Biostatistical Aspects of Genome-Wide Association Studies

Andreas Ziegler[*,1], **Inke R. König**[1], and **John R. Thompson**[2]

[1] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

[2] Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester, 2nd Floor, Adrian Building, University Road, Leicester, LE1 7RH, UK

Received 30 October 2007, revised 4 December 2007, accepted 5 December 2007

Summary

To search the entire human genome for association is a novel and promising approach to unravelling the genetic basis of complex genetic diseases. In these genome-wide association studies (GWAs), several hundreds of thousands of single nucleotide polymorphisms (SNPs) are analyzed at the same time, posing substantial biostatistical and computational challenges. In this paper, we discuss a number of biostatistical aspects of GWAs in detail. We specifically consider quality control issues and show that signal intensity plots are a sine qua condition non in todays GWAs. Approaches to detect and adjust for population stratification are briefly examined. We discuss different strategies aimed at tackling the problem of multiple testing, including adjustment of p-values, the false positive report probability and the false discovery rate. Another aspect of GWAs requiring special attention is the search for gene-gene and gene-environment interactions. We finally describe multistage approaches to GWAs.

Key words: Cochran-Armitage trend test; False discovery rate; False positive report probability; Sequential study design; Signal intensity plot.

1 Introduction

In their last issue in 2006, the News Staff from Science announced genome-wide association studies (GWAs) to be one of the areas to watch in 2007 (The News Staff, 2006). Indeed, after several years of relatively unsuccessful attempts to identify genetic variants responsible for complex and common diseases (Colhoun, McKeigue and Davey Smith, 2003; Hattersley and McCarthy, 2005) such as coronary artery disease (Watkins and Farrall, 2006), the first half of 2007 has seen a surge in publications of GWAs identifying and validating genetic factors for, to name a few, myocardial infarction (Samani et al., 2007), colorectal cancer (Tomlinson et al., 2007; Zanke et al., 2007), breast cancer (Easton et al., 2007; Hunter et al., 2007), and six further common diseases (The Wellcome Trust Case Control Consortium, 2007). However, in addition to their apparent success, GWAs also pose new challenges for biometricians, which will be the focus of this work. Biometricians have started addressing these challenges and new methodological approaches sprout like mushrooms almost every day. We therefore aim at addressing fundamental biostatistical aspects of GWAs. Before delving into these, let us describe the typical scenario of a GWA (Table 1). The aim of any genetic association study is to identify associations between a phenotype on the one hand, which is, in many cases, a binary disease status, and one or more genetic markers on the other. Specifically, GWAs use *single nucleotide polymorphisms* (SNPs) as genetic markers, as they are easy to type and abundant in the human genome.

[*] Corresponding author: e-mail: ziegler@imbs.uni-luebeck.de, Phone: +49 451 500 2780, Fax: +49 451 500 2999

Table 1 Typical scenario for a genome-wide association study.

Study design	Case-control study, i.e., affected cases and healthy controls
Variables	Genetic markers (single nucleotide polymorphisms, SNPs), typically with two different forms A and B
	One of three genotypes per person: AA, AB, or BB
Technology	Affymetrix GeneChip Human Mapping 500K Array Set: ∼500 000 SNPs
	Affymetrix Genome-Wide Human SNP Array 5.0: ∼500 000 SNPs (same SNPs as on array set but on a single chip; ∼420 000 copy number variations added)
	Affymetrix Genome-Wide Human SNP Array 6.0: ∼1 000 000 SNPs
	Illumina HumanHap550 Genotyping BeadChip: ∼550 000 SNPs
	Illumina HumanHap650Y Genotyping BeadChip: ∼650 000 SNPs (the Y indicates that the chip contains ∼100 000 Yoruba specific SNPs)
	Illumina human1M BeadChip: ∼1 000 000 SNPs
Quality control	Exclusion of low-quality SNPs (see Section 3)
	Exlusion of probands with low-quality genotypes (see Section 3)
	Exclusion of probands with contaminated DNA (see Section 3)
Statistical analysis	Assumption of a specific genetic model: dominant (AA versus AB and BB), recessive (AA and AB versus BB), additive (assuming a linear effect for genotypes AA, AB and BB); no assumption of a specific genetic model, resulting in tests with 2 degrees of freedom
	Comparison of genotype frequencies between cases and controls for every SNP

GWAs systematically investigate SNPs in the entire human genome without an a priori specified hypothesis on the location. A GWA can only be successful, however, if the so-called common disease/common variant (CDCV) assumption holds. This popular hypothesis proposes that most of the genetic risk for common, complex diseases is caused by a small to moderate number of disease loci having common variants. These disease loci are assumed to have small effects, and association mapping is appropriate in this circumstance. However, if this CDCV hypothesis is false, e.g., there are many rare variants at a disease locus, association mapping will be rather difficult.

