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

van der Most, Peter Johannes

Chapter 9

Discussion and future perspectives

Methods for genome-wide association analyses have generally emphasised brute force over sophistication: including more variants and using larger samples sizes in order to identify more and more genetic variants associated with a complex trait. However, more sophisticated analysis of genome-wide data could also help in improving the statistical power to detect those variants. In this thesis we developed strategies to improve the quality of the methodology of genome-wide association analysis. We set out to develop and apply novel tools to obtain more and higher-quality results from genome-wide data. We approached this problem from four perspectives:

1. Better quality control (QC) of genome-wide and epigenome-wide association studies (GWAS and EWAS) data by developing software tools (**Chapters 2 and 3**)
2. Use of survival analysis in GWAS by developing new methods and tools (**Chapter 4 and 8**)
3. Estimation of SNP-based heritability, and the extent to which it is explained by known genetic variants, by application of software tools (**Chapters 5, 6, 7 and 8**)
4. Better imputation and analysis of genome-wide data by applying novel methods (**Chapter 7 and 8**)

The Importance of Quality Control

The first step towards obtaining better results is to ensure the quality of the data you already have. Without QC, errors will propagate through the analysis pipeline, influencing results at every step. Although GWAS and EWAS have become familiar methods, they are still complex, multistep analyses with multiple occasions for introducing errors. In this respect, the size of an (epi-)genome-wide analysis is a downside: the sheer number of results, most of which are insignificant and therefore not meaningful, can hide systematic errors. That is why automated QC is essential: only a computer can evaluate the entire dataset, and do so consistently and thoroughly within a reasonable time-frame [1,2]. For this reason we developed the software packages QCGWAS and QCEWAS (**Chapters 2 and 3**).

Another advantage of our automated quality-control software is that it allows you to compare data from multiple sources, since nowadays most GWAS/EWAS studies are meta-analyses that combine many GWAS/EWAS results from different centres. Before the meta-analysis, it is particularly important to run a QC, to ensure that datasets from different sources are actually comparable. If one centre accidentally used an incorrect unit or transformation of the phenotype (for example if they forgot to convert the lipid concentration from milligram/decilitre to millimol/liter, or to log-transform the C-reactive protein levels before analysis), the resulting summary statistics will systematically differ from that of other centres [3]. However, the overall distribution of the effect sizes does reflect the scale of the phenotype, so an error in the phenotype can be detected by comparing the range of effect sizes (preferably filtered for high quality, to exclude unreliable results) between results files [1,2]. Note that studies with larger sample sizes, and hence more power, will yield more accurate effect estimations and thus present a distribution of effect sizes that will be visibly tighter. This is not a sign of incomparable phenotypes; it is the expected outcome. To help distinguish the effect of increased power from real deviations in phenotype scale, both QCGWAS and QCEWAS will sort the results on sample size when plotting the graph comparing the effect-size distribution.

As a third feature, QCGWAS and QCEWAS can remove bad entries and standardize the formatting of the output files, facilitating the meta-analysis. They also check the distribution of quality parameters, like Hardy-Weinberg equilibrium p-values and call rates, to see if these make sense. If they do not, this may indicate a formatting error, and columns may have been switched. Another sanity check is comparing the reported p-value with a p-value calculated from the effect size and standard error. These sometimes do not correspond, either because of formatting, rounding or due to an apparent bug in the analysis software [1,2].

Fourthly, QCGWAS checks marker alleles and allele frequencies against a reference file. Although, ideally speaking, all studies in a meta-analysis should use the same version of the human genome assembly and dbSNP release for imputation, this is not always the case [3]. There are sometimes differences between SNPs of different builds (not the least in their strand orientation) and the allele check ensures that the correct alleles are compared with one another. In case of ambiguous SNPs (SNPs with either A/T or C/G as alleles, which will be mirrored on the opposite strand, making it impossible to determine the strand orientation), QCGWAS will compare the allele frequencies of ambiguous SNPs with a reference, and can correct those that are obviously inverted. Other ambiguous SNPs can be excluded. Fortunately, the wide-spread adoption of the 1000 Genomes reference panel, which standardizes the strand used for all variants, should reduce the need for this [4].

The most importantly QC step is to check the results for p-value inflation via QQ plots and calculating a genomic control lambda value [3,5]. In brief: because the vast majority of genetic markers are not expected to be associated with the phenotype in question, their p-values should follow an uniform distribution. Both the lambda and the QQ plot show how well the p-values actually follow this distribution. A handful of SNPs being more significant than the distribution is expected: these presumably are the truly associated markers, and therefore their p-values do not follow an uniform distribution. However, if large proportions of the dataset are more significant than expected, it indicates a modelling error. This is called p-value inflation, or oversignificance, and can originate from various sources, such as non-independent samples, technical bias or uncorrected population stratification [3]. Insufficient QC can in extreme cases even lead to spurious results being published. A well-known case is the study by Sebastiani and colleagues, which was retracted after receiving severe criticism for, amongst other things, genotyping cases and controls on different platforms. The retraction attributed the positive findings to technical issues and inadequate quality control rather than true associations [6,7].

