yes |
| 5 | 1.02E+08 | SNP   | rs7729395   | Yes | Yes |
| 5 | 1.68E+08 | SNP   | rs3733977   | Yes | Yes |
| 5 | 53272664 | SNP   | rs4865796   | Yes | Yes |
| 5 | 55806751 | SNP   | rs459193    | Yes | Yes |
| 5 | 55860781 | SNP   | rs3936511   | Yes | Yes |
| 5 | 1.58E+08 | SNP   | rs1023667   | Yes | Yes |
| 5 | 76435004 | SNP   | rs7732130   | Yes | Yes |
| 5 | 1.54E+08 | SNP   | rs13166977  | Yes | Yes |
| 5 | 1.56E+08 | SNP   | rs1948759   | Yes | Yes |
| 5 | 1.54E+08 | SNP   | rs6877043   | Yes | Yes |
| 5 | 76425867 | SNP   | rs7708285   | Yes | Yes |
| 5 | 95696585 | SNP   | rs1820176   | Yes | Yes |
| 6 | 20676414 | INDEL | rs34499031  | Yes | Yes |
| 6 | 7231843  | SNP   | rs9379084   | Yes | Yes |
| 6 | 7250270  | SNP   | rs3778321   | Yes | Yes |
| 6 | 20680678 | SNP   | rs9348441   | Yes | Yes |
| 6 | 39034095 | SNP   | rs10305457  | Yes | Yes |
| 6 | 39046794 | SNP   | rs10305492  | Yes | Yes |
| 6 | 43816377 | SNP   | rs2025704   | Yes | Yes |
| 6 | 1.17E+08 | SNP   | rs1761880   | Yes | Yes |
| 6 | 1.53E+08 | SNP   | rs12055786  | Yes | Yes |
| 6 | 31484550 | SNP   | rs116684538 | Yes | Yes |
| 6 | 34222201 | SNP   | rs116141873 | Yes | Yes |
| 6 | 34239989 | SNP   | rs1187129   | Yes | Yes |
| 6 | 41550466 | INDEL | rs5875762   | Yes | Yes |
| 6 | 43757896 | SNP   | rs998584    | Yes | Yes |
| 6 | 43809802 | SNP   | rs9472135   | Yes | Yes |
| 6 | 1.27E+08 | SNP   | rs2745353   | Yes | Yes |
| 6 | 1.4E+08  | SNP   | rs632057    | Yes | Yes |
| 6 | 1.64E+08 | SNP   | rs4709746   | Yes | Yes |

|   |          |     |             |                     |     |
|---|----------|-----|-------------|---------------------|-----|
| 6 | 24978511 | SNP | rs12193223  | Yes                 | Yes |
| 6 | 25578433 | SNP | rs75580845  | Yes                 | Yes |
| 6 | 26091179 | SNP | rs1799945   | Yes                 | Yes |
| 6 | 27941387 | SNP | rs13214703  | Yes                 | Yes |
| 6 | 31555392 | SNP | rs2256974   | Yes                 | Yes |
| 6 | 32381939 | SNP | rs115541994 | Yes                 | Yes |
| 6 | 1.1E+08  | SNP | rs13195517  | Yes                 | Yes |
| 6 | 1.35E+08 | SNP | rs9376090   | Yes                 | Yes |
| 6 | 20703952 | SNP | rs6931514   | Yes                 | Yes |
| 6 | 26093141 | SNP | rs1800562   | Yes                 | Yes |
| 6 | 31522669 | SNP | rs6929796   | Yes                 | Yes |
| 6 | 34236973 | SNP | rs2780215   | Yes                 | Yes |
| 6 | 1.27E+08 | SNP | rs1474696   | Yes                 | Yes |
| 6 | 1.64E+08 | SNP | rs73013411  | Yes                 | Yes |
| 6 | 28984755 | SNP | rs116735744 | Yes                 | Yes |
| 7 | 44231778 | SNP | rs2971669   | Yes                 | Yes |
| 7 | 14919852 | SNP | rs10281892  | Yes                 | Yes |
| 7 | 15064190 | SNP | rs10228796  | Yes                 | Yes |
| 7 | 28189411 | SNP | rs1635852   | Yes                 | Yes |
| 7 | 44145178 | SNP | rs55841377  | Yes                 | Yes |
| 7 | 44229068 | SNP | rs1799884   | Yes                 | Yes |
| 7 | 44235694 | SNP | rs138917529 | Yes                 | Yes |
| 7 | 44787312 | SNP | rs57003656  | <b>Did not pass</b> | Yes |
| 7 | 50786663 | SNP | rs2108349   | Yes                 | Yes |
| 7 | 75836023 | SNP | rs67070387  | Yes                 | Yes |
| 7 | 89854446 | SNP | rs194520    | Yes                 | Yes |
| 7 | 50754679 | SNP | rs2282930   | Yes                 | Yes |
| 7 | 77537964 | SNP | rs848494    | Yes                 | Yes |
| 7 | 1.3E+08  | SNP | rs13234269  | Yes                 | Yes |
| 7 | 15064309 | SNP | rs2191349   | Yes                 | Yes |
| 7 | 44234737 | SNP | rs2908286   | Yes                 | Yes |
| 7 | 73012042 | SNP | rs35332062  | Yes                 | Yes |
| 7 | 98931105 | SNP | rs2301889   | Yes                 | Yes |
| 7 | 1.23E+08 | SNP | rs77300440  | Yes                 | Yes |
| 7 | 15060429 | SNP | rs10231021  | Yes                 | Yes |
| 7 | 44226101 | SNP | rs2971670   | Yes                 | Yes |
| 7 | 73025975 | SNP | rs13234131  | Yes                 | Yes |
| 7 | 1.23E+08 | SNP | rs4731113   | Yes                 | Yes |
| 7 | 15063430 | SNP | rs10487796  | Yes                 | Yes |
| 7 | 44255643 | SNP | rs878521    | Yes                 | Yes |
| 7 | 75654574 | SNP | rs58925536  | Yes                 | Yes |
| 7 | 89800053 | SNP | rs13242882  | Yes                 | Yes |
| 7 | 1.3E+08  | SNP | rs972283    | Yes                 | Yes |
| 8 | 9183596  | SNP | rs4841132   | Yes                 | Yes |
| 8 | 9173209  | SNP | rs7012637   | Yes                 | Yes |
| 8 | 81048558 | SNP | rs13274117  | Yes                 | Yes |

|    |          |       |             |     |     |
|----|----------|-------|-------------|-----|-----|
| 8  | 95960511 | SNP   | rs896854    | Yes | Yes |
| 8  | 1.18E+08 | SNP   | rs13266634  | Yes | Yes |
| 8  | 1.46E+08 | SNP   | rs3757970   | Yes | Yes |
| 8  | 11423072 | SNP   | rs12541800  | Yes | Yes |
| 8  | 23615445 | SNP   | rs13258890  | Yes | Yes |
| 8  | 41516581 | SNP   | rs4317621   | Yes | Yes |
| 8  | 41543675 | SNP   | rs34664882  | Yes | No  |
| 8  | 41630405 | SNP   | rs4737009   | Yes | Yes |
| 8  | 42383084 | SNP   | rs6980507   | Yes | Yes |
| 8  | 1.18E+08 | SNP   | rs11558471  | Yes | Yes |
| 8  | 1.26E+08 | SNP   | rs2954021   | Yes | Yes |
| 8  | 9183358  | SNP   | rs9987289   | Yes | Yes |
| 8  | 81076874 | SNP   | rs12541643  | Yes | Yes |
| 8  | 1.18E+08 | SNP   | rs9650069   | Yes | Yes |
| 8  | 9202317  | INDEL | rs200250298 | Yes | No  |
| 8  | 9021933  | SNP   | rs330945    | Yes | Yes |
| 8  | 9173358  | SNP   | rs7012814   | Yes | Yes |
| 9  | 1.36E+08 | SNP   | rs550057    | Yes | Yes |
| 9  | 682570   | SNP   | rs60405463  | Yes | Yes |
| 9  | 4283137  | SNP   | rs1574285   | Yes | Yes |
| 9  | 22134068 | SNP   | rs10811660  | Yes | Yes |
| 9  | 33344252 | SNP   | rs642157    | Yes | Yes |
| 9  | 1.12E+08 | SNP   | rs16913693  | Yes | Yes |
| 9  | 1.36E+08 | SNP   | rs507666    | Yes | Yes |
| 9  | 1.39E+08 | SNP   | rs3829109   | Yes | Yes |
| 9  | 4295880  | SNP   | rs4339696   | Yes | Yes |
| 9  | 82294364 | SNP   | rs11138325  | Yes | Yes |
| 9  | 1.36E+08 | SNP   | rs8176693   | Yes | Yes |
| 9  | 22134253 | SNP   | rs10811662  | Yes | Yes |
| 9  | 34081331 | SNP   | rs62560863  | Yes | Yes |
| 9  | 79977312 | SNP   | rs7847351   | Yes | Yes |
| 9  | 91495135 | SNP   | rs61750929  | Yes | Yes |
| 9  | 1.11E+08 | SNP   | rs1467311   | Yes | Yes |
| 9  | 1.36E+08 | SNP   | rs649129    | Yes | Yes |
| 9  | 22134094 | SNP   | rs10811661  | Yes | Yes |
| 9  | 80015424 | SNP   | rs12351997  | Yes | Yes |
| 9  | 1.11E+08 | SNP   | rs7042939   | Yes | Yes |
| 9  | 4291928  | SNP   | rs10974438  | Yes | Yes |
| 9  | 1.36E+08 | SNP   | rs75179845  | Yes | Yes |
| 10 | 70982136 | INDEL | rs35696875  | Yes | Yes |
| 10 | 1.15E+08 | SNP   | rs7903146   | Yes | Yes |
| 10 | 26505822 | SNP   | rs2839671   | Yes | Yes |
| 10 | 64914467 | SNP   | rs1046521   | Yes | Yes |
| 10 | 95384152 | SNP   | rs7095788   | Yes | Yes |
| 10 | 1.13E+08 | SNP   | rs35011531  | Yes | Yes |
| 10 | 1.13E+08 | SNP   | rs11195538  | Yes | Yes |

|    |          |       |             |     |     |
|----|----------|-------|-------------|-----|-----|
| 10 | 1.15E+08 | SNP   | rs34872471  | Yes | Yes |
| 10 | 65184717 | SNP   | rs10761762  | Yes | Yes |
| 10 | 89764490 | SNP   | rs12769346  | Yes | Yes |
| 10 | 1.05E+08 | SNP   | rs11191559  | Yes | Yes |
| 10 | 1.23E+08 | SNP   | rs7071062   | Yes | Yes |
| 10 | 12307894 | SNP   | rs11257655  | Yes | Yes |
| 10 | 71002040 | INDEL | rs5785903   | Yes | Yes |
| 10 | 71094504 | SNP   | rs17476364  | Yes | Yes |
| 10 | 1.13E+08 | SNP   | rs12784552  | Yes | Yes |
| 10 | 89680631 | SNP   | rs118164457 | Yes | Yes |
| 11 | 2858440  | SNP   | rs2237896   | Yes | Yes |
| 11 | 17408404 | SNP   | rs5213      | Yes | Yes |
| 11 | 2182224  | SNP   | rs689       | Yes | Yes |
| 11 | 2856658  | SNP   | rs4930011   | Yes | Yes |
| 11 | 2858546  | SNP   | rs2237897   | Yes | Yes |
| 11 | 7575745  | SNP   | rs10743026  | Yes | Yes |
| 11 | 8255408  | SNP   | rs2168101   | Yes | Yes |
| 11 | 45870177 | SNP   | rs10838524  | Yes | Yes |
| 11 | 47340680 | INDEL | rs10717442  | Yes | No  |
| 11 | 47820241 | SNP   | rs34228231  | Yes | Yes |
| 11 | 49313235 | SNP   | rs10769572  | Yes | Yes |
| 11 | 61609750 | SNP   | rs174583    | Yes | Yes |
| 11 | 72460398 | SNP   | rs77464186  | Yes | Yes |
| 11 | 92708710 | SNP   | rs10830963  | Yes | Yes |
| 11 | 63869062 | SNP   | rs2845885   | Yes | Yes |
| 11 | 72350595 | SNP   | rs3781926   | Yes | Yes |
| 11 | 234451   | SNP   | rs4980325   | Yes | Yes |
| 11 | 2184848  | SNP   | rs3842727   | Yes | Yes |
| 11 | 9763094  | SNP   | rs373894    | Yes | Yes |
| 11 | 10508903 | SNP   | rs117706710 | Yes | Yes |
| 11 | 47278502 | SNP   | rs11039154  | Yes | Yes |
| 11 | 61610750 | SNP   | rs174584    | Yes | Yes |
| 11 | 1E+08    | SNP   | rs11224302  | Yes | Yes |
| 11 | 1.19E+08 | SNP   | rs617948    | Yes | Yes |
| 11 | 2181060  | SNP   | rs3842753   | Yes | Yes |
| 11 | 9776567  | SNP   | rs360140    | Yes | Yes |
| 11 | 47363285 | SNP   | rs10838696  | Yes | Yes |
| 11 | 61581656 | SNP   | rs174559    | Yes | Yes |
| 11 | 47293799 | SNP   | rs10501320  | Yes | Yes |
| 11 | 72460694 | SNP   | rs11603349  | Yes | Yes |
| 11 | 17415190 | SNP   | rs4148646   | Yes | Yes |
| 12 | 1.21E+08 | SNP   | rs2649999   | Yes | Yes |
| 12 | 56865338 | SNP   | rs2657879   | Yes | Yes |
| 12 | 57780936 | SNP   | rs12315434  | Yes | Yes |
| 12 | 97848910 | SNP   | rs6538804   | Yes | Yes |
| 12 | 1.22E+08 | SNP   | rs7962128   | Yes | Yes |

|    |          |       |             |              |     |
|----|----------|-------|-------------|--------------|-----|
| 12 | 1.33E+08 | SNP   | rs11610045  | Yes          | Yes |
| 12 | 20582651 | SNP   | rs12369443  | Yes          | Yes |
| 12 | 21696684 | SNP   | rs2054435   | Yes          | Yes |
| 12 | 26474867 | SNP   | rs10842708  | Yes          | Yes |
| 12 | 48202696 | SNP   | rs111264094 | Yes          | Yes |
| 12 | 66371880 | SNP   | rs7968682   | Yes          | Yes |
| 12 | 79157553 | SNP   | rs73343765  | Did not pass | Yes |
| 12 | 1.03E+08 | SNP   | rs1402013   | Yes          | Yes |
| 12 | 1.03E+08 | SNP   | rs35747     | Yes          | Yes |
| 12 | 1.24E+08 | SNP   | rs7975482   | Yes          | Yes |
| 12 | 4328521  | SNP   | rs117233107 | Yes          | Yes |
| 12 | 7075882  | SNP   | rs2110073   | Yes          | Yes |
| 12 | 48486696 | SNP   | rs76261711  | Yes          | Yes |
| 12 | 48512285 | SNP   | rs4760682   | Yes          | Yes |
| 12 | 57645789 | SNP   | rs11172134  | Yes          | Yes |
| 12 | 1.12E+08 | SNP   | rs17696736  | Yes          | Yes |
| 12 | 1.13E+08 | SNP   | rs4766971   | Yes          | Yes |
| 12 | 1.12E+08 | SNP   | rs10774624  | Yes          | Yes |
| 12 | 1.22E+08 | SNP   | rs6489811   | Yes          | Yes |
| 12 | 21699928 | SNP   | rs6487237   | Yes          | Yes |
| 12 | 66351826 | SNP   | rs1351394   | Yes          | Yes |
| 12 | 1.03E+08 | SNP   | rs860598    | Yes          | Yes |
| 12 | 1.24E+08 | SNP   | rs7133378   | Yes          | Yes |
| 13 | 28487599 | SNP   | rs11619319  | Yes          | Yes |
| 13 | 33554302 | SNP   | rs576674    | Yes          | Yes |
| 13 | 57128454 | SNP   | rs10781829  | Yes          | Yes |
| 13 | 1.11E+08 | SNP   | rs9521730   | Yes          | Yes |
| 13 | 1.13E+08 | SNP   | rs76533333  | Yes          | Yes |
| 13 | 1.14E+08 | SNP   | rs1278769   | Yes          | Yes |
| 13 | 1.15E+08 | SNP   | rs7994900   | Yes          | Yes |
| 14 | 38842759 | SNP   | rs112824462 | Yes          | Yes |
| 14 | 90038821 | SNP   | rs8003022   | Yes          | Yes |
| 14 | 1.01E+08 | SNP   | rs12888855  | Yes          | Yes |
| 14 | 65232989 | SNP   | rs59914095  | Yes          | Yes |
| 14 | 73601025 | SNP   | rs147727276 | Yes          | Yes |
| 14 | 90066451 | SNP   | rs17776811  | Yes          | Yes |
| 14 | 1.01E+08 | SNP   | rs1535464   | Yes          | Yes |
| 14 | 65268605 | SNP   | rs2273475   | Yes          | Yes |
| 14 | 73616095 | SNP   | rs10151436  | Yes          | Yes |
| 14 | 1.01E+08 | INDEL | rs33952550  | Yes          | Yes |
| 14 | 90055468 | SNP   | rs35889227  | Yes          | Yes |
| 15 | 62330633 | SNP   | rs11634163  | Yes          | Yes |
| 15 | 62391608 | SNP   | rs7163757   | Yes          | Yes |
| 15 | 75102923 | SNP   | rs11630478  | Yes          | Yes |
| 15 | 77747276 | SNP   | rs11633054  | Yes          | Yes |
| 15 | 99271135 | SNP   | rs6598541   | Yes          | Yes |

|    |          |       |             |     |     |
|----|----------|-------|-------------|-----|-----|
| 15 | 65849552 | SNP   | rs368777    | Yes | Yes |
| 15 | 77832762 | SNP   | rs7177055   | Yes | Yes |
| 15 | 81366859 | SNP   | rs8034233   | Yes | Yes |
| 15 | 65822777 | SNP   | rs452306    | Yes | Yes |
| 15 | 75102851 | SNP   | rs12594062  | Yes | Yes |
| 15 | 77747190 | SNP   | rs7178572   | Yes | Yes |
| 15 | 62332980 | SNP   | rs17271305  | Yes | Yes |
| 16 | 300641   | SNP   | rs9940149   | Yes | Yes |
| 16 | 4014282  | SNP   | rs2238435   | Yes | Yes |
| 16 | 53494617 | SNP   | rs2024449   | Yes | Yes |
| 16 | 293562   | SNP   | rs11248914  | Yes | Yes |
| 16 | 324479   | SNP   | rs3752435   | Yes | Yes |
| 16 | 11437607 | SNP   | rs9926463   | Yes | Yes |
| 16 | 28531287 | SNP   | rs4787458   | Yes | Yes |
| 16 | 31404571 | SNP   | rs8050500   | Yes | Yes |
| 16 | 53825905 | SNP   | rs56137030  | Yes | Yes |
| 16 | 68832943 | SNP   | rs1862748   | Yes | Yes |
| 16 | 88775220 | SNP   | rs4238685   | Yes | Yes |
| 16 | 88853729 | SNP   | rs837763    | Yes | Yes |
| 16 | 11443183 | SNP   | rs11643024  | Yes | Yes |
| 16 | 28590030 | SNP   | rs7190771   | Yes | Yes |
| 16 | 68818390 | SNP   | rs7198799   | Yes | Yes |
| 17 | 7129898  | SNP   | rs72839768  | Yes | Yes |
| 17 | 17391443 | SNP   | rs8072971   | Yes | Yes |
| 17 | 17344122 | SNP   | rs1880900   | Yes | Yes |
| 17 | 27588980 | INDEL | rs143960646 | Yes | No  |
| 17 | 45479446 | SNP   | rs12452315  | Yes | Yes |
| 17 | 46123932 | SNP   | rs3764400   | Yes | Yes |
| 17 | 7537098  | SNP   | rs858516    | Yes | Yes |
| 17 | 8158273  | SNP   | rs6503094   | Yes | Yes |
| 17 | 17740281 | SNP   | rs13306736  | Yes | Yes |
| 17 | 27183104 | SNP   | rs9914988   | Yes | Yes |
| 17 | 57506092 | SNP   | rs2060779   | Yes | Yes |
| 17 | 76121864 | SNP   | rs2748427   | Yes | Yes |
| 17 | 80689036 | SNP   | rs9909940   | Yes | No  |
| 17 | 80950648 | INDEL | rs141786856 | Yes | No  |
| 17 | 17622666 | SNP   | rs8075153   | Yes | Yes |
| 17 | 7185779  | SNP   | rs117643180 | Yes | Yes |
| 18 | 60845884 | SNP   | rs12454712  | Yes | Yes |
| 18 | 43774444 | SNP   | rs28671200  | Yes | Yes |
| 19 | 46182304 | SNP   | rs10423928  | Yes | Yes |
| 19 | 46196634 | SNP   | rs2302593   | Yes | Yes |
| 19 | 7125519  | SNP   | rs1799815   | Yes | Yes |
| 19 | 7970635  | SNP   | rs4804833   | Yes | Yes |
| 19 | 33894846 | SNP   | rs10422861  | Yes | Yes |
| 19 | 40696513 | SNP   | rs339525    | Yes | Yes |

|    |          |       |             |              |     |
|----|----------|-------|-------------|--------------|-----|
| 19 | 47592967 | INDEL | rs200172871 | Yes          | Yes |
| 19 | 17257802 | SNP   | rs17533945  | Yes          | Yes |
| 19 | 33037212 | SNP   | rs12978547  | Yes          | Yes |
| 19 | 33073431 | SNP   | rs4499344   | Yes          | Yes |
| 19 | 46160323 | SNP   | rs55872740  | Yes          | Yes |
| 19 | 33072085 | SNP   | rs10405535  | Yes          | Yes |
| 19 | 46181392 | SNP   | rs1800437   | Yes          | Yes |
| 19 | 33899065 | SNP   | rs731839    | Yes          | Yes |
| 20 | 22562326 | INDEL | rs3833331   | Yes          | Yes |
| 20 | 39832628 | SNP   | rs17265513  | Yes          | Yes |
| 20 | 6623374  | SNP   | rs979012    | Yes          | Yes |
| 20 | 42337875 | SNP   | rs285171    | Yes          | Yes |
| 20 | 45581777 | SNP   | rs1999536   | Yes          | Yes |
| 20 | 22557099 | SNP   | rs6113722   | Yes          | Yes |
| 20 | 25054083 | SNP   | rs62217197  | Yes          | Yes |
| 20 | 55990405 | SNP   | rs737092    | Yes          | Yes |
| 20 | 45582472 | SNP   | rs1206760   | Yes          | Yes |
| 21 | 24493294 | INDEL | rs200678953 | Yes          | Yes |
| 22 | 42908066 | SNP   | rs184389108 | Yes          | Yes |
| 22 | 30343186 | SNP   | rs39713     | Yes          | Yes |
| 22 | 37462936 | SNP   | rs855791    | Yes          | Yes |
| 22 | 43141907 | SNP   | rs2143923   | Yes          | Yes |
| X  | 66565671 | SNP   | rs2497942   | Yes          | No  |
| X  | 1.1E+08  | SNP   | rs12007422  | Did not pass | No  |
| X  | 1.5E+08  | SNP   | rs184506746 | Did not pass | No  |
| X  | 1.52E+08 | SNP   | rs182859029 | Did not pass | No  |
| X  | 1.53E+08 | SNP   | rs138245006 | Did not pass | No  |
| X  | 1.53E+08 | INDEL | rs199546838 | Did not pass | No  |
| X  | 1.54E+08 | SNP   | rs140801471 | Yes          | No  |
| X  | 1.54E+08 | SNP   | rs149309160 | Yes          | No  |