GWAs are in contrast to so-called candidate gene association studies, where SNPs are analyzed in candidate genes or regions (Ziegler and König, 2006). These candidates are plausible because of biological function or previous study results, and some of them have strong genetic effects. The approach is primarily indirect as we do not necessarily suppose that any of the SNPs are causal but rather that they may be in close vicinity to functional mutations and therefore associated with the disease because of *linkage disequilibrium* (LD). The functional variant might not even be a SNP, but an inversion, an *insertion*, a *deletion* or a *copy number variation* (CNV). For the systematic search of the entire human genome with a length of approximately 3.3×10^9 base-pairs, in subjects of European ancestry only approximately 300 000 well-chosen SNPs, termed tagSNPs, are required to capture most of the genomic variation. Here, it is important to note that chips are designed in different ways. For example, for the 1M BeadChip Illumina specifically chose 950 000 tagSNPs from the HapMap project (see the *HapMap project*) and 100 000 non HapMap SNPs. In comparison, Affymetrix used an "unbiased selection of 482 000 SNPs from the SNP Array 5.0" for the SNP Array 6.0, thus not following the tagSNP approach based on the HapMap project.

Compared with gene expression studies where typically the sample size is relatively small, large sample sizes are utilized in GWAs. Because the SNP chips currently are more expensive than the expression chips, the total cost of GWAs is high. Therefore, some groups use specific sequential study designs for minimizing the total study cost (see Section 5).

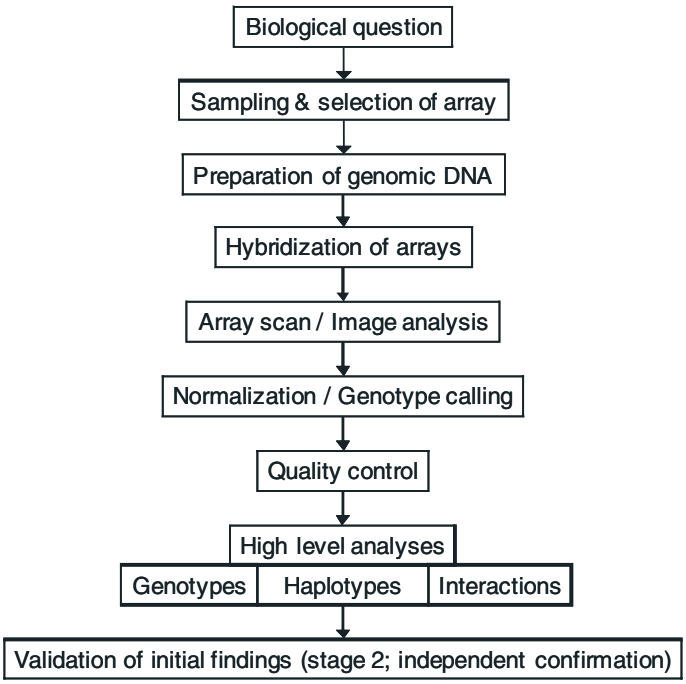


Figure 1 Succession of design, experimental and data analysis steps in a genome-wide association study.

A sequence of steps has to be taken in a successful GWA (Figure 1). First, subjects have to be recruited, their phenotypes have to be determined, and their DNA needs to be extracted. According to specific protocols (see, e.g., Affymetrix, 2006), the DNA needs to be prepared before hybridization. This preparation process is not of primary importance for data analysis at later stages and includes, e.g., a polymerase chain reaction step. The following steps are similar to those of gene expression studies. After washing and staining, the array is scanned, and the raw optical images, called the `.dat` files, are obtained. The pixel values are used for calculating the intensities, which are then stored in the automatically generated `.cel` files. Here, a single representative intensity value is obtained per cell (feature) of the image. Unfortunately, the `.dat` files are usually not kept because of their size, approximately 1GB per chip for the Affymetrix Genome-Wide Human SNP Array 6.0. The biostatistical analysis team typically receives the `.cel` files from the laboratory. However, spot intensity could serve as an indicator for chip quality as for most microarray platforms (Hartmann, 2005).