Finally, we attempted to make the QC procedures user-friendly, and wrote an extensive documentation, so that the packages are usable even by people with only a cursory understanding of R and QC.

However, there are also a few issues with both packages, QCGWAS in particular. The first is that they assume a continuous phenotype, meaning that the result estimates take the form of an effect size and standard error. In case of analysis of dichotomous phenotypes, which usually report odds ratios and confidence intervals, these will need to be transformed to a continuously distributed effect size in order to be processed by QCGWAS. This is easily accomplished by natural-log transformation, but we intend to automate this in future versions of the software.

Secondly, the packages do not automatically correct problems (besides removing bad data); and offer no options for doing so. This was a design choice, intended to keep the packages simple and easy to use, and to avoid inadvertent corrections. However, this means that when a problem is encountered, the users either need to correct it themselves, or have to report it to the original analyst. One suggestion, made by Prins [8], was to include a positive-control option if available for the trait under study. QCGWAS could compare the results of the GWAS file under QC to previous results reported for example on the GWAS catalog website [9] or from previous (consistent and replicated) candidate gene association studies. This would serve as a quick consistency check and may detect a multitude of problems.

Another suggestion by Prins [8], was to improve the strand-alignment checking of ambiguous SNPs. He suggested to check the alignment of such a SNP by comparing the results to nearby, non-ambiguous SNPs that are in LD. This is a worthwhile suggestion, but as said before, made less urgent by the wide-spread adoption of the 1000 Genomes reference panel [4] for imputation

However, the 1000 Genomes reference panel forms the fourth and most important issue for the current version of QCGWAS: the package was designed for HapMap-imputed GWASs, which means that it automatically rejects non-SNP markers. GWAS using 1000 Genomes imputed data also contain insertions and deletions, which are thus excluded by default. An updated version of QCGWAS is needed to fix this deficiency, but that leads to a further problem. Whereas HapMap-based results files typically contain 2.5 to 3 million variants and take between 5 and 15 minutes and between 2 and 3 GB of RAM to run a QC on; 1000 Genomes based results contain upwards of 8-30 million variants, taking over 40 minutes and 20 GB of RAM on a 64-bit PC. This is a consequence of the core R program: it handles large datasets by loading them into the memory in their entirety. The R language is also a scripting language, and therefore executes code more slowly than a programming language. However, it is possible to incorporate code written in the C++ programming language into R, and use parallel computing techniques [10], to speed up the package. In addition, there are R packages available that improve the memory usage for big datasets [11,12].

In conclusion, automated QC of GWAS and EWAS results files saves time both during the QC itself and by finding errors before they become a problem. It may even detect errors that would otherwise have gone unnoticed and reduced the quality of the research results. The QCGWAS and QCEWAS packages provide an easy-to-use method to run an automated and thorough QC. QCGWAS has been successfully applied by other researchers [13,14]. For future versions of the packages, we intend to incorporate the simpler handling of case-control results and, for QCGWAS, data generated using more recent genome references. QCEWAS is already capable of processing the latest methylation array (Illumina MethylationEPIC BeadChip) assaying 850,000 methylation sites, but it will benefit from faster processing and optimized memory usage. Finally, we will consider adding options that compare the reported effect sizes with those expected from earlier GWAS or EWAS, and more sophisticated strand checking through LD.

The Use of Survival Analysis in Genome-Wide Association Studies

Survival analysis is the name given to a category of statistical tests that deal with "censored data". As the name implies, the typical use of survival analysis is to determine whether factor X influences the survival time of patients with condition Y. However, survival analysis can be used for any time-to-event value: not just survival but also, for example, patient recovery or the onset of menopause. The difficulty with such phenotypes is that the starting or ending point of the time-to-event, or both, may be unknown. For example: we may not know when the disease started; only when it was diagnosed. Similarly, if the patient survived for the entire duration of the study, the end point is unknown. In such a case, we cannot simply use the end date of the study as the end point - the patient may have survived for many years after. This type of problem is known as censored data - the actual value is not known, but the interval in which it lies is. In the first example, the actual point lies somewhere before the date at which the disease was diagnosed, while in the second one the actual point is after the termination of the study. The benefits of using survival analysis for appropriate phenotypes in genome-wide analyses are straightforward: it allows for more accurate modelling of the phenotype, while at the same time conserving the sample size [15-17].In **Chapter 8** we used survival analysis on a genome-wide level in order to associate genetic variants with age of initiation of cannabis use. Our analysis found one significant SNP and three significant genes, none of which had been identified by an earlier case-control analysis by the same consortium [18]. This suggests that the age of onset is affected by different factors than cannabis usage. However, this inference is only tentative as we were unable to replicate our findings, presumably due to an underpowered replication stage.

