Development of bioinformatic tools and application of novel statistical methods in genome-wide analysis

van der Most, Peter Johannes

Chapter 1

Introduction

The Genetics of Complex Traits

In the field of genetic epidemiology, a complex trait is defined as a trait with a heritable component that does not follow a Mendelian inheritance pattern. Complex traits are influenced by multiple and sometimes interacting genetic and/or environmental factors. This characteristic makes it hard, if not impossible, to disentangle the genetics behind complex traits with the traditional linkage-based approach of gene mapping, which studies the cosegregation of genetic markers and disease within family pedigrees.

Finding genes for complex traits only became feasible after the introduction of the genome-wide association study (GWAS). In short, a GWAS is a scan for thousands or even millions of genetic variants that are located all over the genome, to test whether they are associated with a phenotype of interest (where the phenotype can be anything from diabetes or body weight to educational level or median income). However, a GWAS does not actually scan the entire genome. Instead, it takes advantage of the principle of linkage disequilibrium (LD) to skim-read the genome, thus drastically reducing the number of tests (and the amount of computer time and memory) required to analyse the genome. LD is the phenomenon that, at population level, two genetic loci on the same chromosome are correlated with one another. When they are in close proximity, LD is likely to be stronger because the farther away they are from each other, the more likely it is that, at some point in the family tree, a genetic recombination occurred that severed the link between the loci [1]. This means that a single genetic variant can be used to *tag* the surrounding region of DNA, as any other nearby variant is likely to be in LD with the tagging variant. Consequently, if a genetic variant that is associated with the trait is not genotyped itself, but is tagged by a nearby variant, the association with the trait will also emerge from this nearby tagging variant. It has been estimated that a selection of 450,000 human single nucleotide polymorphisms (SNPs, the most commonly used type of variant, which is a single-base-pair substitution) is sufficient to tag the majority of known common SNPs (> 10,000,000) in the European population [2], where common means that the minor allele of the SNP has a frequency of at least 5%.

GWASs only became feasible as a result of three scientific advances. Firstly, the Human Genome Project [3, 4] provided a standard sequence of the human DNA. Secondly, the HapMap project [2] identified individual differences in this sequence through mapping common SNPs and determined the LD structure between those SNPs in 270 individuals from three ethnic populations, thus allowing us to determine which SNPs are most useful as tagging variants. These two developments culminated in the final necessary step: the manufacturing of dense genotyping SNP arrays that could rapidly and accurately (and relatively inexpensively) capture the large number of common variants in the entire genome. By combining these technologies with the use of large biobanks (that is, a data repository containing many biological samples for use in research), GWASs became possible [1].

More recently, a variation on the GWAS methodology was developed to study the relation between genome-wide DNA methylation and complex traits. Methylation of CpG sites (CpG means a cytosine nucleotide that is followed by a guanine nucleotide) is the most commonly studied epigenetic mechanism. Unlike in GWAS, where a SNP can have only one of two possible outcomes, the methylation of a CpG site is a

continuous outcome indicating the proportion of cells for which the CpG site is methylated. Beyond that, the methodology of an epigenome-wide association study (EWAS) is similar to that of a GWAS. It uses array technology to determine methylation levels at known CpG sites throughout the genome (early arrays covered 27,000 sites, newer ones cover 450,000 or even 850,000 CpGs), which is then correlated with the phenotype of interest [5].

Development of the (Epi)Genome-Wide Association Study

The past decade has yielded an explosion of GWAS findings that shows no sign of stopping. Starting with two SNPs for the first GWAS on age-related macular degeneration in 2005 [6], the number of SNP-trait association identified by GWAS nowadays increased to 33,811 (unique SNP-trait associations reported in 2,866 publications in the GWAS catalog [7] as of 14/4/2017). Similar developments can be expected for the EWASs. However, this is still a relatively new field, albeit one with an exponentially growing output.