Supplementary Table 7 The RMSE of regression coefficient between different imputation methods in 5,000 individuals and three phenotypes.

| Correlation | Variant       | Missing mechanisms | proportion (%) | Complete Cases | SI MICE  | MI MICE  | SI AMELIAII | MI AMELIAII | k-NN     | QRILC    | RF       |
|-------------|---------------|--------------------|----------------|----------------|----------|----------|-------------|-------------|----------|----------|----------|
| High        | Common        | MAR                | 1              | 1.08E-03       | 3.22E-04 | 3.22E-04 | 8.03E-04    | 3.47E-04    | 1.64E-05 | 2.04E-03 | 4.37E-04 |
| High        | Common        | MAR                | 5              | 1.15E-03       | 8.88E-04 | 8.88E-04 | 1.54E-03    | 1.28E-03    | 2.40E-03 | 3.19E-03 | 1.57E-03 |
| High        | Common        | MAR                | 20             | 5.92E-03       | 3.45E-03 | 3.45E-03 | 3.79E-03    | 2.68E-03    | 3.70E-03 | 7.94E-03 | 1.04E-03 |
| High        | Common        | MAR                | 50             | 9.08E-03       | 9.01E-03 | 9.01E-03 | 6.97E-03    | 7.21E-03    | 8.20E-03 | 1.19E-02 | 5.09E-03 |
| High        | Low frequency | MAR                | 1              | 2.70E-04       | 4.43E-05 | 4.43E-05 | 9.99E-05    | 6.03E-05    | 2.46E-04 | 2.92E-04 | 8.42E-05 |
| High        | Low frequency | MAR                | 5              | 1.02E-03       | 2.49E-04 | 2.49E-04 | 5.27E-04    | 2.64E-04    | 1.37E-03 | 9.53E-04 | 6.74E-04 |
| High        | Low frequency | MAR                | 20             | 3.34E-03       | 2.19E-03 | 2.19E-03 | 3.00E-03    | 7.28E-04    | 9.67E-04 | 1.69E-03 | 3.36E-03 |
| High        | Low frequency | MAR                | 50             | 3.81E-03       | 5.59E-04 | 5.59E-04 | 3.22E-03    | 1.31E-03    | 3.47E-03 | 1.29E-03 | 2.92E-03 |
| High        | Rare          | MAR                | 1              | 1.25E-04       | 1.76E-04 | 1.76E-04 | 6.31E-05    | 7.23E-06    | 4.29E-04 | 6.13E-04 | 2.93E-05 |
| High        | Rare          | MAR                | 5              | 1.52E-04       | 1.11E-04 | 1.11E-04 | 5.42E-05    | 1.42E-04    | 3.13E-05 | 2.70E-05 | 5.88E-05 |
| High        | Rare          | MAR                | 20             | 3.90E-04       | 2.81E-04 | 2.81E-04 | 1.18E-04    | 6.16E-05    | 5.18E-04 | 5.23E-05 | 2.72E-05 |
| High        | Rare          | MAR                | 50             | 8.33E-04       | 6.72E-04 | 6.72E-04 | 5.48E-04    | 7.81E-04    | 3.63E-04 | 4.87E-04 | 4.92E-04 |
| Low         | Common        | MAR                | 1              | 1.05E-03       | 3.41E-04 | 3.41E-04 | 6.88E-04    | 5.00E-04    | 2.83E-04 | 5.13E-04 | 5.27E-04 |
| Low         | Common        | MAR                | 5              | 2.06E-03       | 1.70E-03 | 1.47E-03 | 2.16E-03    | 1.12E-03    | 1.32E-03 | 1.99E-03 | 9.55E-04 |
| Low         | Common        | MAR                | 20             | 1.91E-03       | 1.47E-03 | 3.41E-04 | 1.79E-03    | 1.04E-03    | 1.61E-03 | 1.77E-03 | 1.70E-03 |
| Low         | Common        | MAR                | 50             | 1.04E-02       | 5.06E-03 | 1.47E-03 | 5.19E-03    | 4.68E-03    | 5.15E-03 | 5.60E-03 | 5.44E-03 |
| Low         | Low frequency | MAR                | 1              | 2.70E-04       | 2.14E-03 | 2.14E-03 | 2.13E-03    | 2.00E-03    | 2.03E-03 | 1.98E-03 | 2.08E-03 |
| Low         | Low frequency | MAR                | 5              | 1.02E-03       | 1.97E-03 | 1.97E-03 | 1.91E-03    | 2.13E-03    | 1.85E-03 | 1.87E-03 | 2.13E-03 |
| Low         | Low frequency | MAR                | 20             | 3.34E-03       | 2.15E-03 | 2.15E-03 | 1.81E-03    | 2.24E-03    | 2.68E-03 | 3.04E-03 | 1.96E-03 |
| Low         | Low frequency | MAR                | 50             | 3.81E-03       | 3.01E-03 | 3.01E-03 | 2.13E-03    | 2.96E-03    | 2.11E-03 | 2.60E-03 | 1.84E-03 |

|      |               |      |    |          |          |          |          |          |          |          |          |
|------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| Low  | Rare          | MAR  | 1  | 7.68E-05 | 6.90E-05 | 6.90E-05 | 9.76E-05 | 1.27E-05 | 1.36E-04 | 2.48E-04 | 3.27E-05 |
| Low  | Rare          | MAR  | 5  | 7.76E-05 | 2.62E-05 | 2.62E-05 | 1.03E-04 | 5.69E-05 | 4.69E-05 | 2.50E-05 | 1.34E-05 |
| Low  | Rare          | MAR  | 20 | 2.72E-04 | 1.40E-04 | 1.40E-04 | 4.30E-04 | 1.57E-04 | 4.56E-04 | 2.44E-04 | 2.87E-04 |
| Low  | Rare          | MAR  | 50 | 1.89E-04 | 4.41E-04 | 4.41E-04 | 1.79E-05 | 1.06E-04 | 1.07E-04 | 9.44E-05 | 4.14E-04 |
| High | Common        | MCAR | 1  | 2.37E-04 | 4.53E-04 | 4.53E-04 | 2.34E-03 | 2.22E-04 | 8.47E-04 | 3.24E-03 | 1.44E-04 |
| High | Common        | MCAR | 5  | 2.31E-03 | 2.47E-03 | 2.47E-03 | 1.87E-03 | 1.40E-03 | 7.80E-04 | 2.08E-03 | 2.27E-03 |
| High | Common        | MCAR | 20 | 4.03E-03 | 4.04E-03 | 4.04E-03 | 1.28E-03 | 1.81E-03 | 2.99E-03 | 3.73E-03 | 2.25E-03 |
| High | Common        | MCAR | 50 | 8.90E-03 | 9.20E-03 | 9.20E-03 | 6.65E-03 | 4.91E-03 | 8.51E-03 | 9.79E-03 | 8.55E-03 |
| High | Low frequency | MCAR | 1  | 2.33E-03 | 2.07E-03 | 2.07E-03 | 4.37E-04 | 2.02E-03 | 2.06E-03 | 2.16E-03 | 2.04E-03 |
| High | Low frequency | MCAR | 5  | 1.84E-03 | 2.21E-03 | 2.21E-03 | 1.97E-03 | 1.99E-03 | 2.22E-03 | 2.53E-03 | 1.98E-03 |
| High | Low frequency | MCAR | 20 | 3.50E-03 | 3.13E-03 | 3.13E-03 | 2.58E-03 | 2.80E-03 | 4.33E-03 | 3.63E-03 | 2.46E-03 |
| High | Low frequency | MCAR | 50 | 3.42E-03 | 2.95E-03 | 2.95E-03 | 2.30E-03 | 3.57E-03 | 4.30E-03 | 2.74E-03 | 3.80E-03 |
| High | Rare          | MCAR | 1  | 1.45E-05 | 5.49E-06 | 5.50E-06 | 2.58E-04 | 6.78E-06 | 3.00E-05 | 1.32E-04 | 7.08E-06 |
| High | Rare          | MCAR | 5  | 1.75E-04 | 9.32E-06 | 9.32E-06 | 2.22E-04 | 1.85E-04 | 3.74E-04 | 8.23E-04 | 1.55E-04 |
| High | Rare          | MCAR | 20 | 1.52E-04 | 1.88E-04 | 1.88E-04 | 2.53E-04 | 1.53E-04 | 2.88E-04 | 6.02E-04 | 1.52E-04 |
| High | Rare          | MCAR | 50 | 1.58E-03 | 9.44E-05 | 9.44E-05 | 3.06E-04 | 6.07E-04 | 1.01E-03 | 3.25E-04 | 1.29E-03 |
| Low  | Common        | MCAR | 1  | 4.78E-04 | 1.47E-04 | 1.47E-04 | 7.85E-04 | 2.32E-04 | 6.77E-04 | 2.39E-04 | 6.17E-04 |
| Low  | Common        | MCAR | 5  | 8.49E-04 | 1.43E-03 | 2.69E-03 | 1.38E-04 | 6.71E-04 | 2.58E-03 | 3.85E-03 | 1.26E-03 |
| Low  | Common        | MCAR | 20 | 2.26E-03 | 2.69E-03 | 1.47E-04 | 3.36E-03 | 2.41E-03 | 3.22E-03 | 2.37E-03 | 2.07E-03 |
| Low  | Common        | MCAR | 50 | 3.05E-03 | 3.60E-03 | 2.69E-03 | 2.53E-03 | 3.70E-03 | 6.04E-03 | 7.76E-03 | 2.15E-03 |
| Low  | Low frequency | MCAR | 1  | 9.76E-05 | 6.77E-05 | 6.77E-05 | 6.88E-04 | 3.49E-05 | 2.26E-04 | 3.74E-04 | 4.44E-05 |
| Low  | Low frequency | MCAR | 5  | 4.08E-04 | 3.99E-04 | 3.99E-04 | 6.47E-04 | 5.06E-04 | 1.25E-03 | 1.74E-03 | 4.83E-04 |
| Low  | Low frequency | MCAR | 20 | 4.13E-04 | 6.59E-04 | 6.59E-04 | 1.20E-03 | 9.19E-04 | 6.88E-04 | 8.76E-04 | 6.93E-04 |

|             |               |      |    |          |          |          |          |          |          |          |          |
|-------------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| <b>Low</b>  | Low frequency | MCAR | 50 | 1.49E-03 | 2.45E-03 | 2.45E-03 | 6.89E-04 | 1.16E-03 | 1.71E-03 | 2.59E-03 | 1.01E-03 |
| <b>Low</b>  | Rare          | MCAR | 1  | 3.44E-06 | 1.36E-06 | 1.38E-06 | 1.80E-04 | 1.53E-06 | 1.80E-04 | 1.80E-04 | 0.00E-00 |
| <b>Low</b>  | Rare          | MCAR | 5  | 1.65E-04 | 9.35E-05 | 9.35E-05 | 2.10E-05 | 1.45E-05 | 2.10E-05 | 2.10E-05 | 3.44E-06 |
| <b>Low</b>  | Rare          | MCAR | 20 | 3.82E-04 | 5.41E-04 | 5.41E-04 | 5.48E-04 | 3.20E-04 | 5.48E-04 | 5.48E-04 | 1.65E-04 |
| <b>Low</b>  | Rare          | MCAR | 50 | 7.21E-04 | 2.39E-04 | 2.39E-04 | 2.81E-04 | 2.17E-04 | 2.81E-04 | 2.81E-04 | 3.82E-04 |
| <b>High</b> | Common        | MNAR | 1  | 1.48E-03 | 1.67E-03 | 1.67E-03 | 1.50E-03 | 1.11E-02 | 1.77E-03 | 6.70E-05 | 1.91E-03 |
| <b>High</b> | Common        | MNAR | 5  | 3.60E-03 | 3.54E-03 | 3.54E-03 | 3.48E-03 | 8.34E-03 | 2.32E-03 | 5.56E-04 | 3.90E-03 |
| <b>High</b> | Common        | MNAR | 20 | 3.90E-03 | 5.38E-03 | 5.38E-03 | 1.05E-02 | 9.99E-03 | 1.08E-02 | 2.11E-03 | 9.15E-03 |
| <b>High</b> | Common        | MNAR | 50 | 9.17E-03 | 6.16E-03 | 6.16E-03 | 7.87E-03 | 1.35E-02 | 1.37E-02 | 5.26E-03 | 9.10E-03 |
| <b>High</b> | Low frequency | MNAR | 1  | 5.05E-04 | 3.97E-04 | 3.97E-04 | 5.93E-04 | 6.01E-04 | 2.66E-04 | 8.76E-05 | 5.00E-04 |
| <b>High</b> | Low frequency | MNAR | 5  | 6.86E-04 | 5.22E-04 | 5.22E-04 | 8.56E-04 | 7.58E-04 | 4.94E-04 | 3.64E-05 | 7.77E-04 |
| <b>High</b> | Low frequency | MNAR | 20 | 2.07E-03 | 7.28E-04 | 7.29E-04 | 8.34E-04 | 1.35E-03 | 1.49E-03 | 2.39E-04 | 8.98E-04 |
| <b>High</b> | Low frequency | MNAR | 50 | 4.94E-03 | 1.90E-03 | 1.90E-03 | 2.66E-03 | 2.32E-03 | 1.21E-03 | 2.70E-04 | 1.66E-03 |
| <b>High</b> | Rare          | MNAR | 1  | 1.92E-05 | 8.24E-06 | 8.28E-06 | 7.69E-06 | 8.87E-06 | 7.20E-06 | 5.17E-06 | 1.09E-05 |
| <b>High</b> | Rare          | MNAR | 5  | 1.00E-04 | 5.70E-05 | 5.70E-05 | 4.78E-05 | 4.76E-05 | 2.69E-05 | 8.71E-06 | 6.73E-05 |
| <b>High</b> | Rare          | MNAR | 20 | 1.65E-04 | 7.94E-05 | 7.94E-05 | 8.72E-05 | 1.38E-04 | 2.24E-04 | 7.43E-05 | 1.65E-04 |
| <b>High</b> | Rare          | MNAR | 50 | 4.11E-04 | 1.11E-03 | 1.11E-03 | 1.53E-03 | 5.77E-04 | 1.11E-03 | 2.31E-04 | 9.86E-04 |
| <b>Low</b>  | Common        | MNAR | 1  | 5.60E-04 | 6.39E-04 | 6.39E-04 | 9.08E-04 | 3.45E-04 | 6.47E-04 | 9.48E-05 | 1.04E-03 |
| <b>Low</b>  | Common        | MNAR | 5  | 1.68E-03 | 1.29E-03 | 3.68E-04 | 1.53E-03 | 4.29E-04 | 1.51E-03 | 2.30E-04 | 2.23E-03 |
| <b>Low</b>  | Common        | MNAR | 20 | 2.54E-03 | 3.68E-04 | 5.23E-03 | 2.28E-03 | 2.16E-03 | 2.33E-03 | 1.52E-03 | 1.44E-03 |
| <b>Low</b>  | Common        | MNAR | 50 | 5.08E-03 | 5.23E-03 | 5.23E-03 | 6.32E-03 | 4.92E-03 | 6.25E-03 | 1.93E-03 | 7.33E-03 |
| <b>Low</b>  | Low frequency | MNAR | 1  | 5.05E-04 | 1.96E-03 | 1.96E-03 | 1.97E-03 | 1.96E-03 | 1.97E-03 | 2.00E-03 | 1.94E-03 |
| <b>Low</b>  | Low frequency | MNAR | 5  | 6.86E-04 | 1.71E-03 | 1.71E-03 | 1.55E-03 | 1.63E-03 | 1.74E-03 | 2.00E-03 | 1.62E-03 |

|            |               |      |    |          |          |          |          |          |          |          |          |
|------------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| <b>Low</b> | Low frequency | MNAR | 20 | 2.07E-03 | 1.64E-03 | 1.64E-03 | 1.30E-03 | 1.23E-03 | 1.81E-03 | 1.81E-03 | 1.75E-03 |
| <b>Low</b> | Low frequency | MNAR | 50 | 4.94E-03 | 1.80E-03 | 1.80E-03 | 2.71E-03 | 2.40E-03 | 3.41E-03 | 1.56E-03 | 3.22E-03 |
| <b>Low</b> | Rare          | MNAR | 1  | 1.78E-05 | 6.46E-06 | 6.43E-06 | 7.33E-06 | 6.43E-06 | 5.63E-06 | 6.48E-07 | 8.65E-06 |
| <b>Low</b> | Rare          | MNAR | 5  | 1.82E-04 | 1.43E-04 | 1.43E-04 | 1.98E-04 | 1.83E-04 | 8.52E-05 | 9.62E-05 | 2.10E-04 |
| <b>Low</b> | Rare          | MNAR | 20 | 1.82E-04 | 1.43E-04 | 1.43E-04 | 3.59E-04 | 2.64E-04 | 8.52E-05 | 7.33E-05 | 2.19E-04 |
| <b>Low</b> | Rare          | MNAR | 50 | 1.25E-03 | 3.68E-04 | 3.68E-04 | 9.45E-04 | 3.59E-04 | 6.52E-04 | 1.37E-04 | 5.01E-04 |

Supplementary Table 8 The RMSE of regression coefficient between different imputation methods in 5,000 individuals and 30 phenotypes.