A previous analysis in the same cohort had already investigated genetic effects on cannabis usage, as a dichotomous phenotype [18]. However, a survival analysis generally offers more statistical power than a logistic model [19], presumably because it is a more accurate model of the phenotype. Strictly speaking, though, these two models analyse different phenotypes: a case-control analysis investigates the difference between those with and without the event at the end of the study, while a survival analysis investigates the time to the event. A case-control analysis also assumes that a control individual (e.g. a person who did not use cannabis) will never experience the event, while in fact he/she might do so after the end of the study. With survival analysis, the data of control individuals are treated as censored, allowing for the event to take place after the end of the study. This, however, relates to an often-overlooked caveat of survival analysis, namely that all unaffected subjects are treated similarly. In reality, these unaffected subjects may consist of two distinct groups: those that haven't experienced the event yet, and those that will never experience the event. In a traditional survival analysis involving a terminal disease, this is obviously not an issue. However, in our analysis it is conceivable that there exists a qualitative difference between those that will never use cannabis, and those that started using it later in life (i.e. after the data were collected). There is a way to address this problem by combining a logistic and survival analysis (the *cure survival model*), but the added computational burden makes this unsuitable for genome-wide analyses [20]. Thus, survival analysis is a viable method for GWAS analyses.

Another application of survival analysis is for biomarkers that are measured with an assay that has a limit of detection (LOD). Such data may also be regarded as censored data [17]. We developed the R package lodGWAS to run a GWAS analysis for biomarkers with a LOD using survival analysis (**Chapter 4**). Although methods exists to deal with measurements outside of LOD [21,22], these generally involve either excluding samples outside of detection limit, or replacing them with a constant value (e.g. the actual LOD) [23]. Measurements outside of LOD are inaccurate, but not invalid (i.e. they still indicate whether the effect is large or small, just not how large or small). Removing these values, or replacing them with another inaccurate value, causes a loss of information, in particular when a substantial portion of the samples fall outside of LOD. However, as mentioned above, values outside of LOD can be viewed as censored data: the actual value is unknown but it falls within a known interval. Hence, they can be analysed using survival analysis [17]. This results in more accurate modelling of the data while conserving the sample size, and can be used for any variable subject to LOD [15-17]. Furthermore the above-mentioned caveat of the two distinct types of censored data in survival analysis does not apply, as we know that all samples have a real value for this biomarker. As such, our package lodGWAS is a valuable addition to the available GWAS tools.

Methods to Estimate Heritability: Genetic Risk Scores, Polygenic Risk Scores and GREML-GCTA

Heritability is defined as the proportion of phenotypic variation that is due to genetics. In genetic epidemiology, often only the additive component of the total heritability is taken into account. This is called the "narrow-sense heritability" as opposed to the "broad-sense heritability", which also includes non-additive components such as dominance [24]. In this thesis, we employed three methods to determine the part of the heritability that can be explained by genetic markers. Each method estimates a different subset of the total heritability. The methods are genetic risk scores (**Chapter 5**), polygenic risk scores (**Chapters 7** and **8**) and genomic restricted maximum likelihood (GREML, **Chapters 5** and **6**).

Genetic risk scores (GRS) are calculated using significantly associated genetic markers known from previous GWASs [25,26]. A GRS could be described as (an estimate of) the combined genetic effect on that trait, but including only genes that are genome-wide significantly associated. A polygenic risk score (PRS) is a GRS that also includes less-significant genetic markers. The hypothesis of a PRS is that truly associated genetic markers are still hidden in the noise, but have not been detected due to limited power. A PRS on a trait could be described as the combined genetic effect of all genes that *may* be associated. Finally, the GREML-method from the Genome-wide Complex Trait Analysis (GCTA) software package does not look at the effect of individual SNPs, but instead uses genotyped or imputed genetic markers to calculate relatedness (i.e. the genetic similarity) between unrelated individuals, and then correlates the genetic similarity with phenotypic similarity [27,28]. It tests the hypothesis that individuals who are genetically more closely related will also be phenotypically more similar. As the genetic relationship matrix is calculated from the genotyped SNPs on a DNA array, which typically does not contain rare variants, this method estimates the combined genetic effect of all common SNPs (common as opposed to rare variants), hence it is called *common SNP heritability*. This common SNP heritability can be taken as an upper bound of the amount of heritability that can be explained by the markers detectable through a GWAS [27,29].