The first GWASs used relatively simple analytical methods, as these are easier to program and take up less computer time and resources. The large amount of both input and output involved in the average GWAS (~2.5 million SNPs for a HapMap imputed dataset, and >15 million SNPs for a 1000 Genomes imputed dataset [8, 9]) also emphasises brute force over sophistication. Furthermore, there is the drive to analyse ever larger samples, as the increased statistical power will allow us to find variants with ever smaller genetic effect sizes. This drive has also led to a proliferation of meta-analyses of GWASs (meta-GWASs), which increase the statistical power by combining the GWAS results of multiple cohorts for a single outcome [10-13].

However, upscaling the sample is not the only option to find more associated genetic variants. New statistical methods and software tools open up more sophisticated avenues to analyse genome-wide data. As such it is important to carefully consider the various options in order to obtain more or better information from a genome-wide dataset.

Aims of the thesis

Hence, in this thesis, we sought to develop novel ways to obtain more and better-quality results from genome-wide data, and apply them. We approached this problem from several angles:

- The development of software tools for the quality control of GWAS and EWAS data
- The application of software tools to determine heritability explained by genetic variants
- The development of methods to employ survival analysis in GWAS
- The application of novel methods for imputation and analysis of genetic data

The individual chapters, and how they fit into the overall thesis, are listed in the below table.

Chapters	Software tools		Statistical methods	
	Development	Application	Development	Application
2: QCGWAS	√			
3: QCEWAS	√			
4: lodGWAS	√		√	
5: Missing Heritability of Complex Traits		√	√	√
6: Heritability of Neuroticism		√		√
7: Kidney Function				√
8: Cannabis - age at first use			√	√

Outline

Section A: Development of Software Tools for Analysis and Quality Control of GWAS and EWAS

The sheer size of the average GWAS results file makes a manual quality control (QC) impractical. However, there is always the potential for an analysis to contain errors (ranging from using a bad model or the wrong unit of measurement to issues of file formatting), so it is important to establish that the results data are valid, of high quality, and comparable between cohorts. This is particularly important in the context of meta-analysis, where the researcher will combine data from multiple sources. Although other software packages for automated quality control of GWAS files existed, we felt that these were insufficiently thorough and not well-documented. We also wanted a software package that could prepare the files for meta-analysis. Therefore, we developed the software package QCGWAS, which automates the QC, generates a detailed log and various graphs to allow a thorough quality check, and can also compare the results of multiple files to check if they are compatible for a meta-analysis. This package is described in **Chapter 2**.

Similar problems existed for conducting a meta-analysis of EWAS. As far as we were aware, no tool was available for running a QC over and preparing EWAS files for meta-analysis. Therefore, using QCGWAS as a basis, we developed the software package QCEWAS to address this. QCEWAS is described in **Chapter 3**.

In **Chapter 4**, we present a software package to perform a genome-wide analysis of biomarkers with a limit-of-detection (LOD) restriction. The LOD is the level below (or above) which the assay of the biomarker cannot provide accurate results. Current GWAS methods did not adequately handle values outside of the LOD. To solve this problem, we developed the software package lodGWAS, which uses survival analysis techniques to accurately model these values in a GWAS analysis.

Section B: Application of Software Tools and Novel Statistical Methods

For many complex traits there is a gap between the total heritability and the genetic variance explained by known genetic variants. This is known as the *missing heritability* problem [14]. In recent years, many novel SNPs associations with complex traits have been identified by meta-analyses carried out in large consortia. In order to determine whether the missing heritability is decreasing, we replicated all known SNPs associated with 32 complex traits in five disease categories (anthropometric, cardiovascular & renal, metabolic, haematology & inflammation, and lung function) in the Lifelines Cohort Study (n≈13,300, [15, 16]). Subsequently, these SNPs were combined into a genetic risk score for each trait and used to determine the amount of heritability explained by the known genetic variants. To determine the remaining missing heritability, we did not rely on estimates of total heritability from family and twin studies, but employed a new software method, genomic-relatedness-matrix restricted maximum likelihood (GREML) in the genome-wide complex trait analysis (GCTA) software package, that uses GWAS data of unrelated individuals to estimate the proportion of variance of a trait that can be explained by all common SNPs (as opposed to *known* variants only) [17]. By comparing this with the variance explained by the genetic risk score, we can determine how much of the common-SNP heritability is still missing. This is described in **Chapter 5**. In this chapter we also present a new method to control for overlap between discovery and replication samples. Our genetic risk scores are based on the results of large meta analyses, which sometimes included the Lifelines cohort. If Lifelines cohort data were used both for constructing the genetic risk score and validation of that score, this would cause the estimates of the variance explained by the genetic risk score to be inflated [18]. By "subtracting" the Lifelines effect from our genetic risk score, we were able to prevent this.