| Correlation | Variant       | Missing mechanisms | proportion (%) | Complete Cases | SI MICE  | MI MICE  | SI AMELIAII | MI AMELIAII | k-NN     | QRILC    | RF       |
|-------------|---------------|--------------------|----------------|----------------|----------|----------|-------------|-------------|----------|----------|----------|
| High        | Common        | MAR                | 1              | 8.77E-04       | 8.73E-04 | 5.91E-04 | 7.58E-04    | 5.48E-04    | 9.18E-04 | 2.08E-03 | 5.29E-04 |
| High        | Common        | MAR                | 5              | 1.89E-03       | 1.30E-03 | 1.02E-03 | 1.47E-03    | 1.26E-03    | 1.37E-03 | 2.67E-03 | 1.10E-03 |
| High        | Common        | MAR                | 20             | 4.81E-03       | 2.78E-03 | 2.63E-03 | 4.04E-03    | 2.63E-03    | 3.83E-03 | 3.92E-03 | 3.59E-03 |
| High        | Common        | MAR                | 50             | 6.52E-03       | 4.77E-03 | 3.09E-03 | 3.13E-03    | 3.08E-03    | 3.73E-03 | 3.44E-03 | 4.67E-03 |
| High        | Low frequency | MAR                | 1              | 5.63E-04       | 5.71E-04 | 4.02E-04 | 4.72E-04    | 3.42E-04    | 2.63E-04 | 1.44E-03 | 2.65E-04 |
| High        | Low frequency | MAR                | 1              | 5.63E-04       | 5.71E-04 | 4.02E-04 | 4.72E-04    | 3.42E-04    | 2.63E-04 | 1.44E-03 | 2.65E-04 |
| High        | Low frequency | MAR                | 5              | 1.69E-03       | 1.28E-03 | 1.34E-03 | 1.33E-03    | 1.42E-03    | 1.67E-03 | 1.96E-03 | 1.04E-03 |
| High        | Low frequency | MAR                | 20             | 2.60E-03       | 4.38E-03 | 3.47E-03 | 1.99E-03    | 3.33E-03    | 1.96E-03 | 2.24E-03 | 9.56E-04 |
| High        | Low frequency | MAR                | 50             | 5.02E-03       | 7.91E-03 | 7.39E-03 | 3.44E-03    | 7.82E-03    | 3.33E-03 | 4.95E-03 | 2.79E-03 |
| High        | Rare          | MAR                | 1              | 5.63E-04       | 9.14E-04 | 9.14E-04 | 9.14E-04    | 9.12E-04    | 9.13E-04 | 9.12E-04 | 9.13E-04 |
| High        | Rare          | MAR                | 5              | 1.69E-03       | 9.15E-04 | 9.15E-04 | 9.20E-04    | 9.15E-04    | 9.15E-04 | 9.10E-04 | 9.17E-04 |
| High        | Rare          | MAR                | 20             | 2.60E-03       | 9.21E-04 | 9.02E-04 | 9.21E-04    | 9.10E-04    | 9.13E-04 | 8.81E-04 | 9.08E-04 |
| High        | Rare          | MAR                | 50             | 5.02E-03       | 9.72E-04 | 9.15E-04 | 9.13E-04    | 9.23E-04    | 9.34E-04 | 9.17E-04 | 9.39E-04 |
| Low         | Common        | MAR                | 1              | 2.01E-04       | 1.07E-04 | 1.05E-04 | 1.50E-04    | 8.48E-05    | 1.11E-04 | 1.03E-04 | 8.36E-05 |
| Low         | Common        | MAR                | 5              | 4.99E-04       | 2.96E-04 | 2.71E-04 | 3.11E-04    | 3.32E-04    | 2.47E-04 | 2.11E-04 | 2.50E-04 |
| Low         | Common        | MAR                | 20             | 1.87E-03       | 1.45E-03 | 1.06E-03 | 8.98E-04    | 1.18E-03    | 1.01E-03 | 9.84E-04 | 1.05E-03 |
| Low         | Common        | MAR                | 50             | 2.63E-03       | 1.63E-03 | 1.52E-03 | 1.40E-03    | 1.65E-03    | 1.67E-03 | 1.17E-03 | 1.68E-03 |
| Low         | Low frequency | MAR                | 1              | 2.63E-04       | 2.77E-04 | 1.75E-04 | 2.83E-04    | 1.95E-04    | 2.54E-04 | 2.03E-04 | 2.02E-04 |
| Low         | Low frequency | MAR                | 5              | 2.77E-04       | 2.53E-04 | 2.10E-04 | 2.20E-04    | 2.09E-04    | 1.52E-04 | 2.13E-04 | 2.10E-04 |
| Low         | Low frequency | MAR                | 20             | 6.81E-04       | 4.70E-04 | 3.89E-04 | 7.34E-04    | 4.52E-04    | 3.22E-04 | 5.65E-04 | 3.38E-04 |

|             |               |      |    |          |          |          |          |          |          |          |          |
|-------------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| <b>Low</b>  | Low frequency | MAR  | 50 | 1.09E-03 | 6.10E-04 | 2.99E-04 | 5.76E-04 | 3.91E-04 | 5.14E-04 | 7.14E-04 | 6.10E-04 |
| <b>Low</b>  | Rare          | MAR  | 1  | 1.59E-04 | 1.66E-04 | 9.35E-05 | 1.54E-04 | 1.09E-04 | 1.59E-04 | 1.53E-04 | 8.86E-05 |
| <b>Low</b>  | Rare          | MAR  | 5  | 1.63E-04 | 1.45E-04 | 1.17E-04 | 1.45E-04 | 1.05E-04 | 8.52E-05 | 1.47E-04 | 8.80E-05 |
| <b>Low</b>  | Rare          | MAR  | 20 | 4.17E-04 | 2.30E-04 | 2.40E-04 | 3.29E-04 | 2.71E-04 | 3.53E-04 | 2.78E-04 | 3.31E-04 |
| <b>Low</b>  | Rare          | MAR  | 50 | 6.13E-04 | 4.73E-04 | 4.05E-04 | 2.69E-04 | 3.53E-04 | 3.23E-04 | 2.04E-04 | 3.69E-04 |
| <b>High</b> | Common        | MCAR | 1  | 9.68E-04 | 8.52E-04 | 5.38E-04 | 9.92E-04 | 6.78E-04 | 5.82E-04 | 8.56E-04 | 6.37E-04 |
| <b>High</b> | Common        | MCAR | 5  | 2.29E-03 | 1.82E-03 | 1.56E-03 | 1.92E-03 | 1.47E-03 | 1.66E-03 | 3.41E-03 | 1.49E-03 |
| <b>High</b> | Common        | MCAR | 20 | 4.56E-03 | 4.41E-03 | 3.57E-03 | 4.10E-03 | 2.78E-03 | 3.34E-03 | 6.18E-03 | 3.08E-03 |
| <b>High</b> | Common        | MCAR | 50 | 1.12E-02 | 5.72E-03 | 4.90E-03 | 6.45E-03 | 3.82E-03 | 5.92E-03 | 8.44E-03 | 5.15E-03 |
| <b>High</b> | Low frequency | MCAR | 1  | 7.14E-04 | 5.37E-04 | 4.64E-04 | 6.60E-04 | 4.38E-04 | 4.51E-04 | 5.55E-04 | 4.42E-04 |
| <b>High</b> | Low frequency | MCAR | 5  | 2.00E-03 | 1.31E-03 | 1.36E-03 | 1.82E-03 | 1.34E-03 | 1.34E-03 | 3.16E-03 | 1.20E-03 |
| <b>High</b> | Low frequency | MCAR | 20 | 2.99E-03 | 1.94E-03 | 2.16E-03 | 2.39E-03 | 2.24E-03 | 1.93E-03 | 4.27E-03 | 1.87E-03 |
| <b>High</b> | Low frequency | MCAR | 50 | 6.51E-03 | 5.06E-03 | 3.54E-03 | 3.25E-03 | 3.96E-03 | 3.35E-03 | 4.85E-03 | 3.06E-03 |
| <b>High</b> | Rare          | MCAR | 1  | 2.81E-04 | 2.78E-04 | 1.26E-04 | 1.84E-04 | 1.83E-04 | 1.92E-04 | 5.66E-04 | 1.42E-04 |
| <b>High</b> | Rare          | MCAR | 5  | 1.02E-03 | 8.79E-04 | 6.28E-04 | 6.13E-04 | 6.62E-04 | 6.65E-04 | 1.11E-03 | 5.78E-04 |
| <b>High</b> | Rare          | MCAR | 20 | 1.94E-03 | 1.52E-03 | 9.42E-04 | 1.55E-03 | 1.21E-03 | 1.02E-03 | 1.94E-03 | 1.00E-03 |
| <b>High</b> | Rare          | MCAR | 50 | 4.19E-03 | 2.69E-03 | 1.95E-03 | 2.02E-03 | 1.83E-03 | 2.36E-03 | 3.45E-03 | 1.83E-03 |
| <b>Low</b>  | Common        | MCAR | 1  | 2.32E-04 | 1.65E-04 | 1.79E-04 | 2.37E-04 | 1.72E-04 | 2.43E-04 | 5.54E-04 | 1.62E-04 |
| <b>Low</b>  | Common        | MCAR | 5  | 7.95E-04 | 5.90E-04 | 4.25E-04 | 4.22E-04 | 4.78E-04 | 5.14E-04 | 5.42E-04 | 4.89E-04 |
| <b>Low</b>  | Common        | MCAR | 20 | 1.76E-03 | 9.75E-04 | 8.02E-04 | 1.05E-03 | 7.86E-04 | 7.41E-04 | 8.65E-04 | 7.13E-04 |
| <b>Low</b>  | Common        | MCAR | 50 | 2.82E-03 | 1.78E-03 | 1.23E-03 | 1.59E-03 | 1.06E-03 | 1.24E-03 | 1.60E-03 | 1.24E-03 |
| <b>Low</b>  | Low frequency | MCAR | 1  | 1.28E-05 | 9.59E-06 | 6.61E-06 | 6.28E-06 | 5.59E-06 | 1.13E-05 | 7.11E-05 | 5.26E-06 |
| <b>Low</b>  | Low frequency | MCAR | 5  | 4.77E-04 | 4.11E-04 | 3.16E-04 | 4.47E-04 | 2.91E-04 | 4.35E-04 | 3.60E-04 | 3.15E-04 |

|             |               |      |    |          |          |          |          |          |          |          |          |
|-------------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| <b>Low</b>  | Low frequency | MCAR | 20 | 5.12E-04 | 5.82E-04 | 3.57E-04 | 4.33E-04 | 3.76E-04 | 3.74E-04 | 4.94E-04 | 3.49E-04 |
| <b>Low</b>  | Low frequency | MCAR | 50 | 9.27E-04 | 7.25E-04 | 5.37E-04 | 6.22E-04 | 4.58E-04 | 5.21E-04 | 5.19E-04 | 5.18E-04 |
| <b>Low</b>  | Rare          | MCAR | 1  | 1.54E-05 | 1.06E-05 | 8.93E-06 | 9.30E-06 | 8.13E-06 | 1.32E-05 | 1.54E-05 | 9.61E-06 |
| <b>Low</b>  | Rare          | MCAR | 5  | 4.19E-05 | 1.37E-05 | 1.06E-05 | 1.99E-05 | 1.21E-05 | 2.86E-05 | 7.24E-05 | 1.80E-05 |
| <b>Low</b>  | Rare          | MCAR | 20 | 4.22E-04 | 4.16E-04 | 2.90E-04 | 2.43E-04 | 2.93E-04 | 3.53E-04 | 3.95E-04 | 2.65E-04 |
| <b>Low</b>  | Rare          | MCAR | 50 | 7.63E-04 | 4.34E-04 | 3.49E-04 | 3.46E-04 | 3.71E-04 | 3.10E-04 | 3.83E-04 | 3.26E-04 |
| <b>High</b> | Common        | MNAR | 1  | 1.59E-04 | 9.80E-03 | 9.77E-03 | 9.79E-03 | 9.79E-03 | 9.76E-03 | 9.76E-03 | 9.77E-03 |
| <b>High</b> | Common        | MNAR | 5  | 1.63E-04 | 1.01E-02 | 9.84E-03 | 9.84E-03 | 9.66E-03 | 9.81E-03 | 9.77E-03 | 9.88E-03 |
| <b>High</b> | Common        | MNAR | 20 | 4.17E-04 | 1.04E-02 | 1.04E-02 | 1.08E-02 | 9.52E-03 | 1.05E-02 | 9.64E-03 | 1.06E-02 |
| <b>High</b> | Common        | MNAR | 50 | 6.13E-04 | 9.46E-03 | 9.37E-03 | 9.75E-03 | 9.74E-03 | 9.39E-03 | 9.97E-03 | 1.06E-02 |
| <b>High</b> | Low frequency | MNAR | 1  | 7.54E-04 | 4.40E-04 | 4.18E-04 | 3.76E-04 | 3.97E-04 | 4.07E-04 | 7.92E-05 | 4.40E-04 |
| <b>High</b> | Low frequency | MNAR | 5  | 1.52E-03 | 8.15E-04 | 7.20E-04 | 8.14E-04 | 1.43E-03 | 7.65E-04 | 1.81E-04 | 7.96E-04 |
| <b>High</b> | Low frequency | MNAR | 20 | 3.25E-03 | 1.82E-03 | 1.66E-03 | 1.85E-03 | 2.95E-03 | 1.67E-03 | 3.65E-04 | 1.72E-03 |
| <b>High</b> | Low frequency | MNAR | 50 | 5.83E-03 | 3.20E-03 | 2.76E-03 | 2.87E-03 | 7.09E-03 | 2.82E-03 | 9.29E-04 | 2.90E-03 |
| <b>High</b> | Rare          | MNAR | 1  | 6.57E-04 | 4.74E-04 | 4.43E-04 | 5.16E-04 | 3.80E-04 | 3.34E-04 | 1.59E-04 | 4.43E-04 |
| <b>High</b> | Rare          | MNAR | 5  | 7.77E-04 | 6.91E-04 | 5.96E-04 | 6.38E-04 | 5.98E-04 | 6.44E-04 | 2.48E-04 | 6.11E-04 |
| <b>High</b> | Rare          | MNAR | 20 | 1.69E-03 | 1.22E-03 | 1.24E-03 | 1.29E-03 | 1.14E-03 | 1.15E-03 | 5.81E-04 | 1.21E-03 |
| <b>High</b> | Rare          | MNAR | 50 | 4.67E-03 | 2.76E-03 | 2.41E-03 | 2.76E-03 | 2.19E-03 | 2.40E-03 | 1.34E-03 | 2.44E-03 |
| <b>Low</b>  | Common        | MNAR | 1  | 2.13E-04 | 1.39E-04 | 1.34E-04 | 1.67E-04 | 1.52E-04 | 1.10E-04 | 2.39E-05 | 1.25E-04 |
| <b>Low</b>  | Common        | MNAR | 5  | 6.67E-04 | 3.74E-04 | 3.66E-04 | 4.35E-04 | 3.61E-04 | 3.82E-04 | 6.35E-05 | 4.09E-04 |
| <b>Low</b>  | Common        | MNAR | 20 | 1.60E-03 | 8.12E-04 | 7.90E-04 | 7.60E-04 | 7.78E-04 | 7.93E-04 | 1.44E-04 | 8.21E-04 |
| <b>Low</b>  | Common        | MNAR | 50 | 3.29E-03 | 1.09E-03 | 1.27E-03 | 1.43E-03 | 1.21E-03 | 1.38E-03 | 4.37E-04 | 1.41E-03 |
| <b>Low</b>  | Low frequency | MNAR | 1  | 4.05E-04 | 2.89E-04 | 3.37E-04 | 3.25E-04 | 3.11E-04 | 3.05E-04 | 1.08E-04 | 3.09E-04 |

|            |               |      |    |          |          |          |          |          |          |          |          |
|------------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| <b>Low</b> | Low frequency | MNAR | 5  | 4.17E-04 | 4.27E-04 | 3.75E-04 | 3.19E-04 | 3.43E-04 | 3.41E-04 | 1.03E-04 | 3.47E-04 |
| <b>Low</b> | Low frequency | MNAR | 20 | 6.72E-04 | 4.39E-04 | 4.86E-04 | 6.07E-04 | 4.92E-04 | 7.80E-04 | 2.45E-04 | 4.57E-04 |
| <b>Low</b> | Low frequency | MNAR | 50 | 1.01E-03 | 6.77E-04 | 6.25E-04 | 8.48E-04 | 6.83E-04 | 7.56E-04 | 4.32E-04 | 7.78E-04 |
| <b>Low</b> | Rare          | MNAR | 1  | 1.20E-05 | 5.74E-06 | 5.74E-06 | 5.95E-06 | 5.57E-06 | 6.30E-06 | 1.09E-06 | 6.25E-06 |
| <b>Low</b> | Rare          | MNAR | 5  | 2.35E-04 | 1.16E-04 | 1.27E-04 | 1.36E-04 | 9.96E-05 | 1.13E-04 | 1.69E-05 | 1.23E-04 |
| <b>Low</b> | Rare          | MNAR | 20 | 2.70E-04 | 1.19E-04 | 1.22E-04 | 1.20E-04 | 1.13E-04 | 1.12E-04 | 3.84E-05 | 1.28E-04 |
| <b>Low</b> | Rare          | MNAR | 50 | 6.50E-04 | 2.16E-04 | 2.01E-04 | 2.29E-04 | 1.91E-04 | 2.08E-04 | 7.41E-05 | 2.28E-04 |

Supplementary Table 9 The RMSE of regression coefficient between different imputation methods in 5,000 individuals and 120 phenotypes.

| Correlation | Variant       | Missing mechanisms | proportion (%) | Complete Cases | SI MICE  | MI MICE  | SI AMELIAII | MI AMELIAII | k-NN     | QRILC    | RF       |
|-------------|---------------|--------------------|----------------|----------------|----------|----------|-------------|-------------|----------|----------|----------|
| High        | Common        | MAR                | 1              | 9.70E-04       | 6.54E-04 | 6.54E-04 | 7.70E-04    | 4.93E-04    | 5.17E-04 | 1.04E-03 | 4.95E-04 |
| High        | Common        | MAR                | 5              | 2.45E-03       | 2.22E-03 | 2.22E-03 | 2.04E-03    | 1.56E-03    | 1.70E-03 | 1.92E-03 | 1.47E-03 |
| High        | Common        | MAR                | 20             | 4.72E-03       | 3.39E-03 | 3.39E-03 | 3.42E-03    | 3.02E-03    | 2.78E-03 | 2.93E-03 | 2.80E-03 |
| High        | Common        | MAR                | 50             | 8.08E-03       | 5.90E-03 | 5.90E-03 | 7.32E-03    | 6.22E-03    | 5.31E-03 | 6.42E-03 | 4.95E-03 |
| High        | Low frequency | MAR                | 1              | 6.91E-04       | 4.91E-04 | 4.91E-04 | 5.01E-04    | 3.39E-04    | 3.77E-04 | 5.92E-04 | 3.39E-04 |
| High        | Low frequency | MAR                | 5              | 1.65E-03       | 1.33E-03 | 1.33E-03 | 1.12E-03    | 1.01E-03    | 1.14E-03 | 1.00E-03 | 8.92E-04 |
| High        | Low frequency | MAR                | 20             | 3.89E-03       | 2.89E-03 | 2.89E-03 | 2.64E-03    | 2.39E-03    | 1.98E-03 | 2.47E-03 | 2.15E-03 |
| High        | Low frequency | MAR                | 50             | 7.23E-03       | 5.91E-03 | 5.91E-03 | 4.68E-03    | 4.16E-03    | 3.21E-03 | 3.75E-03 | 3.32E-03 |
| High        | Rare          | MAR                | 1              | 3.15E-04       | 2.27E-04 | 2.27E-04 | 2.66E-04    | 2.13E-04    | 1.96E-04 | 3.25E-04 | 1.87E-04 |
| High        | Rare          | MAR                | 5              | 8.13E-04       | 5.93E-04 | 5.93E-04 | 6.01E-04    | 5.16E-04    | 5.12E-04 | 6.48E-04 | 5.16E-04 |
| High        | Rare          | MAR                | 20             | 1.85E-03       | 1.45E-03 | 1.45E-03 | 1.83E-03    | 1.36E-03    | 1.30E-03 | 1.78E-03 | 1.31E-03 |
| High        | Rare          | MAR                | 50             | 3.72E-03       | 2.85E-03 | 2.85E-03 | 3.32E-03    | 2.68E-03    | 2.48E-03 | 2.20E-03 | 2.36E-03 |
| Low         | Common        | MAR                | 1              | 3.15E-04       | 4.49E-04 | 4.49E-04 | 4.50E-04    | 4.50E-04    | 4.50E-04 | 4.50E-04 | 4.50E-04 |
| Low         | Common        | MAR                | 5              | 8.13E-04       | 4.46E-04 | 4.46E-04 | 4.51E-04    | 4.49E-04    | 4.48E-04 | 4.50E-04 | 4.49E-04 |
| Low         | Common        | MAR                | 20             | 1.85E-03       | 4.56E-04 | 4.56E-04 | 4.64E-04    | 4.59E-04    | 4.57E-04 | 4.57E-04 | 4.54E-04 |
| Low         | Common        | MAR                | 50             | 3.72E-03       | 4.76E-04 | 4.76E-04 | 4.99E-04    | 4.96E-04    | 4.81E-04 | 4.52E-04 | 4.79E-04 |
| Low         | Low frequency | MAR                | 1              | 3.15E-04       | 4.22E-04 | 4.22E-04 | 4.25E-04    | 4.22E-04    | 4.23E-04 | 4.23E-04 | 4.22E-04 |
| Low         | Low frequency | MAR                | 5              | 8.13E-04       | 4.21E-04 | 4.21E-04 | 4.23E-04    | 4.24E-04    | 4.25E-04 | 4.23E-04 | 4.23E-04 |
| Low         | Low frequency | MAR                | 20             | 1.85E-03       | 4.27E-04 | 4.27E-04 | 4.31E-04    | 4.25E-04    | 4.23E-04 | 4.16E-04 | 4.24E-04 |
| Low         | Low frequency | MAR                | 50             | 3.72E-03       | 4.25E-04 | 4.25E-04 | 4.27E-04    | 4.30E-04    | 4.25E-04 | 4.19E-04 | 4.25E-04 |