Comparing the results of GRS, PRS and GREML-GCTA shows how much heritability has been accounted for, and how much is still to be found. In **Chapter 5**, we used GRS and GREML-GCTA to estimate the known and missing heritability of 32 complex traits in the Lifelines Cohort Study. As in earlier studies [27,29], we found that GRS accounted for only a fraction of the total common SNP heritability. This suggests that there are many common variants that are not (yet) identified by GWAS studies. This may be a consequence of the restrictive nature of a GRS: only significant or replicated markers are included. Even large meta-analyses have only a limited power to detect less common variants, or those with small effect sizes. Based on this hypothesis, a polygenic risk score (PRS) analysis has been suggested, where the risk scores include a larger number of SNPs meeting increasingly more lenient significance thresholds. If a PRS accounts for a greater part of the total heritability than a GRS, this suggests that true associations are still hidden among the suggestive SNPs.

There are, however, two caveats to the heritability measured by the above methods. Firstly, all three methods model only the additive genetic effect (narrow-sense heritability) [28,30], so dominance or interactions between markers are not well modelled. Nevertheless, it has been argued that non-additive effects are less relevant in individuals that are not closely related [24]. Furthermore, previous studies of complex traits have found no significant effects of dominance or interaction [28,31]. This is confirmed by our own findings in chapter 5, where, using the same method as Zhu *et al.* [28], we found no significant dominance effects in the Lifelines cohort for any of the 32 complex traits analysed.

A second caveat is that all three methods rely on SNPs that are either assayed directly by a genotyping chip (which contain mostly, if not exclusively, common variants), or that are in linkage disequilibrium (LD) with such variants. The introduction of newer reference panels allows for the imputation of rarer SNPs, which has made this less of an issue for GRS and PRS, but the SNPs in **Chapter 5** were all derived from the older HapMap reference. This applies particularly to GREML-GCTA, which only looks at common SNPs. Therefore, even when disregarding dominance, interaction and gene-gene effects, common SNP heritability is not the same as total heritability.

When estimating the heritability of height and BMI in the Lifelines Cohort Study (**Chapter 5**), the common SNP heritability accounted for 49% and 25%, respectively, of the phenotypic variability. While an improvement over earlier findings by Yang *et al.* [32], whom reported 42% and 16%, respectively, these values still fall short of the heritability estimates reported by twin and family studies (80-90% for height, 42-80% for BMI). However, the same group recently adapted their GREML method to handle (imputed data based on) whole-genome sequencing data, including rare as well as common markers [33]. They called this method linkage-disequilibrium and minor-allele-frequency stratified GREML, or GREML-LDMS. Application to height and BMI yielded heritability estimates of 56% and 27%, respectively, suggesting that there are indeed undetected low frequent genetic variants contributing to the heritability. On the other hand, they argue that family-based studies may overestimate the total heritability as a result of shared environmental effects, and that, therefore, the heritability gap is very small. This also implies that the majority of the missing heritability is likely accounted for by rare SNPs, rather than dominance or interaction effects. Similarly, Zaitlen *et al.* [31], using a method that combines closely and distantly related samples, reported lower heritability for multiple complex phenotypes than traditional structural equation modelling in family-based studies. They also concluded that overestimation due to shared environment is responsible for (part of) the gap between common SNP heritability and family-based heritability estimates. The remainder is therefore more likely to be due to low frequency SNPs.

As a side note, we confirmed that the standard error of the heritability estimate of GREML-GCTA is strongly linked to sample size, being roughly equal to 300 divided by the sample size. As our EGCUT dataset in

Chapter 6, in which we estimated common SNP heritabilities for neuroticism and its underlying facets using GREML-GCTA, contained just over 3200 samples with phenotypes, this should result in a standard error of 9%. However, even in cohorts consisting of "unrelated" samples, GREML-GCTA still removes 15-20% of the samples due to close genetic similarity. As such, and given that the final heritability estimates generally ranged between 5-20%, the starting sample size of 3200 samples was underpowered.

In conclusion, these methods contribute to our understanding of the genetic mechanisms behind complex traits, showing us where the remaining part of the heritability is likely to be found.

Genome Reference Panels: the 1000 Genomes Project

In both **Chapters 7 and 8**, we described a meta-analysis of GWASs using a new, larger genomic reference panel derived from the 1000 Genomes Project [4,34,35] for imputation. The benefits of using this panel have been briefly discussed in **Chapter 7**, but deserve more attention. The first, and most obvious, benefit is that the 1000 Genomes Project was based on sequencing the DNA of individuals rather than genotyping them and therefore includes more genetic markers. Particularly the imputation of rare variants is important, as rare SNPs likely account for a substantial part of the gap between heritability found in family-based studies, and heritability reported by GREML-GCTA (see above). In addition, more individuals were sequenced allowing for more accurate phasing of alleles.