In **Chapter 6**, we use GREML to investigate how much of neuroticism's heritability can be explained by common SNPs in an Estonian and a Dutch cohort. Neuroticism is a complex trait that is substantially heritable, but for which few genetic variants have yet been identified. With GREML, we could not only determine the common-SNP heritability of neuroticism, but also test whether the heritability differed between the Estonian and Dutch cohorts. Furthermore, we endeavoured to obtain more precise estimates of the common-SNP heritability than previous studies by using the same method to score neuroticism in both cohorts (as opposed to harmonizing scores of multiple methods between cohorts), looking at the subscales of neuroticism as well as the overall scale, using residual scores (corrected for the effect of other facet/domain scores), and investigating whether neuroticism as reported by a knowledgeable other person (a spouse, relative, or close friend) is more or less heritable than self-reported neuroticism.

In **Chapter 7**, we describe a meta-GWAS of the estimated glomerular filtration rate calculated from serum creatinine (eGFRcrea). Previous efforts for this phenotype identified 53 associated SNPs, but as with other complex traits, these SNPs account for only a fraction of the total heritability of eGFRcrea [11, 19-23]. In addition, it only analysed SNPs imputed using the HapMap project reference panel [2]. In this analysis, we used a newer reference panel, the 1000 Genomes Project, which is based on more samples, includes more variants, and provides better tagging, particularly of low-frequency variants [8]. In addition, polygenic risk score analysis was performed to determine the maximal variance that can be explained by SNPs, by testing multiple genetic risk scores composed of SNPs meeting different significance thresholds [24]. In this way, we hoped to

determine whether there are additional genetic variants that explain part of the variance, but did not reach genome-wide significance in the GWAS due to limited statistical power to detect small effects.

In **Chapter 8**, we describe a meta-GWAS of the age of onset of cannabis usage. Cannabis usage is associated with a variety of adverse outcomes. This occurs more frequently with an early onset of cannabis use [25]. Young onset of cannabis use is also associated with increased odds of abuse of other substances [26-29]. However, previous studies into the heritability of age-of-initiation of cannabis use have yielded contradictory results. One possible explanation may be the different ways of expressing the phenotype. Does one analyse it as a continuous variable, or as an ordinal one (e.g. "young", "middle aged", "old", "never")? The problem with the first solution is that it cannot include never-users, while the latter assumes that never-users won't become users afterwards. In this chapter we solve the problem of analysing it as a continuous variable by employing survival analysis. By treating the age of the never-user as a censored data point (i.e., at that point he or she was not a user) it can be accurately modelled in the analyses, which increases sample size and statistical power.