|      |               |      |    |          |          |          |          |          |          |          |          |
|------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| Low  | Rare          | MAR  | 1  | 1.90E-04 | 1.67E-04 | 1.67E-04 | 1.86E-04 | 1.38E-04 | 1.22E-04 | 1.26E-04 | 1.23E-04 |
| Low  | Rare          | MAR  | 5  | 2.07E-04 | 1.76E-04 | 1.76E-04 | 1.69E-04 | 1.43E-04 | 1.37E-04 | 1.45E-04 | 1.34E-04 |
| Low  | Rare          | MAR  | 20 | 4.94E-04 | 4.15E-04 | 4.15E-04 | 3.92E-04 | 3.36E-04 | 2.85E-04 | 3.60E-04 | 2.87E-04 |
| Low  | Rare          | MAR  | 50 | 7.61E-04 | 3.99E-04 | 3.99E-04 | 4.82E-04 | 4.14E-04 | 3.73E-04 | 4.23E-04 | 3.66E-04 |
| High | Common        | MCAR | 1  | 9.68E-04 | 8.52E-04 | 5.38E-04 | 9.92E-04 | 6.78E-04 | 5.82E-04 | 8.56E-04 | 6.37E-04 |
| High | Common        | MCAR | 5  | 2.29E-03 | 1.82E-03 | 1.56E-03 | 1.92E-03 | 1.47E-03 | 1.66E-03 | 3.41E-03 | 1.49E-03 |
| High | Common        | MCAR | 20 | 4.56E-03 | 4.41E-03 | 3.57E-03 | 4.10E-03 | 2.78E-03 | 3.34E-03 | 6.18E-03 | 3.08E-03 |
| High | Common        | MCAR | 50 | 1.12E-02 | 5.72E-03 | 4.90E-03 | 6.45E-03 | 3.82E-03 | 5.92E-03 | 8.44E-03 | 5.15E-03 |
| High | Low frequency | MCAR | 1  | 7.14E-04 | 5.37E-04 | 4.64E-04 | 6.60E-04 | 4.38E-04 | 4.51E-04 | 5.55E-04 | 4.42E-04 |
| High | Low frequency | MCAR | 5  | 2.00E-03 | 1.31E-03 | 1.36E-03 | 1.82E-03 | 1.34E-03 | 1.34E-03 | 3.16E-03 | 1.20E-03 |
| High | Low frequency | MCAR | 20 | 2.99E-03 | 1.94E-03 | 2.16E-03 | 2.39E-03 | 2.24E-03 | 1.93E-03 | 4.27E-03 | 1.87E-03 |
| High | Low frequency | MCAR | 50 | 6.51E-03 | 5.06E-03 | 3.54E-03 | 3.25E-03 | 3.96E-03 | 3.35E-03 | 4.85E-03 | 3.06E-03 |
| High | Rare          | MCAR | 1  | 2.81E-04 | 2.78E-04 | 1.26E-04 | 1.84E-04 | 1.83E-04 | 1.92E-04 | 5.66E-04 | 1.42E-04 |
| High | Rare          | MCAR | 5  | 1.02E-03 | 8.79E-04 | 6.28E-04 | 6.13E-04 | 6.62E-04 | 6.65E-04 | 1.11E-03 | 5.78E-04 |
| High | Rare          | MCAR | 20 | 1.94E-03 | 1.52E-03 | 9.42E-04 | 1.55E-03 | 1.21E-03 | 1.02E-03 | 1.94E-03 | 1.00E-03 |
| High | Rare          | MCAR | 50 | 4.19E-03 | 2.69E-03 | 1.95E-03 | 2.02E-03 | 1.83E-03 | 2.36E-03 | 3.45E-03 | 1.83E-03 |
| Low  | Common        | MCAR | 1  | 2.32E-04 | 1.65E-04 | 1.79E-04 | 2.37E-04 | 1.72E-04 | 2.43E-04 | 5.54E-04 | 1.62E-04 |
| Low  | Common        | MCAR | 5  | 7.95E-04 | 5.90E-04 | 4.25E-04 | 4.22E-04 | 4.78E-04 | 5.14E-04 | 5.42E-04 | 4.89E-04 |
| Low  | Common        | MCAR | 20 | 1.76E-03 | 9.75E-04 | 8.02E-04 | 1.05E-03 | 7.86E-04 | 7.41E-04 | 8.65E-04 | 7.13E-04 |
| Low  | Common        | MCAR | 50 | 2.82E-03 | 1.78E-03 | 1.23E-03 | 1.59E-03 | 1.06E-03 | 1.24E-03 | 1.60E-03 | 1.24E-03 |
| Low  | Low frequency | MCAR | 1  | 1.28E-05 | 9.59E-06 | 6.61E-06 | 6.28E-06 | 5.59E-06 | 1.13E-05 | 7.11E-05 | 5.26E-06 |
| Low  | Low frequency | MCAR | 5  | 4.77E-04 | 4.11E-04 | 3.16E-04 | 4.47E-04 | 2.91E-04 | 4.35E-04 | 3.60E-04 | 3.15E-04 |
| Low  | Low frequency | MCAR | 20 | 5.12E-04 | 5.82E-04 | 3.57E-04 | 4.33E-04 | 3.76E-04 | 3.74E-04 | 4.94E-04 | 3.49E-04 |

|             |               |      |    |          |          |          |          |          |          |          |          |
|-------------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| <b>Low</b>  | Low frequency | MCAR | 50 | 9.27E-04 | 7.25E-04 | 5.37E-04 | 6.22E-04 | 4.58E-04 | 5.21E-04 | 5.19E-04 | 5.18E-04 |
| <b>Low</b>  | Rare          | MCAR | 1  | 1.54E-05 | 1.06E-05 | 8.93E-06 | 9.30E-06 | 8.13E-06 | 1.32E-05 | 1.54E-05 | 9.61E-06 |
| <b>Low</b>  | Rare          | MCAR | 5  | 4.19E-05 | 1.37E-05 | 1.06E-05 | 1.99E-05 | 1.21E-05 | 2.86E-05 | 7.24E-05 | 1.80E-05 |
| <b>Low</b>  | Rare          | MCAR | 20 | 4.22E-04 | 4.16E-04 | 2.90E-04 | 2.43E-04 | 2.93E-04 | 3.53E-04 | 3.95E-04 | 2.65E-04 |
| <b>Low</b>  | Rare          | MCAR | 50 | 7.63E-04 | 4.34E-04 | 3.49E-04 | 3.46E-04 | 3.71E-04 | 3.10E-04 | 3.83E-04 | 3.26E-04 |
| <b>High</b> | Common        | MNAR | 1  | 1.59E-04 | 9.80E-03 | 9.77E-03 | 9.79E-03 | 9.79E-03 | 9.76E-03 | 9.76E-03 | 9.77E-03 |
| <b>High</b> | Common        | MNAR | 5  | 1.63E-04 | 1.01E-02 | 9.84E-03 | 9.84E-03 | 9.66E-03 | 9.81E-03 | 9.77E-03 | 9.88E-03 |
| <b>High</b> | Common        | MNAR | 20 | 4.17E-04 | 1.04E-02 | 1.04E-02 | 1.08E-02 | 9.52E-03 | 1.05E-02 | 9.64E-03 | 1.06E-02 |
| <b>High</b> | Common        | MNAR | 50 | 6.13E-04 | 9.46E-03 | 9.37E-03 | 9.75E-03 | 9.74E-03 | 9.39E-03 | 9.97E-03 | 1.06E-02 |
| <b>High</b> | Low frequency | MNAR | 1  | 7.54E-04 | 4.40E-04 | 4.18E-04 | 3.76E-04 | 3.97E-04 | 4.07E-04 | 7.92E-05 | 4.40E-04 |
| <b>High</b> | Low frequency | MNAR | 5  | 1.52E-03 | 8.15E-04 | 7.20E-04 | 8.14E-04 | 1.43E-03 | 7.65E-04 | 1.81E-04 | 7.96E-04 |
| <b>High</b> | Low frequency | MNAR | 20 | 3.25E-03 | 1.82E-03 | 1.66E-03 | 1.85E-03 | 2.95E-03 | 1.67E-03 | 3.65E-04 | 1.72E-03 |
| <b>High</b> | Low frequency | MNAR | 50 | 5.83E-03 | 3.20E-03 | 2.76E-03 | 2.87E-03 | 7.09E-03 | 2.82E-03 | 9.29E-04 | 2.90E-03 |
| <b>High</b> | Rare          | MNAR | 1  | 6.57E-04 | 4.74E-04 | 4.43E-04 | 5.16E-04 | 3.80E-04 | 3.34E-04 | 1.59E-04 | 4.43E-04 |
| <b>High</b> | Rare          | MNAR | 5  | 7.77E-04 | 6.91E-04 | 5.96E-04 | 6.38E-04 | 5.98E-04 | 6.44E-04 | 2.48E-04 | 6.11E-04 |
| <b>High</b> | Rare          | MNAR | 20 | 1.69E-03 | 1.22E-03 | 1.24E-03 | 1.29E-03 | 1.14E-03 | 1.15E-03 | 5.81E-04 | 1.21E-03 |
| <b>High</b> | Rare          | MNAR | 50 | 4.67E-03 | 2.76E-03 | 2.41E-03 | 2.76E-03 | 2.19E-03 | 2.40E-03 | 1.34E-03 | 2.44E-03 |
| <b>Low</b>  | Common        | MNAR | 1  | 2.13E-04 | 1.39E-04 | 1.34E-04 | 1.67E-04 | 1.52E-04 | 1.10E-04 | 2.39E-05 | 1.25E-04 |
| <b>Low</b>  | Common        | MNAR | 5  | 6.67E-04 | 3.74E-04 | 3.66E-04 | 4.35E-04 | 3.61E-04 | 3.82E-04 | 6.35E-05 | 4.09E-04 |
| <b>Low</b>  | Common        | MNAR | 20 | 1.60E-03 | 8.12E-04 | 7.90E-04 | 7.60E-04 | 7.78E-04 | 7.93E-04 | 1.44E-04 | 8.21E-04 |
| <b>Low</b>  | Common        | MNAR | 50 | 3.29E-03 | 1.09E-03 | 1.27E-03 | 1.43E-03 | 1.21E-03 | 1.38E-03 | 4.37E-04 | 1.41E-03 |
| <b>Low</b>  | Low frequency | MNAR | 1  | 4.05E-04 | 2.89E-04 | 3.37E-04 | 3.25E-04 | 3.11E-04 | 3.05E-04 | 1.08E-04 | 3.09E-04 |
| <b>Low</b>  | Low frequency | MNAR | 5  | 4.17E-04 | 4.27E-04 | 3.75E-04 | 3.19E-04 | 3.43E-04 | 3.41E-04 | 1.03E-04 | 3.47E-04 |

|            |               |      |    |           |           |           |           |           |           |           |           |
|------------|---------------|------|----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| <b>Low</b> | Low frequency | MNAR | 20 | 6.72E -04 | 4.39E -04 | 4.86E -04 | 6.07E -04 | 4.92E -04 | 7.80E -04 | 2.45E -04 | 4.57E -04 |
| <b>Low</b> | Low frequency | MNAR | 50 | 1.01E -03 | 6.77E -04 | 6.25E -04 | 8.48E -04 | 6.83E -04 | 7.56E -04 | 4.32E -04 | 7.78E -04 |
| <b>Low</b> | Rare          | MNAR | 1  | 1.20E -05 | 5.74E -06 | 5.74E -06 | 5.95E -06 | 5.57E -06 | 6.30E -06 | 1.09E -06 | 6.25E -06 |
| <b>Low</b> | Rare          | MNAR | 5  | 2.35E -04 | 1.16E -04 | 1.27E -04 | 1.36E -04 | 9.96E -05 | 1.13E -04 | 1.69E -05 | 1.23E -04 |
| <b>Low</b> | Rare          | MNAR | 20 | 2.70E -04 | 1.19E -04 | 1.22E -04 | 1.20E -04 | 1.13E -04 | 1.12E -04 | 3.84E -05 | 1.28E -04 |
| <b>Low</b> | Rare          | MNAR | 50 | 6.50E -04 | 2.16E -04 | 2.01E -04 | 2.29E -04 | 1.91E -04 | 2.08E -04 | 7.41E -05 | 2.28E -04 |

Supplementary Table 10 The RMSE of regression coefficient between different imputation methods in 50,000 individuals and three phenotypes.

| Correlation | Variant       | Missing mechanisms | proportion (%) | Complete Cases | SI MICE  | MI MICE  | SI AMELIAII | MI AMELIAII | k-NN     | QRILC    | RF       |
|-------------|---------------|--------------------|----------------|----------------|----------|----------|-------------|-------------|----------|----------|----------|
| High        | Common        | MAR                | 1              | 6.01E-05       | 1.04E-05 | 1.04E-05 | 9.46E-06    | 4.39E-05    | 1.81E-05 | 3.13E-05 | 3.27E-05 |
| High        | Common        | MAR                | 5              | 3.07E-04       | 1.72E-05 | 1.75E-05 | 2.14E-05    | 1.85E-04    | 1.20E-04 | 3.88E-05 | 1.55E-04 |
| High        | Common        | MAR                | 20             | 1.32E-03       | 5.72E-05 | 5.72E-05 | 4.66E-05    | 1.06E-03    | 6.45E-04 | 1.41E-04 | 6.94E-04 |
| High        | Common        | MAR                | 50             | 3.01E-05       | 1.08E-04 | 1.08E-04 | 8.79E-05    | 2.62E-03    | 1.85E-03 | 1.49E-03 | 2.01E-03 |
| High        | Low frequency | MAR                | 1              | 4.35E-05       | 5.80E-06 | 5.79E-06 | 4.72E-06    | 2.06E-05    | 1.57E-05 | 1.53E-05 | 2.19E-05 |
| High        | Low frequency | MAR                | 5              | 2.16E-04       | 1.63E-05 | 1.63E-05 | 1.54E-05    | 1.44E-04    | 9.65E-05 | 4.99E-05 | 1.11E-04 |
| High        | Low frequency | MAR                | 20             | 9.31E-04       | 3.17E-05 | 3.17E-05 | 3.24E-05    | 5.36E-04    | 4.64E-04 | 1.50E-04 | 4.92E-04 |
| High        | Low frequency | MAR                | 50             | 2.08E-05       | 5.88E-05 | 5.88E-05 | 6.21E-05    | 2.07E-03    | 1.40E-03 | 3.17E-04 | 1.44E-03 |
| High        | Rare          | MAR                | 1              | 3.27E-05       | 3.51E-06 | 3.52E-06 | 3.73E-06    | 8.37E-06    | 7.15E-06 | 1.29E-05 | 1.23E-05 |
| High        | Rare          | MAR                | 5              | 1.62E-04       | 9.86E-06 | 9.88E-06 | 8.75E-06    | 6.25E-05    | 4.96E-05 | 1.74E-05 | 6.42E-05 |
| High        | Rare          | MAR                | 20             | 4.55E-04       | 2.19E-05 | 2.19E-05 | 2.00E-05    | 3.32E-04    | 2.62E-04 | 5.44E-05 | 2.83E-04 |
| High        | Rare          | MAR                | 50             | 1.57E-05       | 4.34E-05 | 4.34E-05 | 3.62E-05    | 1.13E-03    | 7.81E-04 | 5.83E-04 | 8.18E-04 |
| Low         | Common        | MAR                | 1              | 3.27E-05       | 1.18E-03 | 1.18E-03 | 1.18E-03    | 1.17E-03    | 1.17E-03 | 1.18E-03 | 1.17E-03 |
| Low         | Common        | MAR                | 5              | 1.62E-04       | 1.17E-03 | 1.17E-03 | 1.17E-03    | 1.15E-03    | 1.16E-03 | 1.17E-03 | 1.15E-03 |
| Low         | Common        | MAR                | 20             | 4.55E-04       | 1.17E-03 | 1.17E-03 | 1.17E-03    | 1.05E-03    | 1.08E-03 | 1.14E-03 | 1.07E-03 |
| Low         | Common        | MAR                | 50             | 1.57E-05       | 1.16E-03 | 1.16E-03 | 1.17E-03    | 9.88E-04    | 9.73E-04 | 1.26E-03 | 9.61E-04 |
| Low         | Low frequency | MAR                | 1              | 3.27E-05       | 2.99E-03 | 2.99E-03 | 2.99E-03    | 2.97E-03    | 2.99E-03 | 2.99E-03 | 2.99E-03 |
| Low         | Low frequency | MAR                | 5              | 1.62E-04       | 2.99E-03 | 2.99E-03 | 2.99E-03    | 2.98E-03    | 2.98E-03 | 2.99E-03 | 2.98E-03 |
| Low         | Low frequency | MAR                | 20             | 4.55E-04       | 2.99E-03 | 2.99E-03 | 2.99E-03    | 2.90E-03    | 2.95E-03 | 3.00E-03 | 2.94E-03 |

|             |               |      |    |          |          |          |          |          |          |          |          |
|-------------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| <b>Low</b>  | Low frequency | MAR  | 50 | 1.57E-05 | 2.98E-03 | 2.98E-03 | 2.99E-03 | 2.83E-03 | 2.88E-03 | 3.09E-03 | 2.86E-03 |
| <b>Low</b>  | Rare          | MAR  | 1  | 1.84E-05 | 7.88E-07 | 7.92E-07 | 1.08E-06 | 1.64E-05 | 1.54E-06 | 2.42E-06 | 2.22E-06 |
| <b>Low</b>  | Rare          | MAR  | 5  | 1.53E-05 | 1.71E-06 | 1.72E-06 | 1.66E-06 | 2.31E-05 | 1.02E-05 | 5.22E-06 | 1.18E-05 |
| <b>Low</b>  | Rare          | MAR  | 20 | 3.05E-05 | 3.97E-06 | 3.98E-06 | 3.49E-06 | 1.53E-04 | 5.13E-05 | 1.71E-05 | 5.52E-05 |
| <b>Low</b>  | Rare          | MAR  | 50 | 9.11E-06 | 8.68E-06 | 8.68E-06 | 6.89E-06 | 2.11E-04 | 1.56E-04 | 3.54E-05 | 1.61E-04 |
| <b>High</b> | Common        | MCAR | 1  | 5.37E-05 | 1.09E-05 | 1.09E-05 | 9.93E-06 | 4.79E-05 | 1.57E-05 | 7.84E-05 | 2.76E-05 |
| <b>High</b> | Common        | MCAR | 5  | 3.25E-04 | 1.74E-05 | 1.74E-05 | 1.83E-05 | 1.68E-04 | 8.47E-05 | 1.10E-04 | 1.39E-04 |
| <b>High</b> | Common        | MCAR | 20 | 1.25E-03 | 5.92E-05 | 5.92E-05 | 4.65E-05 | 8.41E-04 | 5.31E-04 | 1.67E-04 | 6.34E-04 |
| <b>High</b> | Common        | MCAR | 50 | 2.64E-05 | 1.23E-04 | 1.23E-04 | 9.84E-05 | 2.97E-03 | 1.80E-03 | 1.60E-03 | 2.14E-03 |
| <b>High</b> | Low frequency | MCAR | 1  | 5.37E-05 | 2.99E-03 | 2.99E-03 | 2.99E-03 | 2.97E-03 | 2.98E-03 | 3.01E-03 | 2.97E-03 |
| <b>High</b> | Low frequency | MCAR | 5  | 3.25E-04 | 2.98E-03 | 2.98E-03 | 2.99E-03 | 2.93E-03 | 2.94E-03 | 3.01E-03 | 2.93E-03 |
| <b>High</b> | Low frequency | MCAR | 20 | 1.25E-03 | 2.97E-03 | 2.97E-03 | 2.98E-03 | 2.73E-03 | 2.78E-03 | 2.94E-03 | 2.72E-03 |
| <b>High</b> | Low frequency | MCAR | 50 | 2.64E-05 | 2.96E-03 | 2.96E-03 | 2.95E-03 | 2.40E-03 | 2.42E-03 | 3.26E-03 | 2.37E-03 |
| <b>High</b> | Rare          | MCAR | 1  | 2.43E-05 | 3.57E-06 | 3.57E-06 | 3.29E-06 | 1.83E-05 | 4.99E-06 | 3.13E-05 | 1.18E-05 |
| <b>High</b> | Rare          | MCAR | 5  | 1.39E-04 | 7.75E-06 | 7.71E-06 | 7.70E-06 | 5.26E-05 | 4.01E-05 | 4.12E-05 | 5.87E-05 |
| <b>High</b> | Rare          | MCAR | 20 | 5.94E-04 | 1.43E-05 | 1.43E-05 | 1.58E-05 | 2.85E-04 | 2.11E-04 | 7.20E-05 | 2.51E-04 |
| <b>High</b> | Rare          | MCAR | 50 | 1.18E-05 | 5.38E-05 | 5.38E-05 | 4.34E-05 | 1.21E-03 | 7.42E-04 | 6.32E-04 | 8.71E-04 |
| <b>Low</b>  | Common        | MCAR | 1  | 2.43E-05 | 1.18E-03 | 1.18E-03 | 1.18E-03 | 1.17E-03 | 1.17E-03 | 1.19E-03 | 1.17E-03 |
| <b>Low</b>  | Common        | MCAR | 5  | 1.39E-04 | 1.17E-03 | 1.17E-03 | 1.18E-03 | 1.15E-03 | 1.16E-03 | 1.18E-03 | 1.15E-03 |
| <b>Low</b>  | Common        | MCAR | 20 | 5.94E-04 | 1.17E-03 | 1.17E-03 | 1.17E-03 | 1.05E-03 | 1.09E-03 | 1.16E-03 | 1.07E-03 |
| <b>Low</b>  | Common        | MCAR | 50 | 1.18E-05 | 1.15E-03 | 1.15E-03 | 1.16E-03 | 1.04E-03 | 9.53E-04 | 1.28E-03 | 9.32E-04 |
| <b>Low</b>  | Low frequency | MCAR | 1  | 1.18E-05 | 9.49E-07 | 9.40E-07 | 1.17E-06 | 3.05E-06 | 1.82E-06 | 8.68E-06 | 3.11E-06 |
| <b>Low</b>  | Low frequency | MCAR | 5  | 2.33E-05 | 2.52E-06 | 2.51E-06 | 2.78E-06 | 3.04E-05 | 9.84E-06 | 1.18E-05 | 1.58E-05 |

