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

van der Most, Peter Johannes

Summary

In the current thesis, we looked at ways to improve the quality, rather than quantity, of the results of genome-wide association analyses. We sought to develop and apply novel ways to obtain more and higher-quality results. In the first part of this thesis, we describe three software tools for (epi-)genome-wide association analyses. The second part concerns the application of novel software tools and statistical methods to analyse genome-wide datasets.

Chapter 2 and 3 described QCGWAS and QCEWAS, R packages that automate the quality control of genome- and epigenome-wide association study (GWAS and EWAS) results files, respectively. The advantages of these packages are that they perform a thorough and comprehensive quality control (in less time than it would take to do even a simple manual check); are flexible, allowing for a great deal of user customization; and can generate ready-made files for a meta-analysis, as well as performing various checks to ensure that the files are compatible.

Chapter 4 described lodGWAS, which is an R package for genome-wide analysis of phenotypes such as biomarkers with a limit of detection (LOD). Although there are alternative methods for analysing biomarkers with a LOD in a GWAS, these generally rely on fixing or excluding measurements beyond the LOD. By using a survival analysis, and treating these measurements as censored data, lodGWAS can accurately model the biomarker, without the loss of information associated with the other methods. lodGWAS is flexible and easy to use, providing a simple and elegant way for GWAS analysis of biomarkers with a LOD.

In **chapter 5** we used genetic risk scores (GRS) and genomic restricted maximum likelihood (GREML) methods to estimate the amount of common SNP heritability accounted for by the known genetic markers, and the amount that is still missing, for 32 complex disease traits in the Lifelines cohort. The majority of previously-associated single-nucleotide polymorphisms (SNPs) (median = 75%) were significantly associated with their respective traits. Both the GRSs and the broad-sense common-SNP heritability estimates were significant for all traits, with weighted GRSs generally explaining more variance than unweighted ones. However, the variance explained by the weighted GRSs accounted for only 10.7%, on average, of the common-SNP heritabilities of the 32 complex disease traits. Dominance effects of common SNPs were small and not significant for any of the traits. In terms of clinical relevance, this demonstrated that the GRSs of these complex disease traits are not yet accurate enough to be used for personalized predictive medicine. In terms of methodology, an important innovation presented in this chapter was the tool that we developed to remove the effect from a specific cohort from an effect size calculated in a meta-analysis that included results from that same cohort. Using the effect size without this correction in the GRSs would have caused inflated results, as we would be validating the effect size in (part of) the same data that had been used to estimate it in the first place. Our tool allowed us to use corrected effect sizes and apply the GRSs to Lifelines without overestimating the proportion of trait variance explained.

In **chapter 6**, we used GREML to investigate in the Estonian Biobank (EGCUT) and Dutch Lifelines cohorts how much of neuroticism's heritability can be explained by common SNPs. We looked not only at the overall neuroticism score, but also at neuroticism facet scores, residual scores, and self-reported neuroticism vs.

neuroticism reported by a spouse, relative or friend. In addition, we tested for differences between the two cohorts through a gene-environment interaction analysis, where the cohort was considered as the environmental factor. In EGCUT, heritability estimates for self-reported neuroticism (both domain and facet) ranged from 0.08 to 0.20, but only the impulsiveness facet ($h^2 = 0.20$) was significant. In Lifelines all estimates (0.07-0.16) were significant except for the impulsiveness facet ($h^2 = 0.03$). Heritability estimates for residual scores or neuroticism reported by a spouse, relative or friend (both of which were only available in EGCUT) were not significant. The lack of significant findings can be at least partially attributed to the small sample size of EGCUT, which rendered it somewhat underpowered for the analysis. Finally, we used a new technique available in GREML to carry out a gene-environment interaction analysis to determine whether there were differences between the two cohorts. No significant interactions were found except for the impulsivity facet, where the inclusion of the interaction reduced the heritability estimate common to both populations to near zero ($h^2 = 0.001$, n.s.).

Chapter 7 described a large-scale meta-GWAS of kidney function, quantified as the estimated glomerular filtration rate (eGFR), using a new genome reference panel, the 1000 Genomes project, for imputation. The study replicated 39 previously identified loci, and found 10 novel loci, six of which were not tagged by previous imputation panels. This demonstrates the benefits of using a more detailed reference panel for imputation. In addition, we calculated polygenic risk scores to determine whether there are additional associated genetic variants that could not be identified in our GWAS due to limited statistical power to detect small effects. The polygenic risk score analysis was applied in the TRAILS and NESDA cohorts and explained a maximum of 2.2% of variance in eGFR.

In **chapter 8**, we combined a traditional survival analysis with GWAS, and performed a meta-analysis of such GWASs to find genetic markers related to the age of initiation of cannabis use. A single locus (containing five significant SNPs) was found in the Calcium-transporting ATPase (*ATP2C2*) gene. Gene-based tests also identified *ATP2C2*, as well as *ECT2L* and *RAD51B*. Unfortunately, the lead SNP in *ATP2C2* could not be replicated in a small replication sample (n = 4,478), but it remained significant in the combined analysis. Tests of SNP-based heritability and polygenic risk scores did not yield significant results.

Finally, **chapter 9** provided a general discussion of the novel tools and methods described in this thesis. Taken together, these methods allow researchers to run more sophisticated, rather than simply larger, genome-wide analyses.

In future, we expect that there will be two main challenges for the field of Genetic Epidemiology. The first is to find the remaining missing heritability. Before the availability of whole exome or genome sequencing data becomes widespread this will likely be primarily accomplished by means of array based GWASs with larger sample sizes and imputed against more detailed reference genome panels. The second challenge is how to effectively leverage the vast amounts of omics data already collected to determine the functional mechanisms of the identified genetic association effects. Unfortunately, the sheer size of the genome-wide datasets is a problem in itself, in terms of both collecting and analysing the data. However, here too there are solutions, for example the use of genome-wide summary statistics as a substitute for individual-level data when investigating downstream effects. These methods are another example of employing more sophisticated methods, rather than larger datasets, to do genetic-epidemiologic research.