When image analysis has been completed and the `.cel` files are available for all subjects, the, signal intensities are normalized and genotypes are called, i.e., signal intensities are converted into genotypes. The genotype calling step is important, and we describe the basic ideas in Section 3. After genotype calling, intensive quality control–low level analysis–needs to be performed (Section 3), which is then followed by high level statistical analyses. Usually, genotype or allele frequencies between cases and controls are compared SNP by SNP to test for association (Section 4). Both simple and more complex frequentist and Bayesian approaches are employed, and computational efficiency is an important issue here. The findings of these analyses are then carried forward to a second stage of analyses (Section 5) or to an independent confirmation study.

In the remainder of this work, we concentrate on GWAs performed using the Affymetrix chip technology because the authors have more experience with this platform than with the Illumina Bead-

Chip technology. Nevertheless, many biostatistical aspects are comparable, e.g., the quality control procedures. To help biostatisticians who are not familiar with the terms used in the area of genetic epidemiology, Table 2 provides a short glossary.

Table 2 Glossary.

Assortative mating	Also termed assortative pairing; takes place when individuals tend to mate individuals who are similar to themselves (positive assortative mating) or dissimilar (negative assortative mating).
CNP, CNV	Copy-number polymorphism, copy-number variation; a normal variation in DNA due to variation in the number of copies of a sequence within the DNA. CNV is sometimes defined in terms of segment length, e.g., the DNA segment presenting a variable copy number has to be at least 1 kb. CNP is sometimes defined in terms of frequency, e.g., there have to be at least two variants with $\geq 1\%$ frequency.
Deletion	A mutation in which at least one base pair is removed from a DNA sequence.
Epistasis	From the greek word epistanai meaning "to stop"; for a detailed discussion, see Cordell (2002). It is a masking effect whereby a variant or allele at one locus prevents the variant at another locus from manifesting its effect. It is therefore a special case of *interaction*. Unfortunately, gene-gene *interaction* and epistasis are increasingly used synonymously. Epistasis is different from dominance: with epistasis a variant in one gene masks the expression of a different gene. With dominance, one variant of a gene masks the expression of another variant of the same gene.
Haplotype	The particular combination of alleles on the same chromosome typically inherited together.
Hardy-Weinberg equilibrium	(HWE) Holds at a locus when the two alleles within an individual are statistically independent.
Insertion	A mutation in which at least one base pair is added in a DNA sequence.
Interaction	Statistical interaction is the deviation from additivity in the effect of two independent variables on a dependent variable. Unfortunately, gene-gene interaction and *epistasis* are sometimes used synonymously.
Linkage disequilibrium	(LD) The situation in which alleles of two loci do not arise independently of one another. The term "association" unfortunately is often used synonymously. Linkage disequilibrium can be due to the fact that two loci are linked, i.e., in close vicinity. It may also arise at loci on different chromosomes, e.g., because of selection or population stratification.
Population stratification	Also termed population heterogeneity or confounding by ethnicity. Population stratification refers to a situation where subgroups of individuals are on average more related to each other than to other members of the wider population. For example, cases are often considered to be closer related than controls, leading to different allele frequencies between cases and controls caused by this systematic difference in ancestry rather than association of genes with disease.
Selection	Any process that alters allele frequency in a directed fashion without mutation or immigration.
Sequencing	Process of determining the nucleotide order of a given DNA fragment.
SNP	Single nucleotide polymorphism; SNPs are variations at only a single base, meaning that one base is substituted by another.
TaqMan	TaqMan® measures the accumulation of a DNA sequence, i.e., a polymerase chain reaction (PCR) product via fluorophore in real-time during the exponential stages of the PCR.

2 Challenges of Genome-Wide Association Studies

In comparison with candidate gene association studies and, indeed, compared with most genetic or clinical studies, GWAs are special in a number of ways. What first comes to mind is the sheer amount of data, because on currently applied technologies, hundreds of thousands or even a million genotypes are assessed per individual. This has a number of consequences. The first is a logistical issue, because computer capacities need to be large, both with regard to storage and to CPU time for data analysis. For example, the uncompressed `.cel` files of ∼1000 subjects require ∼150 GB of storage, and a typical analysis directory of such a GWA, without `.cel` files but with called genotypes, requires more than 100 GB. The genotype calling of ∼2000 subjects with the Genome-Wide Human SNP Array 6.0 requires ∼10 days on an Intel Xeon Dual Quad Core E5345 with 2.3 GHz and 16 GB RAM.