|      |               |      |    |          |          |          |          |          |          |          |          |
|------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| Low  | Low frequency | MCAR | 20 | 9.67E-05 | 5.08E-06 | 5.12E-06 | 4.31E-06 | 1.40E-04 | 5.62E-05 | 2.03E-05 | 6.99E-05 |
| Low  | Low frequency | MCAR | 50 | 5.95E-06 | 1.25E-05 | 1.25E-05 | 1.26E-05 | 2.77E-04 | 2.04E-04 | 1.77E-04 | 2.38E-04 |
| Low  | Rare          | MCAR | 1  | 8.31E-06 | 6.58E-07 | 6.55E-07 | 7.13E-07 | 6.66E-07 | 1.61E-06 | 3.85E-06 | 2.26E-06 |
| Low  | Rare          | MCAR | 5  | 4.37E-05 | 1.56E-06 | 1.56E-06 | 1.85E-06 | 1.76E-06 | 8.50E-06 | 6.96E-06 | 1.14E-05 |
| Low  | Rare          | MCAR | 20 | 1.73E-04 | 3.64E-06 | 3.64E-06 | 4.94E-06 | 2.49E-05 | 4.17E-05 | 1.44E-05 | 4.97E-05 |
| Low  | Rare          | MCAR | 50 | 4.07E-06 | 9.90E-06 | 9.90E-06 | 9.57E-06 | 2.01E-04 | 1.61E-04 | 5.11E-05 | 1.71E-04 |
| High | Common        | MNAR | 1  | 6.85E-05 | 7.19E-06 | 7.19E-06 | 7.82E-06 | 6.93E-05 | 8.00E-06 | 4.14E-06 | 8.11E-06 |
| High | Common        | MNAR | 5  | 3.76E-04 | 3.55E-05 | 3.54E-05 | 3.39E-05 | 1.71E-04 | 4.20E-05 | 9.39E-06 | 4.24E-05 |
| High | Common        | MNAR | 20 | 1.35E-03 | 1.57E-04 | 1.57E-04 | 1.41E-04 | 1.16E-03 | 1.60E-04 | 2.75E-05 | 1.90E-04 |
| High | Common        | MNAR | 50 | 3.36E-05 | 6.96E-04 | 6.96E-04 | 5.54E-04 | 3.61E-03 | 7.35E-04 | 1.98E-04 | 9.08E-04 |
| High | Low frequency | MNAR | 1  | 4.01E-05 | 6.39E-06 | 6.39E-06 | 6.34E-06 | 6.73E-06 | 6.68E-06 | 2.78E-06 | 6.72E-06 |
| High | Low frequency | MNAR | 5  | 2.72E-04 | 2.08E-05 | 2.08E-05 | 2.07E-05 | 2.40E-05 | 2.26E-05 | 5.77E-06 | 2.28E-05 |
| High | Low frequency | MNAR | 20 | 1.08E-03 | 8.04E-05 | 8.04E-05 | 7.99E-05 | 9.40E-05 | 9.16E-05 | 1.04E-05 | 9.28E-05 |
| High | Low frequency | MNAR | 50 | 1.99E-05 | 2.65E-04 | 2.65E-04 | 2.57E-04 | 3.28E-04 | 3.14E-04 | 2.25E-05 | 3.18E-04 |
| High | Rare          | MNAR | 1  | 1.12E-05 | 4.56E-06 | 4.55E-06 | 4.35E-06 | 5.99E-06 | 4.78E-06 | 2.47E-06 | 4.83E-06 |
| High | Rare          | MNAR | 5  | 1.74E-04 | 1.65E-05 | 1.65E-05 | 1.56E-05 | 2.10E-05 | 1.89E-05 | 4.26E-06 | 1.90E-05 |
| High | Rare          | MNAR | 20 | 6.55E-04 | 6.66E-05 | 6.65E-05 | 5.94E-05 | 7.94E-05 | 7.20E-05 | 1.20E-05 | 8.07E-05 |
| High | Rare          | MNAR | 50 | 5.25E-06 | 2.88E-04 | 2.88E-04 | 2.34E-04 | 3.55E-04 | 2.81E-04 | 7.62E-05 | 3.73E-04 |
| Low  | Common        | MNAR | 1  | 1.12E-05 | 1.18E-03 | 1.18E-03 | 1.18E-03 | 1.18E-03 | 1.18E-03 | 1.18E-03 | 1.18E-03 |
| Low  | Common        | MNAR | 5  | 1.74E-04 | 1.17E-03 | 1.17E-03 | 1.17E-03 | 1.17E-03 | 1.17E-03 | 1.18E-03 | 1.17E-03 |
| Low  | Common        | MNAR | 20 | 6.55E-04 | 1.14E-03 | 1.14E-03 | 1.14E-03 | 1.14E-03 | 1.14E-03 | 1.17E-03 | 1.14E-03 |
| Low  | Common        | MNAR | 50 | 5.25E-06 | 1.07E-03 | 1.07E-03 | 1.07E-03 | 1.04E-03 | 1.05E-03 | 1.18E-03 | 1.05E-03 |
| Low  | Low frequency | MNAR | 1  | 1.46E-05 | 1.08E-06 | 1.05E-06 | 1.13E-06 | 3.13E-06 | 1.25E-06 | 6.12E-07 | 1.26E-06 |

|            |               |      |    |           |           |           |           |           |           |           |           |
|------------|---------------|------|----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| <b>Low</b> | Low frequency | MNAR | 5  | 6.62E -05 | 4.20E -06 | 4.20E -06 | 4.17E -06 | 5.26E -06 | 4.96E -06 | 1.13E -06 | 4.97E -06 |
| <b>Low</b> | Low frequency | MNAR | 20 | 1.97E -04 | 1.71E -05 | 1.71E -05 | 1.56E -05 | 1.72E -05 | 1.72E -05 | 3.05E -06 | 2.06E -05 |
| <b>Low</b> | Low frequency | MNAR | 50 | 7.32E -06 | 7.69E -05 | 7.69E -05 | 6.26E -05 | 8.92E -05 | 7.74E -05 | 1.91E -05 | 9.98E -05 |
| <b>Low</b> | Rare          | MNAR | 1  | 8.91E -06 | 7.03E -07 | 6.99E -07 | 6.84E -07 | 7.71E -07 | 7.60E -07 | 2.90E -07 | 7.61E -07 |
| <b>Low</b> | Rare          | MNAR | 5  | 4.79E -05 | 2.41E -06 | 2.41E -06 | 2.38E -06 | 2.42E -06 | 2.62E -06 | 6.20E -07 | 2.64E -06 |
| <b>Low</b> | Rare          | MNAR | 20 | 1.56E -04 | 9.23E -06 | 9.23E -06 | 8.96E -06 | 1.27E -05 | 1.03E -05 | 1.27E -06 | 1.04E -05 |
| <b>Low</b> | Rare          | MNAR | 50 | 4.40E -06 | 3.00E -05 | 3.00E -05 | 2.92E -05 | 3.82E -05 | 3.56E -05 | 2.37E -06 | 3.62E -05 |

Supplementary Table 11 The RMSE of regression coefficient between different imputation methods in 50,000 individuals and 30 phenotypes.

| Correlation | Variant       | Missing mechanisms | proportion (%) | Complete Cases | SI MICE  | MI MICE  | SI AMELIAII | MI AMELIAII | k-NN     | QRILC    | RF       |
|-------------|---------------|--------------------|----------------|----------------|----------|----------|-------------|-------------|----------|----------|----------|
| High        | Common        | MAR                | 1              | 2.98E-04       | 1.25E-04 | 1.25E-04 | 1.58E-04    | 2.52E-04    | 3.66E-04 | 3.30E-04 | 2.20E-04 |
| High        | Common        | MAR                | 5              | 9.74E-04       | 1.16E-03 | 1.16E-03 | 3.73E-04    | 4.18E-04    | 9.47E-04 | 1.17E-03 | 5.00E-04 |
| High        | Common        | MAR                | 20             | 2.07E-03       | 5.70E-04 | 5.70E-04 | 1.16E-03    | 1.86E-03    | 1.77E-03 | 1.47E-03 | 2.16E-03 |
| High        | Common        | MAR                | 50             | 1.57E-03       | 1.66E-03 | 1.66E-03 | 1.54E-03    | 5.51E-04    | 5.81E-04 | 2.07E-03 | 6.92E-04 |
| High        | Low frequency | MAR                | 1              | 3.38E-04       | 1.67E-04 | 1.67E-04 | 1.67E-04    | 1.67E-04    | 1.67E-04 | 1.67E-04 | 1.67E-04 |
| High        | Low frequency | MAR                | 5              | 6.88E-04       | 1.67E-04 | 1.67E-04 | 1.67E-04    | 1.67E-04    | 1.66E-04 | 1.65E-04 | 1.67E-04 |
| High        | Low frequency | MAR                | 20             | 2.00E-03       | 1.66E-03 | 1.66E-03 | 1.66E-03    | 1.66E-03    | 1.63E-03 | 1.64E-03 | 1.67E-03 |
| High        | Low frequency | MAR                | 50             | 4.58E-03       | 1.66E-03 | 1.66E-03 | 1.70E-03    | 1.69E-03    | 1.64E-03 | 1.67E-03 | 1.68E-03 |
| High        | Rare          | MAR                | 1              | 2.19E-04       | 1.12E-04 | 1.12E-04 | 1.35E-04    | 8.81E-05    | 2.29E-04 | 5.64E-04 | 7.75E-05 |
| High        | Rare          | MAR                | 5              | 1.80E-04       | 8.69E-05 | 8.66E-05 | 1.33E-04    | 1.47E-04    | 3.43E-04 | 8.23E-04 | 8.09E-05 |
| High        | Rare          | MAR                | 20             | 6.64E-04       | 8.02E-04 | 8.02E-04 | 5.63E-04    | 4.99E-04    | 4.58E-04 | 5.12E-04 | 5.77E-04 |
| High        | Rare          | MAR                | 50             | 1.17E-03       | 1.17E-03 | 1.17E-03 | 3.50E-04    | 7.22E-04    | 9.47E-04 | 8.01E-04 | 7.01E-04 |
| Low         | Common        | MAR                | 1              | 3.57E-04       | 3.45E-04 | 3.45E-04 | 3.17E-04    | 3.19E-04    | 3.40E-04 | 3.39E-04 | 3.47E-04 |
| Low         | Common        | MAR                | 5              | 3.55E-04       | 3.43E-04 | 3.43E-04 | 3.35E-04    | 3.19E-04    | 3.27E-04 | 3.28E-04 | 3.46E-04 |
| Low         | Common        | MAR                | 20             | 3.31E-04       | 3.30E-04 | 3.30E-04 | 3.62E-04    | 3.79E-04    | 3.49E-04 | 3.34E-04 | 3.51E-04 |
| Low         | Common        | MAR                | 50             | 3.57E-04       | 3.31E-04 | 3.31E-04 | 4.00E-04    | 3.44E-04    | 3.47E-04 | 3.39E-04 | 3.52E-04 |
| Low         | Low frequency | MAR                | 1              | 1.89E-05       | 3.09E-05 | 3.09E-05 | 1.25E-05    | 1.90E-05    | 1.22E-05 | 4.03E-05 | 4.52E-06 |
| Low         | Low frequency | MAR                | 5              | 8.31E-05       | 1.30E-04 | 1.30E-04 | 3.92E-05    | 7.86E-05    | 1.12E-04 | 1.82E-04 | 5.82E-05 |
| Low         | Low frequency | MAR                | 20             | 2.54E-04       | 7.94E-05 | 7.93E-05 | 1.79E-04    | 1.67E-04    | 4.07E-05 | 6.77E-05 | 1.64E-04 |
| Low         | Low frequency | MAR                | 50             | 5.87E-04       | 1.26E-04 | 1.26E-04 | 1.99E-05    | 1.12E-04    | 2.16E-04 | 2.26E-04 | 1.27E-04 |
| Low         | Rare          | MAR                | 1              | 1.74E-04       | 1.82E-04 | 1.82E-04 | 1.83E-04    | 1.82E-04    | 1.83E-04 | 1.77E-04 | 1.83E-04 |
| Low         | Rare          | MAR                | 5              | 1.88E-04       | 2.19E-04 | 2.19E-04 | 1.86E-04    | 1.81E-04    | 1.86E-04 | 2.74E-04 | 1.85E-04 |
| Low         | Rare          | MAR                | 20             | 3.43E-04       | 2.61E-04 | 2.61E-04 | 1.97E-04    | 2.90E-04    | 1.97E-04 | 1.54E-04 | 2.83E-04 |
| Low         | Rare          | MAR                | 50             | 6.53E-04       | 2.49E-04 | 2.49E-04 | 3.89E-04    | 2.20E-04    | 3.89E-04 | 2.89E-04 | 2.69E-04 |
| High        | Common        | MCAR               | 1              | 1.38E-04       | 7.35E-05 | 7.34E-05 | 2.78E-04    | 1.74E-04    | 4.46E-04 | 7.38E-04 | 5.70E-05 |
| High        | Common        | MCAR               | 5              | 3.25E-04       | 6.80E-04 | 6.80E-04 | 3.51E-04    | 4.46E-04    | 3.85E-04 | 1.17E-03 | 4.87E-04 |
| High        | Common        | MCAR               | 20             | 6.80E-04       | 6.83E-04 | 6.83E-04 | 3.86E-04    | 5.81E-04    | 1.33E-03 | 2.12E-03 | 2.10E-04 |

|      |               |      |    |          |          |          |          |          |          |          |          |
|------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| High | Common        | MCAR | 50 | 3.59E-03 | 3.00E-03 | 3.00E-03 | 2.34E-03 | 2.63E-03 | 1.54E-03 | 2.29E-03 | 2.02E-03 |
| High | Low frequency | MCAR | 1  | 3.38E-04 | 1.68E-04 | 1.68E-04 | 1.68E-04 | 1.68E-04 | 1.67E-04 | 1.67E-04 | 1.68E-04 |
| High | Low frequency | MCAR | 5  | 6.88E-04 | 1.67E-04 | 1.67E-04 | 1.67E-04 | 1.67E-04 | 1.68E-04 | 1.68E-04 | 1.67E-04 |
| High | Low frequency | MCAR | 20 | 2.00E-03 | 1.65E-03 | 1.65E-03 | 1.66E-03 | 1.67E-03 | 1.67E-03 | 1.62E-03 | 1.66E-03 |
| High | Low frequency | MCAR | 50 | 4.58E-03 | 1.65E-03 | 1.65E-03 | 1.67E-03 | 1.68E-03 | 1.64E-03 | 1.68E-03 | 1.66E-03 |
| High | Rare          | MCAR | 1  | 4.47E-05 | 8.50E-05 | 8.51E-05 | 4.89E-05 | 3.39E-05 | 8.83E-05 | 5.95E-05 | 5.30E-05 |
| High | Rare          | MCAR | 5  | 2.57E-04 | 1.19E-04 | 1.20E-04 | 2.58E-04 | 1.74E-04 | 3.27E-04 | 1.16E-04 | 1.67E-04 |
| High | Rare          | MCAR | 20 | 3.42E-04 | 4.34E-04 | 4.34E-04 | 7.74E-04 | 5.85E-04 | 5.75E-04 | 3.81E-04 | 4.05E-04 |
| High | Rare          | MCAR | 50 | 6.20E-04 | 6.76E-04 | 6.76E-04 | 8.71E-04 | 8.51E-04 | 6.83E-04 | 5.97E-04 | 7.26E-04 |
| Low  | Common        | MCAR | 1  | 3.48E-04 | 3.43E-04 | 3.43E-04 | 1.25E-04 | 1.24E-04 | 3.41E-04 | 3.46E-04 | 3.44E-04 |
| Low  | Common        | MCAR | 5  | 3.44E-04 | 3.40E-04 | 3.40E-04 | 1.14E-04 | 1.14E-04 | 3.35E-04 | 3.24E-04 | 3.42E-04 |
| Low  | Common        | MCAR | 20 | 3.67E-04 | 3.65E-04 | 3.65E-04 | 3.34E-04 | 3.67E-04 | 3.65E-04 | 3.60E-04 | 3.53E-04 |
| Low  | Common        | MCAR | 50 | 3.39E-04 | 3.28E-04 | 3.28E-04 | 1.19E-04 | 1.90E-04 | 3.46E-04 | 3.24E-04 | 3.34E-04 |
| Low  | Low frequency | MCAR | 1  | 1.83E-05 | 6.44E-06 | 6.42E-06 | 5.28E-06 | 1.51E-05 | 4.77E-05 | 6.33E-05 | 1.88E-06 |
| Low  | Low frequency | MCAR | 5  | 1.56E-04 | 9.19E-05 | 9.20E-05 | 1.02E-04 | 1.10E-04 | 1.13E-04 | 1.03E-04 | 1.20E-04 |
| Low  | Low frequency | MCAR | 20 | 6.11E-05 | 1.35E-04 | 1.35E-04 | 3.89E-05 | 5.23E-05 | 3.65E-04 | 3.59E-04 | 5.40E-05 |
| Low  | Low frequency | MCAR | 50 | 2.71E-04 | 1.99E-04 | 1.99E-04 | 2.14E-04 | 1.60E-04 | 1.35E-04 | 3.02E-04 | 1.66E-04 |
| Low  | Rare          | MCAR | 1  | 1.83E-04 | 1.83E-04 | 1.83E-04 | 1.83E-04 | 1.83E-04 | 1.79E-04 | 1.77E-04 | 1.83E-04 |
| Low  | Rare          | MCAR | 5  | 1.97E-04 | 1.82E-04 | 1.82E-04 | 2.07E-04 | 1.91E-04 | 1.79E-04 | 2.42E-04 | 1.86E-04 |
| Low  | Rare          | MCAR | 20 | 3.62E-04 | 1.81E-04 | 1.81E-04 | 2.17E-04 | 1.90E-04 | 1.98E-04 | 2.10E-04 | 2.19E-04 |
| Low  | Rare          | MCAR | 50 | 2.57E-04 | 2.01E-04 | 2.01E-04 | 2.59E-04 | 1.99E-04 | 2.91E-04 | 2.54E-04 | 2.08E-04 |
| High | Common        | MNAR | 1  | 3.38E-04 | 4.00E-04 | 4.01E-04 | 3.40E-04 | 3.26E-04 | 2.39E-04 | 9.40E-05 | 3.66E-04 |
| High | Common        | MNAR | 5  | 6.88E-04 | 1.12E-03 | 1.12E-03 | 3.38E-04 | 1.06E-03 | 1.66E-03 | 2.13E-04 | 1.20E-03 |
| High | Common        | MNAR | 20 | 2.00E-03 | 1.36E-03 | 1.36E-03 | 1.13E-03 | 2.50E-03 | 4.09E-03 | 6.06E-04 | 3.46E-03 |
| High | Common        | MNAR | 50 | 4.58E-03 | 9.59E-04 | 9.59E-04 | 2.65E-03 | 3.05E-03 | 1.03E-02 | 1.63E-03 | 8.08E-03 |
| High | Low frequency | MNAR | 1  | 3.78E-04 | 2.94E-04 | 2.94E-04 | 3.89E-04 | 3.27E-04 | 2.16E-04 | 1.71E-05 | 3.75E-04 |
| High | Low frequency | MNAR | 5  | 9.38E-04 | 5.90E-04 | 5.90E-04 | 1.09E-03 | 9.02E-04 | 7.81E-04 | 1.34E-04 | 9.84E-04 |
| High | Low frequency | MNAR | 20 | 2.24E-03 | 2.15E-03 | 2.15E-03 | 3.47E-03 | 3.17E-03 | 3.89E-03 | 3.49E-04 | 3.20E-03 |
| High | Low frequency | MNAR | 50 | 5.18E-03 | 2.63E-03 | 2.63E-03 | 6.23E-03 | 6.91E-03 | 8.85E-03 | 1.65E-03 | 8.40E-03 |
| High | Rare          | MNAR | 1  | 3.36E-05 | 6.79E-05 | 6.80E-05 | 6.77E-05 | 4.94E-05 | 3.63E-05 | 3.33E-05 | 5.42E-05 |