Secondly, the 1000 Genomes Project included more ethnic groups than the previous HapMap reference panel. As allele frequencies vary between ethnicities, variants that are rare in one group may be common in another. Thus, more ethnic diversity allows for better characterization and phasing of these variants. In addition, the greater genetic diversity in African samples results in a larger number of haplotypes, making rare variants easier to impute [35].

A few caveats should be mentioned, though. Firstly, the ability to accurately capture structural variation still remains limited. Secondly, increasing the number of variants also adds to the multiple testing burden [35]. Thirdly, from our experience (not discussed in the papers above), accurate imputation of rare variants still requires a sufficiently high-resolution DNA chip as a starting point. For example the Illumina CytoSNP chip used in Lifelines and TRAILS contained only 250,000 SNPs and produced usable GWAS results for only 8 million SNPs in our analyses, a number that coincides with the number of common SNPs in the final build of 1000 Genomes. In other words, we could not impute the rare SNPs of the 1000 Genomes build because of the limited resolution of our DNA chip.

In conclusion, the 1000 Genomes reference allows for more accurate imputation of the genome than previous references and hence higher power to detect new genetic variants for complex traits.

Future Perspectives

The recent boom in GWAS analyses has yielded a wealth of data, but also presents us with two further challenges. The first is that, despite the advances discussed in **Chapter 5** and in the section on 'Methods to Estimate Heritability: Genetic Risk Scores, Polygenic Risk Scores and GREML-GCTA' above, a substantial chunk of the heritability of complex traits is still missing. Our challenge is how to find the remainder. The second challenge is to turn the data already collected into biological insight and clinically relevant applications.

What we have not found: missing heritability

As discussed above, there is reason to believe that most of the missing heritability is "hiding rather than missing" [29] and resides in rare variants [33]. This means that it can be found with the current tools, if enough statistical power and a sufficiently high-resolution genome imputation are available. Both of these requirements are well understood by the scientific community, and addressing them is an ongoing effort.

The most straightforward method to increase power is by increasing the sample size. A common way to do this is by organizing a consortium of existing cohorts, and combining their results into a meta-analysis (as was done in **Chapters 7 and 8**). As the number of cohorts with data available on a specific phenotype is limited, the need for greater sample sizes may lead to consortia with a similar scope to join forces as well, for example the CARDIoGRAM and C4D consortia. Another way to increase sample size is by using electronic health records to supplement phenotype data [36].

The use of high-resolution genome references is also discussed above. **Chapter 7** demonstrates the benefits of using the 1000 Genomes Reference over the older HapMap reference. However, the 1000 Genomes Reference will be superseded itself by the Haplotype Reference Consortium project [37]. Imputation using the data of the Haplotype Reference Consortium, which includes the 1000 Genomes dataset, has already been successfully applied in GWAS [38-40]. In time, the cost of sequencing an entire genome may fall to a point where it becomes a feasible alternative to imputation. However, currently the combination of a high-resolution SNP array and imputation against a large reference dataset of fully sequenced genomes yields more statistical power for the same price [33]. Alternatively, one can sequence only the coding parts of the genome (the exome), which detects genetic variants that may directly alter the protein structure. This is a cheaper alternative than whole-genome sequencing, although it will not detect any variants outside the exome, such as those that are involved in gene regulation.

However, we do want to caution that, in our experience, a high-resolution genome reference is less effective if the initial genotyping has been carried out on a low-resolution DNA chip. This is not necessarily a problem, but it does limit the value of the newer genome references for cohorts that carried out their genotyping with older SNP arrays.

In addition to increasing genome resolution, it may also be worthwhile to look at marker types that are ignored in conventional GWAS. One of the restrictions of conventional GWAS software is that it is designed for biallelic markers. However, there are various polyallelic markers known, most importantly copy number variants (CNV). CNV have already been associated with several traits [41,42], but, as yet CNV analyses are conducted separately from a standard GWAS. For GWAS to become truly comprehensive, it would need to incorporate these markers as well. Another often ignored part of the genome are the X- and Y-chromosomes. Part of this lies in the difference between allosomal and autosomal markers: the former have only one instance of a marker per male person. Imputation and GWAS software were initially not suitable for these chromosomes, as they expect two alleles per marker. Furthermore, statistical power is much lower (in particular for the Y chromosome, which is only present in men), which renders these analyses of little value unless sample sizes are very large.