References

1. Visscher PM, Brown MA, McCarthy MI, Yang J: Five Years of GWAS Discovery. *Am J Hum Genet* 2012; **90:** 7-24.
2. Altshuler D, Brooks L, Chakravarti A *et al:* A haplotype map of the human genome. *Nature* 2005; **437:** 1299-1320.
3. Lander E, Int Human Genome Sequencing Consortium, Linton L *et al:* Initial sequencing and analysis of the human genome. *Nature* 2001; **409:** 860-921.
4. Venter J, Adams M, Myers E *et al:* The sequence of the human genome. *Science* 2001; **291:** 1304-1351.
5. Rakyan VK, Down TA, Balding DJ, Beck S: Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics* 2011; **12:** 529-541.
6. Klein R, Zeiss C, Chew E *et al:* Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308:** 385-389.
7. Welter D, MacArthur J, Morales J *et al:* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 2014; **42:** D1001.
8. Altshuler DM, Durbin RM, Abecasis GR *et al:* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491:** 56-65.
9. Altshuler DM, Durbin RM, Abecasis GR *et al:* A global reference for human genetic variation. *Nature* 2015; **526:** 68-74.
10. Ehret GB, Munroe PB, Rice KM *et al:* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 2011; **478:** 103-109.
11. Pattaro C, Koettgen A, Teumer A *et al:* Genome-Wide Association and Functional Follow-Up Reveals New Loci for Kidney Function. *Plos Genetics* 2012; **8:** e1002584.
12. Psaty BM, O'Donnell CJ, Gudnason V *et al:* Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Design of Prospective Meta-Analyses of Genome-Wide Association Studies From 5 Cohorts. *Circulation-Cardiovascular Genetics* 2009; **2:** 73-80.
13. Speliotes EK, Willer CJ, Berndt SI *et al:* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010; **42:** 937-948.
14. Manolio TA, Collins FS, Cox NJ *et al:* Finding the missing heritability of complex diseases. *Nature* 2009; **461:** 747-753.
15. Scholtens S, Smidt N, Swertz MA *et al:* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol* 2015; **44:** 1172-1180.
16. Stolk RP, Rosmalen JGM, Postma DS *et al:* Universal risk factors for multifactorial diseases - LifeLines: a three-generation population-based study. *Eur J Epidemiol* 2008; **23:** 67-74.
17. Yang J, Lee SH, Goddard ME, Visscher PM: GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am J Hum Genet* 2011; **88:** 76-82.
18. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM: Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* 2013; **14:** 507-515.
19. Chambers JC, Zhang W, Lord GM *et al:* Genetic loci influencing kidney function and chronic kidney disease. *Nat Genet* 2010; **42:** 373-375.
20. Chasman DI, Fuchsberger C, Pattaro C *et al:* Integration of genome-wide association studies with biological knowledge identifies six novel genes related to kidney function. *Hum Mol Genet* 2012; **21:** 5329-5343.
21. Koettgen A, Glazer NL, Dehghan A *et al:* Multiple loci associated with indices of renal function and chronic kidney disease. *Nat Genet* 2009; **41:** 712-717.
22. Koettgen A, Pattaro C, Boeger CA *et al:* New loci associated with kidney function and chronic kidney disease. *Nat Genet* 2010; **42:** 376-384.
23. Pattaro C, Teumer A, Gorski M *et al:* Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat Commun* 2016; **7:** 10023.
24. Purcell SM, Wray NR, Stone JL *et al:* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460:** 748-752.
25. Hall W, Degenhardt L: Adverse health effects of non-medical cannabis use. *Lancet* 2009; **374:** 1383-1391.
26. Agrawal A, Grant JD, Waldron M *et al:* Risk for initiation of substance use as a function of age of onset of cigarette, alcohol and cannabis use: Findings in a Midwestern female twin cohort. *Prev Med* 2006; **43:** 125-128.
27. Kokkevi A, Gabhainn SN, Spyropoulou M, Risk Behav Focus Grp HBSC: Early initiation of cannabis use: A cross-national European perspective. *Journal of Adolescent Health* 2006; **39:** 712-719.
28. Lynskey MT, Agrawal A, Henders A, Nelson EC, Madden PAF, Martin NG: An Australian Twin Study of Cannabis and Other Illicit Drug Use and Misuse, and Other Psychopathology. *Twin Research and Human Genetics* 2012; **15:** 631-641.
29. Lynskey M, Vink J, Boomsma D: Early onset cannabis use and progression to other drug use in a sample of Dutch twins. *Behav Genet* 2006; **36:** 195-200.