|             |               |      |    |          |          |          |          |          |          |          |          |
|-------------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| <b>High</b> | Rare          | MNAR | 5  | 3.41E-04 | 4.22E-04 | 4.23E-04 | 4.99E-04 | 3.16E-04 | 4.97E-04 | 5.98E-05 | 3.80E-04 |
| <b>High</b> | Rare          | MNAR | 20 | 8.37E-04 | 4.08E-04 | 4.08E-04 | 7.30E-04 | 7.84E-04 | 8.53E-04 | 1.12E-04 | 6.52E-04 |
| <b>High</b> | Rare          | MNAR | 50 | 1.23E-03 | 5.21E-04 | 5.21E-04 | 9.52E-04 | 6.10E-04 | 2.07E-03 | 4.36E-04 | 1.73E-03 |
| <b>Low</b>  | Common        | MNAR | 1  | 5.64E-05 | 7.16E-05 | 7.16E-05 | 3.44E-05 | 3.47E-05 | 2.36E-05 | 3.24E-05 | 8.15E-05 |
| <b>Low</b>  | Common        | MNAR | 5  | 2.65E-04 | 2.15E-04 | 2.15E-04 | 3.47E-04 | 3.46E-04 | 1.60E-04 | 1.11E-05 | 2.57E-04 |
| <b>Low</b>  | Common        | MNAR | 20 | 5.89E-04 | 5.45E-04 | 5.45E-04 | 3.65E-04 | 3.44E-04 | 7.79E-04 | 1.46E-04 | 6.67E-04 |
| <b>Low</b>  | Common        | MNAR | 50 | 7.68E-04 | 5.28E-04 | 5.28E-04 | 3.46E-04 | 3.67E-04 | 1.35E-03 | 2.32E-04 | 1.09E-03 |
| <b>Low</b>  | Low frequency | MNAR | 1  | 2.11E-05 | 5.28E-06 | 5.27E-06 | 5.57E-06 | 4.86E-06 | 5.57E-06 | 4.73E-07 | 5.49E-06 |
| <b>Low</b>  | Low frequency | MNAR | 5  | 8.64E-05 | 6.41E-05 | 6.41E-05 | 7.44E-05 | 7.06E-05 | 7.44E-05 | 4.24E-05 | 7.74E-05 |
| <b>Low</b>  | Low frequency | MNAR | 20 | 1.23E-04 | 1.73E-04 | 1.73E-04 | 2.08E-04 | 1.90E-04 | 2.08E-04 | 4.99E-05 | 1.76E-04 |
| <b>Low</b>  | Low frequency | MNAR | 50 | 4.13E-04 | 5.49E-04 | 5.49E-04 | 1.61E-04 | 3.60E-04 | 1.61E-04 | 8.94E-05 | 1.55E-04 |
| <b>Low</b>  | Rare          | MNAR | 1  | 3.42E-04 | 3.52E-04 | 3.52E-04 | 5.53E-06 | 4.07E-06 | 3.53E-04 | 3.53E-04 | 3.45E-04 |
| <b>Low</b>  | Rare          | MNAR | 5  | 3.69E-04 | 3.73E-04 | 3.73E-04 | 1.11E-04 | 7.71E-05 | 3.63E-04 | 3.44E-04 | 3.69E-04 |
| <b>Low</b>  | Rare          | MNAR | 20 | 4.36E-04 | 4.22E-04 | 4.22E-04 | 1.19E-04 | 7.70E-05 | 3.98E-04 | 3.47E-04 | 4.17E-04 |
| <b>Low</b>  | Rare          | MNAR | 50 | 6.78E-04 | 5.00E-04 | 5.00E-04 | 2.05E-04 | 2.94E-04 | 6.14E-04 | 3.44E-04 | 6.25E-04 |

Supplementary Table 12 The RMSE of regression coefficient between different imputation methods in 50,000 individuals and 120 phenotypes.

| Correlation | Variant       | Missing mechanisms | proportion (%) | Complete Cases | SI MICE  | MI MICE  | SI AMELIAII | MI AMELIAII | k-NN     | QRILC    | RF       |
|-------------|---------------|--------------------|----------------|----------------|----------|----------|-------------|-------------|----------|----------|----------|
| High        | Common        | MAR                | 1              | 3.19E-04       | 2.95E-04 | 2.95E-04 | 3.06E-04    | 2.45E-04    | 2.44E-04 | 2.64E-04 | 2.36E-04 |
| High        | Common        | MAR                | 5              | 8.29E-04       | 7.09E-04 | 7.09E-04 | 5.57E-04    | 4.84E-04    | 4.08E-04 | 4.61E-04 | 4.00E-04 |
| High        | Common        | MAR                | 20             | 1.81E-03       | 1.32E-03 | 1.32E-03 | 1.35E-03    | 1.10E-03    | 1.07E-03 | 1.11E-03 | 1.07E-03 |
| High        | Common        | MAR                | 50             | 2.80E-03       | 1.58E-03 | 1.58E-03 | 2.07E-03    | 1.76E-03    | 1.47E-03 | 2.14E-03 | 1.56E-03 |
| High        | Low frequency | MAR                | 1              | 2.07E-04       | 1.96E-04 | 1.96E-04 | 1.88E-04    | 1.38E-04    | 1.31E-04 | 1.54E-04 | 1.34E-04 |
| High        | Low frequency | MAR                | 5              | 5.42E-04       | 5.79E-04 | 5.79E-04 | 4.88E-04    | 3.53E-04    | 3.18E-04 | 3.82E-04 | 3.23E-04 |
| High        | Low frequency | MAR                | 20             | 1.31E-03       | 1.54E-03 | 1.54E-03 | 7.64E-04    | 8.69E-04    | 6.98E-04 | 7.30E-04 | 7.05E-04 |
| High        | Low frequency | MAR                | 50             | 2.24E-03       | 3.44E-03 | 3.44E-03 | 1.39E-03    | 1.21E-03    | 1.20E-03 | 1.34E-03 | 1.23E-03 |
| High        | Rare          | MAR                | 1              | 1.43E-04       | 1.27E-04 | 1.27E-04 | 1.06E-04    | 9.23E-05    | 8.68E-05 | 1.10E-04 | 8.08E-05 |
| High        | Rare          | MAR                | 5              | 2.96E-04       | 2.34E-04 | 2.34E-04 | 2.93E-04    | 1.85E-04    | 1.72E-04 | 2.20E-04 | 1.69E-04 |
| High        | Rare          | MAR                | 20             | 6.93E-04       | 5.20E-04 | 5.20E-04 | 5.77E-04    | 4.21E-04    | 3.91E-04 | 4.75E-04 | 3.93E-04 |
| High        | Rare          | MAR                | 50             | 1.08E-03       | 7.73E-04 | 7.73E-04 | 8.08E-04    | 7.92E-04    | 6.77E-04 | 6.68E-04 | 6.67E-04 |
| Low         | Common        | MAR                | 1              | 1.12E-04       | 8.52E-05 | 8.52E-05 | 9.29E-05    | 6.66E-05    | 6.71E-05 | 1.66E-04 | 6.02E-05 |
| Low         | Common        | MAR                | 5              | 2.14E-04       | 1.54E-04 | 1.54E-04 | 1.31E-04    | 1.30E-04    | 1.14E-04 | 1.73E-04 | 1.18E-04 |
| Low         | Common        | MAR                | 20             | 4.83E-04       | 5.20E-04 | 5.20E-04 | 3.62E-04    | 3.09E-04    | 2.92E-04 | 3.18E-04 | 2.85E-04 |
| Low         | Common        | MAR                | 50             | 8.41E-04       | 8.73E-04 | 8.73E-04 | 5.99E-04    | 5.27E-04    | 4.71E-04 | 4.54E-04 | 4.46E-04 |
| Low         | Low frequency | MAR                | 1              | 3.11E-05       | 2.37E-05 | 2.37E-05 | 1.90E-05    | 1.71E-05    | 2.11E-05 | 5.14E-04 | 1.69E-05 |
| Low         | Low frequency | MAR                | 5              | 7.30E-05       | 5.78E-05 | 5.78E-05 | 5.25E-05    | 4.70E-05    | 3.83E-05 | 5.24E-04 | 4.11E-05 |
| Low         | Low frequency | MAR                | 20             | 1.87E-04       | 1.08E-04 | 1.08E-04 | 1.29E-04    | 1.09E-04    | 1.05E-04 | 5.16E-04 | 1.03E-04 |
| Low         | Low frequency | MAR                | 50             | 3.33E-04       | 2.17E-04 | 2.17E-04 | 2.00E-04    | 1.94E-04    | 1.58E-04 | 4.95E-04 | 1.60E-04 |
| Low         | Rare          | MAR                | 1              | 2.10E-05       | 1.96E-05 | 1.96E-05 | 1.84E-05    | 1.33E-05    | 1.12E-05 | 1.26E-05 | 1.13E-05 |
| Low         | Rare          | MAR                | 5              | 5.37E-05       | 3.92E-05 | 3.92E-05 | 4.44E-05    | 3.45E-05    | 3.26E-05 | 3.43E-05 | 3.25E-05 |
| Low         | Rare          | MAR                | 20             | 1.15E-04       | 7.39E-05 | 7.39E-05 | 8.56E-05    | 5.97E-05    | 6.19E-05 | 8.04E-05 | 5.81E-05 |
| Low         | Rare          | MAR                | 50             | 2.32E-04       | 1.70E-04 | 1.70E-04 | 1.76E-04    | 1.56E-04    | 1.36E-04 | 1.22E-04 | 1.36E-04 |
| High        | Common        | MCAR               | 1              | 3.06E-04       | 2.77E-04 | 2.77E-04 | 3.09E-04    | 2.01E-04    | 2.19E-04 | 2.30E-04 | 1.98E-04 |
| High        | Common        | MCAR               | 5              | 7.63E-04       | 7.25E-04 | 7.25E-04 | 6.50E-04    | 5.50E-04    | 5.55E-04 | 6.86E-04 | 4.71E-04 |
| High        | Common        | MCAR               | 20             | 1.62E-03       | 1.20E-03 | 1.20E-03 | 1.26E-03    | 9.63E-04    | 8.85E-04 | 8.69E-04 | 8.35E-04 |

|      |               |      |    |          |          |          |          |          |          |          |          |
|------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| High | Common        | MCAR | 50 | 3.47E-03 | 2.12E-03 | 2.12E-03 | 2.38E-03 | 2.05E-03 | 1.88E-03 | 1.58E-03 | 1.79E-03 |
| High | Low frequency | MCAR | 1  | 2.08E-04 | 1.84E-04 | 1.84E-04 | 2.20E-04 | 1.34E-04 | 1.75E-04 | 2.22E-04 | 1.26E-04 |
| High | Low frequency | MCAR | 5  | 5.72E-04 | 4.98E-04 | 4.98E-04 | 4.30E-04 | 3.80E-04 | 3.56E-04 | 4.18E-04 | 3.07E-04 |
| High | Low frequency | MCAR | 20 | 1.10E-03 | 1.42E-03 | 1.42E-03 | 8.83E-04 | 7.69E-04 | 7.01E-04 | 9.95E-04 | 6.96E-04 |
| High | Low frequency | MCAR | 50 | 2.32E-03 | 3.21E-03 | 3.21E-03 | 1.57E-03 | 1.40E-03 | 1.15E-03 | 1.32E-03 | 1.23E-03 |
| High | Rare          | MCAR | 1  | 1.19E-04 | 1.00E-04 | 9.99E-05 | 1.11E-04 | 8.26E-05 | 1.13E-04 | 1.50E-04 | 7.93E-05 |
| High | Rare          | MCAR | 5  | 2.96E-04 | 2.15E-04 | 2.15E-04 | 1.99E-04 | 1.82E-04 | 1.65E-04 | 2.47E-04 | 1.59E-04 |
| High | Rare          | MCAR | 20 | 6.70E-04 | 4.63E-04 | 4.63E-04 | 5.12E-04 | 4.28E-04 | 4.09E-04 | 4.47E-04 | 3.95E-04 |
| High | Rare          | MCAR | 50 | 1.20E-03 | 7.75E-04 | 7.75E-04 | 8.87E-04 | 7.22E-04 | 6.91E-04 | 7.94E-04 | 7.94E-04 |
| Low  | Common        | MCAR | 1  | 8.51E-05 | 7.88E-05 | 7.88E-05 | 6.72E-05 | 5.41E-05 | 5.34E-05 | 8.51E-05 | 4.76E-05 |
| Low  | Common        | MCAR | 5  | 2.06E-04 | 1.79E-04 | 1.79E-04 | 1.79E-04 | 1.36E-04 | 1.33E-04 | 2.04E-04 | 1.22E-04 |
| Low  | Common        | MCAR | 20 | 4.44E-04 | 3.81E-04 | 3.81E-04 | 3.60E-04 | 2.63E-04 | 2.61E-04 | 3.17E-04 | 2.37E-04 |
| Low  | Common        | MCAR | 50 | 9.35E-04 | 7.62E-04 | 7.62E-04 | 6.37E-04 | 5.25E-04 | 4.87E-04 | 4.76E-04 | 4.76E-04 |
| Low  | Low frequency | MCAR | 1  | 3.88E-05 | 3.44E-05 | 3.44E-05 | 3.44E-05 | 2.20E-05 | 2.63E-05 | 5.11E-04 | 2.30E-05 |
| Low  | Low frequency | MCAR | 5  | 4.80E-05 | 4.26E-05 | 4.26E-05 | 3.63E-05 | 3.32E-05 | 3.29E-05 | 5.19E-04 | 3.26E-05 |
| Low  | Low frequency | MCAR | 20 | 1.88E-04 | 1.24E-04 | 1.24E-04 | 1.33E-04 | 1.22E-04 | 1.17E-04 | 5.16E-04 | 1.10E-04 |
| Low  | Low frequency | MCAR | 50 | 3.60E-04 | 2.11E-04 | 2.11E-04 | 2.59E-04 | 2.04E-04 | 1.68E-04 | 4.94E-04 | 1.73E-04 |
| Low  | Rare          | MCAR | 1  | 3.74E-06 | 1.93E-06 | 1.93E-06 | 1.90E-06 | 1.70E-06 | 2.98E-06 | 6.08E-06 | 1.83E-06 |
| Low  | Rare          | MCAR | 5  | 6.52E-05 | 4.97E-05 | 4.97E-05 | 5.54E-05 | 3.48E-05 | 3.33E-05 | 4.62E-05 | 3.25E-05 |
| Low  | Rare          | MCAR | 20 | 1.50E-04 | 1.01E-04 | 1.01E-04 | 1.20E-04 | 9.52E-05 | 8.27E-05 | 1.00E-04 | 5.81E-05 |
| Low  | Rare          | MCAR | 50 | 2.71E-04 | 1.81E-04 | 1.81E-04 | 1.93E-04 | 1.72E-04 | 1.61E-04 | 1.37E-04 | 1.36E-04 |
| High | Common        | MNAR | 1  | 2.68E-04 | 1.11E-04 | 1.11E-04 | 1.06E-04 | 1.01E-04 | 9.97E-05 | 3.42E-05 | 3.42E-05 |
| High | Common        | MNAR | 5  | 7.36E-04 | 3.14E-04 | 3.14E-04 | 3.02E-04 | 3.01E-04 | 2.87E-04 | 6.77E-05 | 6.77E-05 |
| High | Common        | MNAR | 20 | 1.79E-03 | 7.89E-04 | 7.89E-04 | 8.11E-04 | 7.92E-04 | 8.53E-04 | 2.72E-04 | 2.72E-04 |
| High | Common        | MNAR | 50 | 3.44E-03 | 1.48E-03 | 1.48E-03 | 1.43E-03 | 1.41E-03 | 1.55E-03 | 6.15E-04 | 6.15E-04 |
| High | Low frequency | MNAR | 1  | 1.88E-04 | 5.09E-05 | 5.10E-05 | 6.09E-05 | 5.30E-05 | 5.37E-05 | 1.13E-05 | 5.48E-05 |
| High | Low frequency | MNAR | 5  | 4.60E-04 | 1.31E-04 | 1.31E-04 | 1.27E-04 | 1.20E-04 | 1.19E-04 | 2.43E-05 | 1.31E-04 |
| High | Low frequency | MNAR | 20 | 1.09E-03 | 3.50E-04 | 3.50E-04 | 3.58E-04 | 3.35E-04 | 3.30E-04 | 5.13E-05 | 3.35E-04 |
| High | Low frequency | MNAR | 50 | 2.39E-03 | 7.15E-04 | 7.15E-04 | 7.23E-04 | 6.47E-04 | 6.53E-04 | 1.54E-04 | 1.54E-04 |
| High | Rare          | MNAR | 1  | 1.04E-04 | 4.59E-05 | 4.59E-05 | 1.03E-04 | 1.06E-04 | 4.22E-05 | 1.78E-05 | 1.68E-03 |

|             |               |      |    |          |          |          |          |          |          |          |          |
|-------------|---------------|------|----|----------|----------|----------|----------|----------|----------|----------|----------|
| <b>High</b> | Rare          | MNAR | 5  | 2.10E-04 | 1.12E-04 | 1.12E-04 | 2.27E-04 | 2.93E-04 | 1.21E-04 | 3.35E-05 | 1.71E-03 |
| <b>High</b> | Rare          | MNAR | 20 | 5.26E-04 | 2.47E-04 | 2.47E-04 | 4.82E-04 | 5.77E-04 | 2.78E-04 | 9.07E-05 | 1.69E-03 |
| <b>High</b> | Rare          | MNAR | 50 | 1.11E-03 | 4.73E-04 | 4.73E-04 | 1.09E-03 | 8.08E-04 | 5.03E-04 | 2.54E-04 | 1.85E-03 |
| <b>Low</b>  | Common        | MNAR | 1  | 7.22E-05 | 2.23E-05 | 2.23E-05 | 2.49E-05 | 2.34E-05 | 2.26E-05 | 5.70E-06 | 1.21E-03 |
| <b>Low</b>  | Common        | MNAR | 5  | 1.67E-04 | 5.41E-05 | 5.41E-05 | 5.42E-05 | 5.33E-05 | 5.34E-05 | 8.58E-06 | 2.32E-05 |
| <b>Low</b>  | Common        | MNAR | 20 | 4.36E-04 | 1.22E-04 | 1.22E-04 | 1.18E-04 | 1.12E-04 | 1.11E-04 | 2.07E-05 | 5.33E-05 |
| <b>Low</b>  | Common        | MNAR | 50 | 8.89E-04 | 2.39E-04 | 2.39E-04 | 2.52E-04 | 2.34E-04 | 2.38E-04 | 5.67E-05 | 2.40E-04 |
| <b>Low</b>  | Low frequency | MNAR | 1  | 3.88E-05 | 5.32E-06 | 5.32E-06 | 7.39E-06 | 9.33E-06 | 7.92E-06 | 2.86E-05 | 2.86E-05 |
| <b>Low</b>  | Low frequency | MNAR | 5  | 4.80E-05 | 2.06E-05 | 2.06E-05 | 2.81E-05 | 2.52E-05 | 2.39E-05 | 5.05E-05 | 4.55E-05 |
| <b>Low</b>  | Low frequency | MNAR | 20 | 1.88E-04 | 5.32E-06 | 5.32E-06 | 7.26E-05 | 7.13E-05 | 6.90E-05 | 1.51E-04 | 1.11E-04 |
| <b>Low</b>  | Low frequency | MNAR | 50 | 3.60E-04 | 7.38E-05 | 7.38E-05 | 1.76E-04 | 1.55E-04 | 1.66E-04 | 2.27E-04 | 1.87E-04 |
| <b>Low</b>  | Rare          | MNAR | 1  | 2.67E-06 | 7.58E-08 | 7.63E-08 | 7.96E-08 | 2.86E-07 | 5.32E-07 | 1.25E-06 | 7.06E-07 |
| <b>Low</b>  | Rare          | MNAR | 5  | 5.74E-06 | 1.47E-07 | 1.40E-07 | 1.37E-07 | 5.50E-06 | 3.05E-06 | 2.00E-06 | 3.91E-06 |
| <b>Low</b>  | Rare          | MNAR | 20 | 2.29E-05 | 3.63E-07 | 3.62E-07 | 3.25E-07 | 2.67E-05 | 1.48E-05 | 4.37E-06 | 1.78E-05 |
| <b>Low</b>  | Rare          | MNAR | 50 | 1.06E-04 | 7.67E-07 | 7.63E-07 | 6.01E-07 | 6.64E-05 | 5.12E-05 | 1.65E-05 | 5.13E-05 |