What we have found: integration & big data

Besides finding the missing heritability, the main future challenge of genetic epidemiology is how to best utilize the data that are already collected. One obvious application is that of personalized predictive genomics. Unfortunately, the results discussed in **Chapter 5** suggest that we cannot accurately predict complex traits yet. This is likely to remain a problem until we have made greater progress in eliminating the missing heritability.

Another important follow-up step is to focus on function, i.e., determining which genes in associated loci are responsible for the association with complex traits, and simultaneously the mechanisms through which these genes affect these traits. As the effects of a single marker are tiny, current methodological developments favour looking at the data as a whole rather than at single markers. For example, these days many GWAS are followed up by some form of gene-based or pathway analysis (as in **Chapters 7 and 8**). In a gene-based analysis SNPs are grouped based on their genes, followed by testing whether certain genes are enriched for significant and suggestive markers. A pathway analysis works similarly, but tests biological pathways for enrichment of genes.

The next step is to go beyond the genome-wide data and look at downstream markers to determine how genetic variants influence the phenotype. A genetic variant that was significant in GWAS is not likely to be the actual cause. Most likely, it is simply in LD with the causal variant. However even the location of the causal variant does not necessarily identify the responsible gene: it may be in a regulatory rather than coding region, which can affect genes some distance away.

Solving this puzzle, and bringing the genotype and phenotype closer, will require integrating genomic, epigenomic and transcriptomic data, and possibly even intermediate outcomes such as proteomics and metabolite levels [43]. As an example, in **Chapter 7** of this thesis, we combined genetic and expression data to test if the significant markers from the GWAS are also associated with the expression of nearby genes. This method is known as expression quantitative trait loci (eQTL) analysis. The same methodology can be applied to methylation (mQTL), protein and even metabolite levels. However, this leads to another challenge: the availability of such data. Methylation and expression levels are tissue-specific. As such, they are usually only measured in a relatively small number of samples for the specific tissues of interest. Most larger databases of eQTL and mQTL data are based on whole blood. This makes it harder to find a suitable cohort with the right tissue to run eQTL or mQTL analysis in. Also, few studies attempt to test every possible eQTL available to them, as it is time consuming and adds hugely to the multiple-testing burden. However, such QTL data might still be valuable to a researcher who is interested in a specific region or marker, and therefore doesn't face this multiple-testing burden. It would be very beneficial if existing cohorts that combined genetic with expression and/or methylation data would run a full eQTL/mQTL analysis and make the results available to the public. This allows outside researchers to look up their own GWAS results for eQTL and mQTL. A good example of this would be the GTEx Consortium, which was founded for this purpose and has made eQTL data of multiple tissues available [44].

The above, however, leads to a further problem: the increasing size of the datasets and results files. It is already infeasible to run a GWAS on a desktop PC; a computer cluster is required to analyse a dataset imputed with 1000 Genomes data within a reasonable time-frame. In order to deal with future massive datasets, the field of genetic epidemiology will likely have to employ the methods used by commercial companies working with big data. A commonly used method is the open source platform Hadoop to store and manage data. It functions by splitting the dataset into small chunks and moving the process to the data rather than the other way round. A similar issue is faced by biobanks: for a variety of reasons (privacy, data protection, network load) it is undesirable that researchers simply download the entire dataset from their server, and run the analysis on their own computer. The solution is for the biobank to set up their own computer environment which the researcher can access from a distance to run his analysis locally. This allows for the data (and hence the patient privacy) to be protected and reduces the network load. However, it also requires an investment of time and money to create this environment and maintain it. As scientific methods are constantly being updated, the environment needs to be flexible enough that a researcher can run non-standard tools. The data itself also need to be accessible and well-documented.

As an alternative to dealing with the ever-increasing file- and sample-size requirements, various software tools are under development that bypass the need for individual-level genome-wide data by using (meta-) GWAS summary statistics. For example, the LD score regression methods [45,46] combines the results of a GWAS with the LD score (the amount of genetic variation that is tagged by an individual marker) to calculate the heritability of a single trait, or the genetic correlation between two traits (that is: the proportion of variance the two traits share due to genetic causes). If two traits are genetically highly correlated, it could be worthwhile to combine the available data for these traits in a multivariate analysis to improve power and find more (shared) genes for both traits.

LD can also be employed to detect secondary variants in GWAS summary statistics [47]. A secondary variant is a genome-wide significant marker that is not in (strong) LD with another significant marker and whose effect is therefore (in part) independent of that marker. More recently, it has been proposed to use haplotypes rather than LD [48], which should aid in the identification of multiple causal variants within the same LD block. These secondary variants could then be treated in a similar way as the primary variants, i.e., help identify the underlying causal gene and when added to genetic risk scores explain more of the phenotypic variation.