Supplementary Table 13 The RMSE of regression coefficient between different imputation methods in 500,000 individuals and three phenotypes.

| Correlation | Variant       | Missing mechanisms | proportion (%) | Complete Cases | SI MICE | MI MICE | SI AMELIAII | MI AMELIAII | k-NN     | QRILC    | RF |
|-------------|---------------|--------------------|----------------|----------------|---------|---------|-------------|-------------|----------|----------|----|
| High        | Common        | MAR                | 1              | 1.53E-05       | -       | -       | 1.29E-06    | 4.73E-06    | 3.17E-05 | 1.18E-04 | -  |
| High        | Common        | MAR                | 5              | 7.44E-05       | -       | -       | 1.41E-06    | 5.36E-05    | 8.48E-05 | 3.94E-04 | -  |
| High        | Common        | MAR                | 20             | 3.41E-04       | -       | -       | 4.76E-06    | 1.62E-04    | 2.09E-04 | 8.76E-04 | -  |
| High        | Common        | MAR                | 50             | 1.19E-03       | -       | -       | 4.97E-06    | 7.08E-04    | 2.50E-04 | 1.27E-03 | -  |
| High        | Low frequency | MAR                | 1              | 3.74E-05       | -       | -       | 1.94E-05    | 2.67E-05    | 2.12E-05 | 1.06E-04 | -  |
| High        | Low frequency | MAR                | 5              | 1.60E-04       | -       | -       | 1.88E-05    | 6.60E-05    | 9.41E-05 | 3.85E-04 | -  |
| High        | Low frequency | MAR                | 20             | 6.16E-04       | -       | -       | 1.65E-05    | 1.08E-04    | 1.83E-04 | 8.62E-04 | -  |
| High        | Low frequency | MAR                | 50             | 1.99E-03       | -       | -       | 1.95E-05    | 3.69E-04    | 2.77E-04 | 1.26E-03 | -  |
| High        | Rare          | MAR                | 1              | 3.00E-06       | -       | -       | 0.00E+00    | 4.80E-06    | 1.30E-05 | 4.51E-05 | -  |
| High        | Rare          | MAR                | 5              | 2.77E-05       | -       | -       | 5.77E-07    | 1.52E-05    | 3.00E-05 | 1.54E-04 | -  |
| High        | Rare          | MAR                | 20             | 1.27E-04       | -       | -       | 2.08E-06    | 6.15E-05    | 5.32E-05 | 3.37E-04 | -  |
| High        | Rare          | MAR                | 50             | 4.44E-04       | -       | -       | 2.83E-06    | 5.62E-05    | 8.24E-05 | 4.87E-04 | -  |
| Low         | Common        | MAR                | 1              | 2.71E-04       | -       | -       | 2.67E-04    | 2.71E-04    | 2.56E-04 | 2.22E-04 | -  |
| Low         | Common        | MAR                | 5              | 2.95E-04       | -       | -       | 2.67E-04    | 2.78E-04    | 2.28E-04 | 1.14E-04 | -  |
| Low         | Common        | MAR                | 20             | 3.94E-04       | -       | -       | 2.67E-04    | 3.28E-04    | 1.95E-04 | 6.92E-05 | -  |
| Low         | Common        | MAR                | 50             | 7.10E-04       | -       | -       | 2.67E-04    | 3.21E-04    | 1.58E-04 | 2.20E-04 | -  |
| Low         | Low frequency | MAR                | 1              | 2.20E-06       | -       | -       | 1.00E-07    | 1.03E-06    | 3.54E-06 | 1.34E-05 | -  |
| Low         | Low frequency | MAR                | 5              | 1.06E-05       | -       | -       | 1.29E-07    | 7.06E-06    | 1.17E-05 | 4.40E-05 | -  |
| Low         | Low frequency | MAR                | 20             | 4.85E-05       | -       | -       | 4.43E-07    | 1.33E-05    | 2.18E-05 | 9.68E-05 | -  |
| Low         | Low frequency | MAR                | 50             | 1.14E-04       | -       | -       | 5.74E-07    | 4.49E-05    | 3.23E-05 | 1.40E-04 | -  |
| Low         | Rare          | MAR                | 1              | 7.92E-05       | -       | -       | 7.69E-05    | 7.71E-05    | 7.69E-05 | 6.34E-05 | -  |
| Low         | Rare          | MAR                | 5              | 8.76E-05       | -       | -       | 7.69E-05    | 7.78E-05    | 7.69E-05 | 3.30E-05 | -  |
| Low         | Rare          | MAR                | 20             | 1.26E-04       | -       | -       | 7.74E-05    | 9.36E-05    | 7.74E-05 | 1.99E-05 | -  |
| Low         | Rare          | MAR                | 50             | 1.91E-04       | -       | -       | 7.70E-05    | 1.39E-04    | 7.70E-05 | 6.30E-05 | -  |
| High        | Common        | MCAR               | 1              | 1.40E-05       | -       | -       | 8.16E-07    | 4.73E-06    | 3.45E-05 | 1.12E-04 | -  |
| High        | Common        | MCAR               | 5              | 7.73E-05       | -       | -       | 2.65E-06    | 1.92E-05    | 1.03E-04 | 3.78E-04 | -  |
| High        | Common        | MCAR               | 20             | 3.42E-04       | -       | -       | 4.51E-06    | 1.66E-04    | 2.11E-04 | 8.47E-04 | -  |

|      |               |      |    |          |   |   |          |          |          |          |   |
|------|---------------|------|----|----------|---|---|----------|----------|----------|----------|---|
| High | Common        | MCAR | 50 | 1.19E-03 | - | - | 6.45E-06 | 2.59E-04 | 3.43E-04 | 1.25E-03 | - |
| High | Low frequency | MCAR | 1  | 3.74E-05 | - | - | 1.89E-05 | 2.09E-05 | 2.22E-05 | 9.99E-05 | - |
| High | Low frequency | MCAR | 5  | 1.60E-04 | - | - | 1.92E-05 | 5.41E-05 | 7.52E-05 | 3.77E-04 | - |
| High | Low frequency | MCAR | 20 | 6.16E-04 | - | - | 2.13E-05 | 1.22E-04 | 1.56E-04 | 8.39E-04 | - |
| High | Low frequency | MCAR | 50 | 1.99E-03 | - | - | 1.41E-05 | 3.72E-04 | 2.07E-04 | 1.24E-03 | - |
| High | Rare          | MCAR | 1  | 6.68E-06 | - | - | 0.00E+00 | 4.80E-06 | 1.17E-05 | 4.45E-05 | - |
| High | Rare          | MCAR | 5  | 2.63E-05 | - | - | 8.16E-07 | 1.08E-05 | 4.25E-05 | 1.47E-04 | - |
| High | Rare          | MCAR | 20 | 1.34E-04 | - | - | 2.45E-06 | 4.80E-06 | 7.80E-05 | 3.25E-04 | - |
| High | Rare          | MCAR | 50 | 4.52E-04 | - | - | 4.00E-06 | 3.38E-04 | 1.03E-04 | 4.83E-04 | - |
| Low  | Common        | MCAR | 1  | 2.71E-04 | - | - | 2.67E-04 | 2.71E-04 | 2.56E-04 | 2.24E-04 | - |
| Low  | Common        | MCAR | 5  | 3.03E-04 | - | - | 2.66E-04 | 2.75E-04 | 2.30E-04 | 1.23E-04 | - |
| Low  | Common        | MCAR | 20 | 3.93E-04 | - | - | 2.66E-04 | 2.71E-04 | 1.99E-04 | 6.11E-05 | - |
| Low  | Common        | MCAR | 50 | 7.33E-04 | - | - | 2.68E-04 | 5.96E-04 | 1.62E-04 | 2.14E-04 | - |
| Low  | Low frequency | MCAR | 1  | 2.13E-06 | - | - | 5.77E-08 | 3.56E-07 | 3.54E-06 | 1.28E-05 | - |
| Low  | Low frequency | MCAR | 5  | 6.80E-06 | - | - | 1.91E-07 | 5.65E-06 | 1.00E-05 | 4.10E-05 | - |
| Low  | Low frequency | MCAR | 20 | 2.82E-05 | - | - | 3.87E-07 | 1.98E-05 | 2.51E-05 | 9.41E-05 | - |
| Low  | Low frequency | MCAR | 50 | 1.31E-04 | - | - | 1.63E-07 | 2.47E-05 | 2.89E-05 | 1.38E-04 | - |
| Low  | Rare          | MCAR | 1  | 8.02E-05 | - | - | 7.70E-05 | 7.71E-05 | 7.33E-05 | 6.45E-05 | - |
| Low  | Rare          | MCAR | 5  | 8.67E-05 | - | - | 7.68E-05 | 8.16E-05 | 6.85E-05 | 3.53E-05 | - |
| Low  | Rare          | MCAR | 20 | 1.11E-04 | - | - | 7.68E-05 | 9.28E-05 | 5.58E-05 | 1.75E-05 | - |
| Low  | Rare          | MCAR | 50 | 1.99E-04 | - | - | 7.76E-05 | 1.10E-04 | 5.39E-05 | 6.15E-05 | - |
| High | Common        | MNAR | 1  | 3.74E-05 | - | - | 1.58E-05 | 2.52E-05 | 8.29E-06 | 8.16E-07 | - |
| High | Common        | MNAR | 5  | 1.60E-04 | - | - | 1.60E-05 | 9.44E-05 | 1.92E-05 | 1.37E-05 | - |
| High | Common        | MNAR | 20 | 6.16E-04 | - | - | 5.87E-05 | 3.97E-04 | 3.03E-05 | 1.42E-04 | - |
| High | Common        | MNAR | 50 | 1.99E-03 | - | - | 1.80E-04 | 1.26E-03 | 6.16E-05 | 6.17E-04 | - |
| High | Low frequency | MNAR | 1  | 3.60E-05 | - | - | 2.07E-05 | 2.81E-05 | 1.41E-05 | 1.15E-06 | - |
| High | Low frequency | MNAR | 5  | 1.58E-04 | - | - | 8.53E-05 | 1.11E-04 | 7.58E-05 | 3.87E-06 | - |
| High | Low frequency | MNAR | 20 | 5.79E-04 | - | - | 2.48E-04 | 3.65E-04 | 3.06E-04 | 2.40E-05 | - |
| High | Low frequency | MNAR | 50 | 1.80E-03 | - | - | 4.90E-04 | 7.97E-04 | 7.74E-04 | 1.55E-04 | - |
| High | Rare          | MNAR | 1  | 1.70E-05 | - | - | 7.05E-06 | 4.80E-06 | 4.36E-06 | 5.77E-07 | - |

|             |               |      |    |          |   |   |          |          |          |          |   |
|-------------|---------------|------|----|----------|---|---|----------|----------|----------|----------|---|
| <b>High</b> | Rare          | MNAR | 5  | 6.47E-05 | - | - | 2.53E-05 | 3.44E-05 | 5.07E-06 | 5.35E-06 | - |
| <b>High</b> | Rare          | MNAR | 20 | 2.52E-04 | - | - | 7.07E-05 | 1.04E-04 | 1.65E-05 | 5.24E-05 | - |
| <b>High</b> | Rare          | MNAR | 50 | 8.19E-04 | - | - | 1.45E-04 | 3.96E-04 | 2.30E-05 | 2.43E-04 | - |
| <b>Low</b>  | Common        | MNAR | 1  | 1.47E-05 | - | - | 2.67E-04 | 2.70E-04 | 5.84E-06 | 3.16E-07 | - |
| <b>Low</b>  | Common        | MNAR | 5  | 6.13E-05 | - | - | 2.65E-04 | 2.80E-04 | 2.41E-05 | 1.30E-06 | - |
| <b>Low</b>  | Common        | MNAR | 20 | 2.31E-04 | - | - | 2.57E-04 | 2.99E-04 | 1.18E-04 | 7.24E-06 | - |
| <b>Low</b>  | Common        | MNAR | 50 | 7.44E-04 | - | - | 2.35E-04 | 3.09E-04 | 2.99E-04 | 5.83E-05 | - |
| <b>Low</b>  | Low frequency | MNAR | 1  | 5.43E-06 | - | - | 1.80E-06 | 1.82E-06 | 1.80E-06 | 1.29E-07 | - |
| <b>Low</b>  | Low frequency | MNAR | 5  | 1.94E-05 | - | - | 6.48E-06 | 9.29E-06 | 6.48E-06 | 1.61E-06 | - |
| <b>Low</b>  | Low frequency | MNAR | 20 | 7.36E-05 | - | - | 2.02E-05 | 2.55E-05 | 2.02E-05 | 1.49E-05 | - |
| <b>Low</b>  | Low frequency | MNAR | 50 | 2.05E-04 | - | - | 3.98E-05 | 9.97E-05 | 3.98E-05 | 6.90E-05 | - |
| <b>Low</b>  | Rare          | MNAR | 1  | 4.50E-06 | - | - | 2.28E-06 | 2.49E-06 | 2.11E-06 | 1.29E-06 | - |
| <b>Low</b>  | Rare          | MNAR | 5  | 1.94E-05 | - | - | 9.27E-06 | 1.22E-05 | 8.97E-06 | 1.29E-06 | - |
| <b>Low</b>  | Rare          | MNAR | 20 | 6.76E-05 | - | - | 2.79E-05 | 3.95E-05 | 3.26E-05 | 2.34E-06 | - |
| <b>Low</b>  | Rare          | MNAR | 50 | 2.04E-04 | - | - | 5.50E-05 | 9.18E-05 | 8.55E-05 | 1.69E-05 | - |

Supplementary Table 14 The RMSE of regression coefficient between different imputation methods in 500,000 individuals and 30 phenotypes.

| Correlation | Variant       | Missing mechanisms | proportion (%) | Complete Cases | SI MICE | MI MICE | SI AMELIAII | MI AMELIAII | k-NN     | QRILC    | RF |
|-------------|---------------|--------------------|----------------|----------------|---------|---------|-------------|-------------|----------|----------|----|
| High        | Common        | MAR                | 1              | 1.76E-05       | -       | -       | 8.56E-07    | 1.52E-05    | 3.15E-06 | 3.93E-05 | -  |
| High        | Common        | MAR                | 5              | 8.44E-05       | -       | -       | 7.82E-07    | 1.10E-05    | 2.44E-05 | 4.19E-05 | -  |
| High        | Common        | MAR                | 20             | 3.91E-04       | -       | -       | 1.34E-06    | 5.62E-05    | 1.70E-04 | 9.30E-05 | -  |
| High        | Common        | MAR                | 50             | 1.13E-03       | -       | -       | 1.82E-06    | 7.39E-05    | 5.72E-04 | 5.06E-04 | -  |
| High        | Low frequency | MAR                | 1              | 1.26E-05       | -       | -       | 6.88E-07    | 8.28E-06    | 3.35E-06 | 1.82E-05 | -  |
| High        | Low frequency | MAR                | 5              | 6.07E-05       | -       | -       | 1.50E-06    | 3.73E-05    | 2.79E-05 | 1.67E-05 | -  |
| High        | Low frequency | MAR                | 20             | 2.79E-04       | -       | -       | 3.11E-06    | 2.74E-04    | 1.31E-04 | 2.06E-05 | -  |
| High        | Low frequency | MAR                | 50             | 8.09E-04       | -       | -       | 3.99E-06    | 4.44E-04    | 4.07E-04 | 1.33E-04 | -  |
| High        | Rare          | MAR                | 1              | 3.83E-06       | -       | -       | 5.05E-07    | 1.01E-05    | 1.41E-06 | 1.45E-05 | -  |
| High        | Rare          | MAR                | 5              | 3.16E-05       | -       | -       | 6.34E-07    | 2.71E-05    | 1.01E-05 | 1.50E-05 | -  |
| High        | Rare          | MAR                | 20             | 1.47E-04       | -       | -       | 1.37E-06    | 8.40E-05    | 6.55E-05 | 3.39E-05 | -  |
| High        | Rare          | MAR                | 50             | 4.31E-04       | -       | -       | 3.22E-06    | 3.11E-04    | 2.20E-04 | 1.91E-04 | -  |
| Low         | Common        | MAR                | 1              | 2.64E-06       | -       | -       | 2.07E-07    | 7.58E-06    | 2.05E-06 | 6.88E-06 | -  |
| Low         | Common        | MAR                | 5              | 2.25E-05       | -       | -       | 4.89E-07    | 3.65E-05    | 9.06E-06 | 6.58E-06 | -  |
| Low         | Common        | MAR                | 20             | 1.05E-04       | -       | -       | 1.33E-06    | 5.87E-05    | 5.08E-05 | 7.87E-06 | -  |
| Low         | Common        | MAR                | 50             | 3.08E-04       | -       | -       | 1.94E-06    | 2.11E-04    | 1.59E-04 | 5.21E-05 | -  |
| Low         | Low frequency | MAR                | 1              | 2.66E-06       | -       | -       | 1.05E-07    | 9.48E-07    | 4.75E-07 | 4.18E-06 | -  |
| Low         | Low frequency | MAR                | 5              | 1.24E-05       | -       | -       | 1.83E-07    | 5.46E-06    | 2.89E-06 | 4.57E-06 | -  |
| Low         | Low frequency | MAR                | 20             | 5.67E-05       | -       | -       | 5.21E-07    | 1.79E-05    | 1.93E-05 | 9.98E-06 | -  |
| Low         | Low frequency | MAR                | 50             | 1.18E-04       | -       | -       | 1.05E-06    | 1.09E-04    | 6.35E-05 | 5.56E-05 | -  |
| Low         | Rare          | MAR                | 1              | 4.53E-07       | -       | -       | 1.03E-04    | 1.02E-04    | 1.02E-04 | 1.04E-04 | -  |
| Low         | Rare          | MAR                | 5              | 1.22E-05       | -       | -       | 1.03E-04    | 1.01E-04    | 1.01E-04 | 1.04E-04 | -  |
| Low         | Rare          | MAR                | 20             | 4.15E-05       | -       | -       | 1.03E-04    | 9.38E-05    | 9.55E-05 | 1.03E-04 | -  |
| Low         | Rare          | MAR                | 50             | 1.12E-04       | -       | -       | 1.03E-04    | 8.77E-05    | 8.57E-05 | 1.14E-04 | -  |
| High        | Common        | MCAR               | 1              | 1.71E-05       | -       | -       | 7.53E-07    | 1.24E-05    | 1.47E-06 | 5.92E-05 | -  |
| High        | Common        | MCAR               | 5              | 8.31E-05       | -       | -       | 7.97E-07    | 1.17E-05    | 1.14E-05 | 1.02E-04 | -  |
| High        | Common        | MCAR               | 20             | 3.92E-04       | -       | -       | 2.07E-06    | 4.18E-05    | 6.47E-05 | 1.23E-04 | -  |