Another useful tool that makes use of GWAS summary statistics is the R package TwoSampleMR, which was developed for Mendelian Randomization (MR) analyses. This approach uses genetic variants in observational epidemiology to make causal inferences about modifiable risk factors for disease and health-related outcomes. The main rationale in Mendelian randomization is that, if a risk factor or biomarker is causally related to a disease outcome, a genetic variant or GRS that is known to be associated with the biomarker should have a similar relation to the outcome as the supposedly causal marker itself. The developers of TwoSampleMR also set up a website MRBase.org ,where researchers can run MR analysis using their own GWAS summary statistics and/or published GWAS results [49]. This allows researchers to test the causal effects of biomarkers such as CRP on disease outcomes through using the GWAS-identified SNPs as instrumental variables [50]. Other promising methods exist that impute expression levels based on GWAS summary statistics, bypassing the need to collect expression data, in order to analyse the relation between gene expression and the phenotype of interest [51-53].

In conclusion, methods using GWAS summary statistics will become increasingly popular to integrate data from different sources in order to unravel the aetiology of complex diseases.

Conclusion

The purpose of this thesis was to develop and apply novel ways to obtain more and higher-quality results from genome-wide association data. We approached this from four different angles: QC of GWAS and EWAS results, use of survival analysis in GWAS, estimation of common-SNP heritability of complex traits, and the use of a more detailed reference genome for imputation. As such, the current thesis has attempted to contribute to genome-wide association analyses by altering the balance between brute force and sophistication in favour of the latter.

References

1. van der Most PJ, Vaez A, Prins BP *et al:* QCGWAS: A flexible R package for automated quality control of genome-wide association results. *Bioinformatics* 2014; **30:** 1185-1186.
2. Van der Most PJ, Kupers LK, Snieder H, Nolte I: QCEWAS: automated quality control of results of epigenome-wide association studies. *Bioinformatics* 2017; **33:** 1243-1245.
3. de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF: Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17:** R122-R128.
4. Altshuler D, Durbin RM, Abecasis GR *et al:* A map of human genome variation from population-scale sequencing. *Nature* 2010; **467:** 1061-1073.
5. Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55:** 997-1004.
6. Sebastiani P, Solovieff N, Puca A *et al:* Editorial Expression of Concern (Retraction of vol 330, pg 912, 2010). *Science* 2011; **333:** 404-404.
7. Ledford H: Paper on genetics of longevity retracted. *Nature* 2011; http://www.nature.com/news/2011/110721/full/news.2011.429.html.
8. Prins BP: Inflammatory Biomarker Genomics. Groningen, University of Groningen, 2016; p. 176.
9. MacArthur J, Bowler E, Cerezo M *et al:* GWAS catalog. 2017; http://www.ebi.ac.uk/gwas/.
10. Eddelbuettel D: High-Performance and Parallel Computing with R. 2017; https://cran.r-project.org/web/views/HighPerformanceComputing.html.

11. Adler D, Glaser C, Nenadic O, Oehlschlagel J, Zucchini W: ff: memory-efficient storage of large data on disk and fast access functions. 2014; https://CRAN.R-project.org/package=ff.

12. Kane MJ, Emerson JW, Haverty P, Determan Jr. C: bigmemory: Manage Massive Matrices with Shared Memory and Memory-Mapped Files. 2016; https://CRAN.R-project.org/package=bigmemory.

13. Barban N, Jansen R, de Vlaming R *et al:* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat Genet* 2016; **48:** 1462-1472.

14. Nolte IM, Munoz ML, Tragante V *et al:* Genetic loci associated with heart rate variability and their effects on cardiac disease risk. *Nature Communications* 2017; accepted for publication.

15. Gillespie BW, Chen Q, Reichert H *et al:* Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator. *Epidemiology* 2010; **21:** S64-S70.

16. Sattar A, Sinha SK, Morris NJ: A Parametric Survival Model When a Covariate is Subject to Left-Censoring. *J Biomet Biostat* 2012; **3**.

17. Dinse GE, Jusko TA, Ho LA *et al:* Accommodating Measurements Below a Limit of Detection: A Novel Application of Cox Regression. *Am J Epidemiol* 2014; **179:** 1018-1024.

18. Stringer S, Minica CC, Verweij KJH *et al:* Genome-wide association study of lifetime cannabis use based on a large meta-analytic sample of 32330 subjects from the International Cannabis Consortium. *Translational Psychiatry* 2016; **6:** e769.

19. van der Net JB, Janssens ACJW, Eijkemans MJC, Kastelein JJP, Sijbrands EJG, Steyerberg EW: Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. *European Journal of Human Genetics* 2008; **16:** 1111-1116.