|      |               |      |    |          |   |   |          |          |          |          |   |
|------|---------------|------|----|----------|---|---|----------|----------|----------|----------|---|
| High | Common        | MCAR | 50 | 1.36E-03 | - | - | 2.18E-06 | 5.03E-05 | 2.77E-04 | 2.75E-04 | - |
| High | Low frequency | MCAR | 1  | 1.20E-05 | - | - | 5.92E-07 | 4.17E-06 | 2.25E-06 | 2.50E-05 | - |
| High | Low frequency | MCAR | 5  | 5.75E-05 | - | - | 1.19E-06 | 3.54E-05 | 1.67E-05 | 4.61E-05 | - |
| High | Low frequency | MCAR | 20 | 2.78E-04 | - | - | 2.52E-06 | 1.31E-04 | 7.80E-05 | 4.33E-05 | - |
| High | Low frequency | MCAR | 50 | 9.61E-04 | - | - | 5.86E-06 | 3.60E-04 | 2.32E-04 | 6.41E-05 | - |
| High | Rare          | MCAR | 1  | 5.09E-06 | - | - | 4.18E-07 | 5.76E-06 | 6.69E-07 | 2.26E-05 | - |
| High | Rare          | MCAR | 5  | 3.36E-05 | - | - | 7.13E-07 | 2.37E-05 | 4.82E-06 | 3.86E-05 | - |
| High | Rare          | MCAR | 20 | 1.51E-04 | - | - | 1.64E-06 | 8.93E-05 | 2.55E-05 | 4.61E-05 | - |
| High | Rare          | MCAR | 50 | 5.12E-04 | - | - | 1.96E-06 | 2.29E-04 | 1.02E-04 | 1.06E-04 | - |
| Low  | Common        | MCAR | 1  | 4.36E-06 | - | - | 2.99E-07 | 3.08E-06 | 1.03E-06 | 9.20E-06 | - |
| Low  | Common        | MCAR | 5  | 2.33E-05 | - | - | 5.04E-07 | 1.52E-05 | 6.07E-06 | 1.83E-05 | - |
| Low  | Common        | MCAR | 20 | 1.02E-04 | - | - | 1.16E-06 | 6.90E-05 | 2.97E-05 | 1.73E-05 | - |
| Low  | Common        | MCAR | 50 | 3.83E-04 | - | - | 1.45E-06 | 1.36E-04 | 8.76E-05 | 2.51E-05 | - |
| Low  | Low frequency | MCAR | 1  | 1.53E-06 | - | - | 1.28E-07 | 1.22E-06 | 2.57E-07 | 6.68E-06 | - |
| Low  | Low frequency | MCAR | 5  | 8.97E-06 | - | - | 2.24E-07 | 4.64E-06 | 1.06E-06 | 1.12E-05 | - |
| Low  | Low frequency | MCAR | 20 | 4.21E-05 | - | - | 5.17E-07 | 2.51E-05 | 7.77E-06 | 1.34E-05 | - |
| Low  | Low frequency | MCAR | 50 | 1.59E-04 | - | - | 6.66E-07 | 4.80E-05 | 3.06E-05 | 3.04E-05 | - |
| Low  | Rare          | MCAR | 1  | 4.53E-07 | - | - | 1.08E-07 | 3.28E-07 | 1.92E-07 | 6.68E-06 | - |
| Low  | Rare          | MCAR | 5  | 1.22E-05 | - | - | 2.40E-07 | 4.39E-06 | 1.35E-06 | 1.12E-05 | - |
| Low  | Rare          | MCAR | 20 | 4.15E-05 | - | - | 4.96E-07 | 2.22E-05 | 7.24E-06 | 1.35E-05 | - |
| Low  | Rare          | MCAR | 50 | 1.12E-04 | - | - | 7.74E-07 | 4.96E-05 | 3.04E-05 | 3.04E-05 | - |
| High | Common        | MNAR | 1  | 1.71E-05 | - | - | 7.53E-07 | 1.24E-05 | 1.47E-06 | 5.92E-05 | - |
| High | Common        | MNAR | 5  | 8.31E-05 | - | - | 7.97E-07 | 1.17E-05 | 1.14E-05 | 1.02E-04 | - |
| High | Common        | MNAR | 20 | 3.92E-04 | - | - | 2.07E-06 | 4.18E-05 | 6.47E-05 | 1.23E-04 | - |
| High | Common        | MNAR | 50 | 1.36E-03 | - | - | 2.18E-06 | 5.03E-05 | 2.77E-04 | 2.75E-04 | - |
| High | Low frequency | MNAR | 1  | 1.20E-05 | - | - | 5.92E-07 | 4.17E-06 | 2.25E-06 | 2.50E-05 | - |
| High | Low frequency | MNAR | 5  | 5.75E-05 | - | - | 1.19E-06 | 3.54E-05 | 1.67E-05 | 4.61E-05 | - |
| High | Low frequency | MNAR | 20 | 2.78E-04 | - | - | 2.52E-06 | 1.31E-04 | 7.80E-05 | 4.33E-05 | - |
| High | Low frequency | MNAR | 50 | 9.61E-04 | - | - | 5.86E-06 | 3.60E-04 | 2.32E-04 | 6.41E-05 | - |
| High | Rare          | MNAR | 1  | 5.09E-06 | - | - | 4.18E-07 | 5.76E-06 | 6.69E-07 | 2.26E-05 | - |

|             |               |      |    |          |   |   |          |          |          |          |   |
|-------------|---------------|------|----|----------|---|---|----------|----------|----------|----------|---|
| <b>High</b> | Rare          | MNAR | 5  | 3.36E-05 | - | - | 7.13E-07 | 2.37E-05 | 4.82E-06 | 3.86E-05 | - |
| <b>High</b> | Rare          | MNAR | 20 | 1.51E-04 | - | - | 1.64E-06 | 8.93E-05 | 2.55E-05 | 4.61E-05 | - |
| <b>High</b> | Rare          | MNAR | 50 | 5.12E-04 | - | - | 1.96E-06 | 2.29E-04 | 1.02E-04 | 1.06E-04 | - |
| <b>Low</b>  | Common        | MNAR | 1  | 4.36E-06 | - | - | 2.99E-07 | 3.08E-06 | 1.03E-06 | 9.20E-06 | - |
| <b>Low</b>  | Common        | MNAR | 5  | 2.33E-05 | - | - | 5.04E-07 | 1.52E-05 | 6.07E-06 | 1.83E-05 | - |
| <b>Low</b>  | Common        | MNAR | 20 | 1.02E-04 | - | - | 1.16E-06 | 6.90E-05 | 2.97E-05 | 1.73E-05 | - |
| <b>Low</b>  | Common        | MNAR | 50 | 3.83E-04 | - | - | 1.45E-06 | 1.36E-04 | 8.76E-05 | 2.51E-05 | - |
| <b>Low</b>  | Low frequency | MNAR | 1  | 1.53E-06 | - | - | 1.28E-07 | 1.22E-06 | 2.57E-07 | 6.68E-06 | - |
| <b>Low</b>  | Low frequency | MNAR | 5  | 8.97E-06 | - | - | 2.24E-07 | 4.64E-06 | 1.06E-06 | 1.12E-05 | - |
| <b>Low</b>  | Low frequency | MNAR | 20 | 4.21E-05 | - | - | 5.17E-07 | 2.51E-05 | 7.77E-06 | 1.34E-05 | - |
| <b>Low</b>  | Low frequency | MNAR | 50 | 1.59E-04 | - | - | 6.66E-07 | 4.80E-05 | 3.06E-05 | 3.04E-05 | - |
| <b>Low</b>  | Rare          | MNAR | 1  | 4.53E-07 | - | - | 1.08E-07 | 3.28E-07 | 1.92E-07 | 6.68E-06 | - |
| <b>Low</b>  | Rare          | MNAR | 5  | 1.22E-05 | - | - | 2.40E-07 | 4.39E-06 | 1.35E-06 | 1.12E-05 | - |
| <b>Low</b>  | Rare          | MNAR | 20 | 4.15E-05 | - | - | 4.96E-07 | 2.22E-05 | 7.24E-06 | 1.35E-05 | - |
| <b>Low</b>  | Rare          | MNAR | 50 | 1.12E-04 | - | - | 7.74E-07 | 4.96E-05 | 3.04E-05 | 3.04E-05 | - |

Supplementary Table 15 The RMSE of regression coefficient between different imputation methods in 500,000 individuals and 120 phenotypes.

| Correlation | Variant       | Missing mechanisms | proportion (%) | Complete Cases | SI MICE | MI MICE | SI AMELIAII | MI AMELIAII | k-NN     | QRILC    | RF |
|-------------|---------------|--------------------|----------------|----------------|---------|---------|-------------|-------------|----------|----------|----|
| High        | Common        | MAR                | 1              | 1.80E-05       | -       | -       | 9.79E-07    | 1.45E-05    | 4.26E-06 | 9.20E-06 | -  |
| High        | Common        | MAR                | 5              | 8.60E-05       | -       | -       | 1.75E-06    | 9.23E-05    | 3.83E-05 | 9.84E-06 | -  |
| High        | Common        | MAR                | 20             | 3.97E-04       | -       | -       | 4.76E-06    | 3.72E-04    | 2.11E-04 | 4.04E-05 | -  |
| High        | Common        | MAR                | 50             | 8.96E-06       | -       | -       | 7.07E-06    | 9.50E-04    | 6.31E-04 | 4.68E-04 | -  |
| High        | Low frequency | MAR                | 1              | 1.28E-05       | -       | -       | 6.83E-07    | 9.33E-06    | 5.41E-06 | 4.65E-06 | -  |
| High        | Low frequency | MAR                | 5              | 6.07E-05       | -       | -       | 1.35E-06    | 5.13E-05    | 3.06E-05 | 1.53E-05 | -  |
| High        | Low frequency | MAR                | 20             | 2.81E-04       | -       | -       | 2.98E-06    | 2.32E-04    | 1.50E-04 | 5.03E-05 | -  |
| High        | Low frequency | MAR                | 50             | 6.35E-06       | -       | -       | 4.74E-06    | 6.45E-04    | 4.51E-04 | 1.04E-04 | -  |
| High        | Rare          | MAR                | 1              | 3.78E-06       | -       | -       | 4.83E-07    | 7.73E-06    | 1.99E-06 | 3.50E-06 | -  |
| High        | Rare          | MAR                | 5              | 3.21E-05       | -       | -       | 8.32E-07    | 2.33E-05    | 1.49E-05 | 3.92E-06 | -  |
| High        | Rare          | MAR                | 20             | 1.49E-04       | -       | -       | 1.44E-06    | 1.24E-04    | 8.12E-05 | 1.53E-05 | -  |
| High        | Rare          | MAR                | 50             | 1.88E-06       | -       | -       | 3.06E-06    | 2.77E-04    | 2.46E-04 | 1.81E-04 | -  |
| Low         | Common        | MAR                | 1              | 2.74E-06       | -       | -       | 2.38E-07    | 4.52E-06    | 2.17E-06 | 1.68E-06 | -  |
| Low         | Common        | MAR                | 5              | 2.30E-05       | -       | -       | 5.00E-07    | 2.14E-05    | 1.17E-05 | 6.14E-06 | -  |
| Low         | Common        | MAR                | 20             | 1.07E-04       | -       | -       | 9.94E-07    | 7.63E-05    | 5.80E-05 | 1.96E-05 | -  |
| Low         | Common        | MAR                | 50             | 1.37E-06       | -       | -       | 2.07E-06    | 2.12E-04    | 1.74E-04 | 3.97E-05 | -  |
| Low         | Low frequency | MAR                | 1              | 2.65E-06       | -       | -       | 1.23E-07    | 9.48E-07    | 5.99E-07 | 3.38E-05 | -  |
| Low         | Low frequency | MAR                | 5              | 1.26E-05       | -       | -       | 2.30E-07    | 6.58E-06    | 4.25E-06 | 3.38E-05 | -  |
| Low         | Low frequency | MAR                | 20             | 5.77E-05       | -       | -       | 5.32E-07    | 2.55E-05    | 2.35E-05 | 3.40E-05 | -  |
| Low         | Low frequency | MAR                | 50             | 1.32E-06       | -       | -       | 1.18E-06    | 9.69E-05    | 7.09E-05 | 6.16E-05 | -  |
| Low         | Rare          | MAR                | 1              | 1.90E-06       | -       | -       | 8.47E-08    | 6.68E-07    | 6.28E-07 | 4.84E-07 | -  |
| Low         | Rare          | MAR                | 5              | 9.02E-06       | -       | -       | 1.38E-07    | 4.62E-06    | 3.43E-06 | 1.73E-06 | -  |
| Low         | Rare          | MAR                | 20             | 4.11E-05       | -       | -       | 3.38E-07    | 1.66E-05    | 1.68E-05 | 5.67E-06 | -  |
| Low         | Rare          | MAR                | 50             | 9.36E-07       | -       | -       | 5.03E-07    | 7.12E-05    | 5.03E-05 | 1.14E-05 | -  |
| High        | Common        | MCAR               | 1              | 1.80E-05       | -       | -       | 9.13E-07    | 1.10E-05    | 3.57E-06 | 2.54E-05 | -  |
| High        | Common        | MCAR               | 5              | 8.59E-05       | -       | -       | 2.36E-06    | 6.45E-05    | 2.93E-05 | 3.36E-05 | -  |
| High        | Common        | MCAR               | 20             | 3.96E-04       | -       | -       | 4.23E-06    | 3.38E-04    | 1.76E-04 | 5.05E-05 | -  |

|      |               |      |    |          |   |   |          |          |          |          |   |
|------|---------------|------|----|----------|---|---|----------|----------|----------|----------|---|
| High | Common        | MCAR | 50 | 8.99E-06 | - | - | 6.96E-06 | 9.07E-04 | 6.46E-04 | 5.12E-04 | - |
| High | Low frequency | MCAR | 1  | 1.25E-05 | - | - | 7.01E-07 | 1.01E-05 | 4.55E-06 | 1.11E-05 | - |
| High | Low frequency | MCAR | 5  | 5.99E-05 | - | - | 1.51E-06 | 4.58E-05 | 2.69E-05 | 1.85E-05 | - |
| High | Low frequency | MCAR | 20 | 2.75E-04 | - | - | 3.41E-06 | 2.07E-04 | 1.31E-04 | 3.85E-05 | - |
| High | Low frequency | MCAR | 50 | 6.20E-06 | - | - | 4.89E-06 | 5.56E-04 | 4.64E-04 | 1.52E-04 | - |
| High | Rare          | MCAR | 1  | 7.18E-06 | - | - | 4.28E-07 | 4.79E-06 | 1.45E-06 | 9.66E-06 | - |
| High | Rare          | MCAR | 5  | 3.74E-05 | - | - | 8.42E-07 | 2.54E-05 | 1.08E-05 | 1.29E-05 | - |
| High | Rare          | MCAR | 20 | 1.44E-04 | - | - | 1.54E-06 | 1.02E-04 | 6.58E-05 | 1.89E-05 | - |
| High | Rare          | MCAR | 50 | 3.56E-06 | - | - | 2.90E-06 | 3.82E-04 | 2.48E-04 | 1.96E-04 | - |
| Low  | Common        | MCAR | 1  | 4.75E-06 | - | - | 2.37E-07 | 3.06E-06 | 1.81E-06 | 4.19E-06 | - |
| Low  | Common        | MCAR | 5  | 2.14E-05 | - | - | 5.85E-07 | 1.87E-05 | 1.07E-05 | 6.90E-06 | - |
| Low  | Common        | MCAR | 20 | 1.10E-04 | - | - | 1.06E-06 | 7.57E-05 | 5.09E-05 | 1.50E-05 | - |
| Low  | Common        | MCAR | 50 | 2.39E-06 | - | - | 2.36E-06 | 2.29E-04 | 1.83E-04 | 5.76E-05 | - |
| Low  | Low frequency | MCAR | 1  | 1.57E-06 | - | - | 1.15E-07 | 1.73E-06 | 3.71E-07 | 3.39E-05 | - |
| Low  | Low frequency | MCAR | 5  | 1.61E-05 | - | - | 1.82E-07 | 2.03E-06 | 3.20E-06 | 3.40E-05 | - |
| Low  | Low frequency | MCAR | 20 | 4.14E-05 | - | - | 3.68E-07 | 3.02E-05 | 1.89E-05 | 3.42E-05 | - |
| Low  | Low frequency | MCAR | 50 | 7.68E-07 | - | - | 8.36E-07 | 9.28E-05 | 7.14E-05 | 6.60E-05 | - |
| Low  | Rare          | MCAR | 1  | 2.67E-06 | - | - | 7.96E-08 | 2.86E-07 | 5.32E-07 | 1.25E-06 | - |
| Low  | Rare          | MCAR | 5  | 5.74E-06 | - | - | 1.37E-07 | 5.50E-06 | 3.05E-06 | 2.00E-06 | - |
| Low  | Rare          | MCAR | 20 | 2.29E-05 | - | - | 3.25E-07 | 2.67E-05 | 1.48E-05 | 4.37E-06 | - |
| Low  | Rare          | MCAR | 50 | 1.33E-06 | - | - | 6.01E-07 | 6.64E-05 | 5.12E-05 | 1.65E-05 | - |
| High | Common        | MNAR | 1  | 2.05E-05 | - | - | 1.85E-06 | 3.72E-06 | 2.10E-06 | 1.14E-06 | - |
| High | Common        | MNAR | 5  | 1.03E-04 | - | - | 9.55E-06 | 1.23E-05 | 1.11E-05 | 1.65E-06 | - |
| High | Common        | MNAR | 20 | 4.71E-04 | - | - | 5.10E-05 | 7.54E-05 | 6.04E-05 | 7.33E-06 | - |
| High | Common        | MNAR | 50 | 1.02E-05 | - | - | 1.85E-04 | 3.10E-04 | 2.36E-04 | 5.91E-05 | - |
| High | Low frequency | MNAR | 1  | 1.73E-05 | - | - | 1.29E-06 | 3.08E-06 | 1.41E-06 | 6.06E-07 | - |
| High | Low frequency | MNAR | 5  | 8.20E-05 | - | - | 5.42E-06 | 6.62E-06 | 6.02E-06 | 1.19E-06 | - |
| High | Low frequency | MNAR | 20 | 3.57E-04 | - | - | 2.46E-05 | 2.82E-05 | 2.85E-05 | 1.48E-06 | - |
| High | Low frequency | MNAR | 50 | 8.55E-06 | - | - | 8.31E-05 | 1.06E-04 | 1.03E-04 | 3.14E-06 | - |
| High | Rare          | MNAR | 1  | 9.45E-06 | - | - | 4.83E-07 | 4.83E-07 | 9.92E-07 | 6.19E-07 | - |

|             |               |      |    |          |  |  |          |          |          |          |  |
|-------------|---------------|------|----|----------|--|--|----------|----------|----------|----------|--|
| <b>High</b> | Rare          | MNAR | 5  | 5.34E-05 |  |  | 1.05E-06 | 8.32E-07 | 5.21E-06 | 6.06E-07 |  |
| <b>High</b> | Rare          | MNAR | 20 | 2.09E-04 |  |  | 3.47E-06 | 1.44E-06 | 2.47E-05 | 3.24E-06 |  |
| <b>High</b> | Rare          | MNAR | 50 | 4.69E-06 |  |  | 6.61E-06 | 3.06E-06 | 9.15E-05 | 2.31E-05 |  |
| <b>Low</b>  | Common        | MNAR | 1  | 6.70E-06 |  |  | 5.08E-07 | 6.56E-07 | 5.44E-07 | 2.05E-07 |  |
| <b>Low</b>  | Common        | MNAR | 5  | 3.50E-05 |  |  | 2.17E-06 | 2.40E-06 | 2.40E-06 | 4.55E-07 |  |
| <b>Low</b>  | Common        | MNAR | 20 | 1.55E-04 |  |  | 9.57E-06 | 1.10E-05 | 1.11E-05 | 4.82E-07 |  |
| <b>Low</b>  | Common        | MNAR | 50 | 3.39E-06 |  |  | 3.24E-05 | 3.93E-05 | 3.99E-05 | 1.20E-06 |  |
| <b>Low</b>  | Low frequency | MNAR | 1  | 1.57E-06 |  |  | 2.20E-07 | 3.71E-07 | 2.54E-07 | 1.07E-06 |  |
| <b>Low</b>  | Low frequency | MNAR | 5  | 1.61E-05 |  |  | 1.13E-06 | 1.35E-06 | 1.34E-06 | 1.08E-06 |  |
| <b>Low</b>  | Low frequency | MNAR | 20 | 4.14E-05 |  |  | 5.86E-06 | 8.28E-06 | 7.00E-06 | 4.21E-06 |  |
| <b>Low</b>  | Low frequency | MNAR | 50 | 7.68E-07 |  |  | 2.08E-05 | 3.29E-05 | 2.66E-05 | 4.45E-05 |  |
| <b>Low</b>  | Rare          | MNAR | 1  | 2.67E-06 |  |  | 7.96E-08 | 2.86E-07 | 5.32E-07 | 1.25E-06 |  |
| <b>Low</b>  | Rare          | MNAR | 5  | 5.74E-06 |  |  | 1.37E-07 | 5.50E-06 | 3.05E-06 | 2.00E-06 |  |
| <b>Low</b>  | Rare          | MNAR | 20 | 2.29E-05 |  |  | 3.25E-07 | 2.67E-05 | 1.48E-05 | 4.37E-06 |  |
| <b>Low</b>  | Rare          | MNAR | 50 | 1.33E-06 |  |  | 6.01E-07 | 6.64E-05 | 5.12E-05 | 1.65E-05 |  |