20. Stringer S, Denys D, Kahn RS, Derks EM: What Cure Models Can Teach us About Genome-Wide Survival Analysis. *Behav Genet* 2016; **46:** 269-280.

21. Lubin J, Colt J, Camann D *et al:* Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect* 2004; **112:** 1691-1696.

22. Armbruster DA, Pry T: Limit of Blank, Limit of Detection and Limit of Quantitation. *Clin Biochem Rev* 2008; **29:** S49.

23. Uh H, Hartgers FC, Yazdanbakhsh M, Houwing-Duistermaat JJ: Evaluation of regression methods when immunological measurements are constrained by detection limits. *Bmc Immunology* 2008; **9:** 59.

24. Visscher PM, Hill WG, Wray NR: Heritability in the genomics era - concepts and misconceptions. *Nat Rev Genet* 2008; **9:** 255-266.

25. Jamshidi Y, Nolte IM, Spector TD, Snieder H: Novel genes for QTc interval. How much heritability is explained, and how much is left to find? *Genome Med* 2010; **2:** 35.

26. 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.

27. Yang J, Benyamin B, McEvoy BP *et al:* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42:** 565-569.

28. Zhu Z, Bakshi A, Vinkhuyzen AAE *et al:* Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. *Am J Hum Genet* 2015; **96:** 377-385.

29. 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.

30. Zuk O, Hechter E, Sunyaev SR, Lander ES: The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* 2012; **109:** 1193-1198.

31. Zaitlen N, Kraft P, Patterson N *et al:* Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genet* 2013; **9:** UNSP e1003520.

32. Yang J, Manolio TA, Pasquale LR *et al:* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* 2011; **43:** 519-525.

33. Yang J, Bakshi A, Zhu Z *et al:* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 2015; **47:** 1114-1120.

34. Altshuler DM, Durbin RM, Abecasis GR *et al:* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491:** 56-65.

35. Altshuler DM, Durbin RM, Abecasis GR *et al:* A global reference for human genetic variation. *Nature* 2015; **526:** 68-74.

36. Hoffmann TJ, Ehret GB, Nandakumar P *et al:* Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat Genet* 2017; **49:** 54-64.

37. McCarthy S, Das S, Kretzschmar W *et al:* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; **48:** 1279-1283.
38. Lesseur C, Diergaarde B, Olshan AF *et al:* Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat Genet* 2016; **48:** 1544-1550.
39. Hernandez-Pacheco N, Flores C, Alonso S *et al:* Identification of a novel locus associated with skin colour in African-admixed populations. *Scientific Reports* 2017; **7:** 44548.
40. Nagy R, Boutin TS, Marten J *et al:* Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Medicine* 2017; **9:** 23.
41. Wheeler E, Huang N, Bochukova EG *et al:* Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nat Genet* 2013; **45:** 513-517.
42. Peterson RE, Maes HH, Lin P *et al:* On the association of common and rare genetic variation influencing body mass index: a combined SNP and CNV analysis. *BMC Genomics* 2014; **15:** 368.
43. Prins BP, Lagou V, Asselbergs FW, Snieder H, Fu J: Genetics of coronary artery disease: Genome-wide association studies and beyond. *Atherosclerosis* 2012; **225:** 1-10.
44. Lonsdale J, Thomas J, Salvatore M *et al:* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013; **45:** 580-585.
45. Bulik-Sullivan B, Finucane HK, Anttila V *et al:* An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015; **47:** 1236-1241.
46. Bulik-Sullivan BK, Loh P, Finucane HK *et al:* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47:** 291-295.
47. Yang J, Ferreira T, Morris AP *et al:* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; **44:** 369-375.
48. Zheng J, Rodriguez S, Laurin C *et al:* HAPRAP: a haplotype-based iterative method for statistical fine mapping using GWAS summary statistics. *Bioinformatics* 2017; **33:** 79-86.
49. Hemani G, Zheng J, Wade KH *et al:* MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv* 2016.
50. Prins BP, Abbasi A, Wong A *et al:* Investigating the Causal Relationship of C-Reactive Protein with 32 Complex Somatic and Psychiatric Outcomes: A Large-Scale Cross-Consortium Mendelian Randomization Study. *Plos Medicine* 2016; **13:** e1001976.
51. Gusev A, Ko A, Shi H *et al:* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016; **48:** 245-252.
52. Zhu Z, Zhang F, Hu H *et al:* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 2016; **48:** 481-487.
53. Barbeira AN, Dickinson SP, Torres JM *et al:* Integrating tissue specific mechanisms into GWAS summary results. *bioRxiv* 2017.