In this line, we need new partners: For clinical studies, we were used to working closely with clinical partners. In genetic epidemiological projects, we have learned to work hand in hand with clinicians and molecular biologists, in order to understand the peculiarities of the variables we are analyzing, the kinds of errors which might occur, and the biological meaning of the results. In GWAs, skills and algorithms from computer science and bioinformatics are also required.

The second consequence relates to the statistical analysis. The specific challenge in GWAs is that the increase in the number of variables is not paralleled by an increase in the numbers of probands, so that we have a large p small n problem, similar to gene expression studies. To make this even more complicated, the SNPs are dependent on each other, thus in LD (Ziegler and König, 2006). Finally, the variables in a GWA are generated in a highly automated way. This poses further challenges for the quality control of the data, which will be the topic of the next section.

3 First Things First – Quality Control in GWAs

In contrast to candidate gene association studies or controlled clinical trials, it is almost impossible in a GWA to scrutinize every single variable for its quality in an individual monitoring procedure. Because of the high number of variables, quality control needs to be highly automated. Some of the quality control depends on the genotyping technology employed, and, as in the previous section, we will focus on the procedures for the Affymetrix platform. Although some divergence exists, a number of quality criteria seem to have been agreed upon and are currently typically in use.

3.1 Genotype calling and signal intensity plots

To understand the quality control, it is important to know how the variables are generated. In brief, for every proband at each SNP, we measure the signal intensities giving the strength of the signal for each of the alleles. If one intensity is high and the other is low, the proband is classified as homozygous, if the intensities for both alleles are equally high, the proband is assigned a heterozygous genotype.

In slightly greater detail, the procedure can be described as follows. Data are normalized and genotypes are called when the `.cel` files for all subjects are available, that is when image analysis has been completed. Genotype calling, i.e., assignment of genotypes to subjects according to their signal intensities is performed automatically using one of the available algorithms, for which the basic ideas are described here. A subject is either homozygous for the A or for the B allele or has a heterozygous genotype, where A and B refer to the minor, i.e., the less frequent, and the major allele, respectively. Thus, rather than using the actual bases, a more general coding system is employed. For the reader, it is important to note that different coding systems are used in different laboratories and research groups, and, in fact, the different coding schemes may lead to confusion in pooled analyses.

In a perfect world, all conditions in the experiment would be identical, all subjects with the same genotypes would have identical signal intensities, and there would be only three different signal intensities (Figure 2a). However, there are many factors affecting the signal intensity at the subject level,

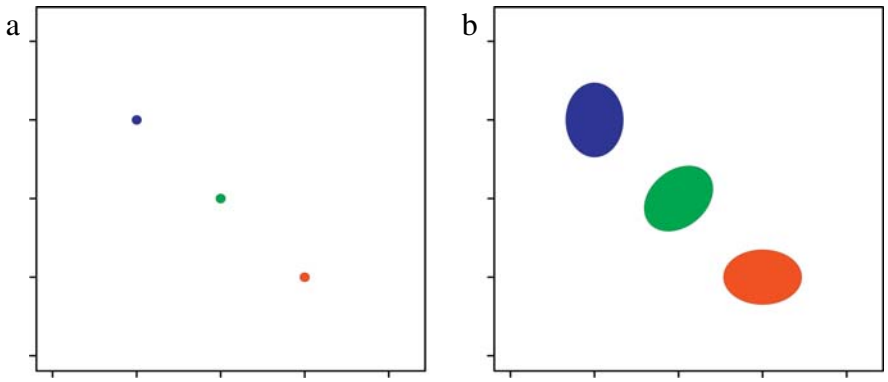


Figure 2 Signal intensity plots. Figure a) shows the idealized situation that all subjects can be assigned uniquely without measurement error to one of the three possible genotypes. Figure b) displays the more realistic scenario of a high quality SNP. Signal intensities of all subjects represent three clouds. The cloud of signal intensities for heterozygous subjects is not exactly between the clouds of the homozygous subjects. A slight shift is observed because of different probe affinities.

including DNA concentration or specific features during the hybridization process so that signal intensities of all subjects for the three genotypes form three clouds (Figure 2b). Furthermore we have to expect that the intensity of heterozygous subjects may not necessarily be between the respective intensity of homozygous subjects. Instead, there might be a slight shift in intensities because of different probe affinities.

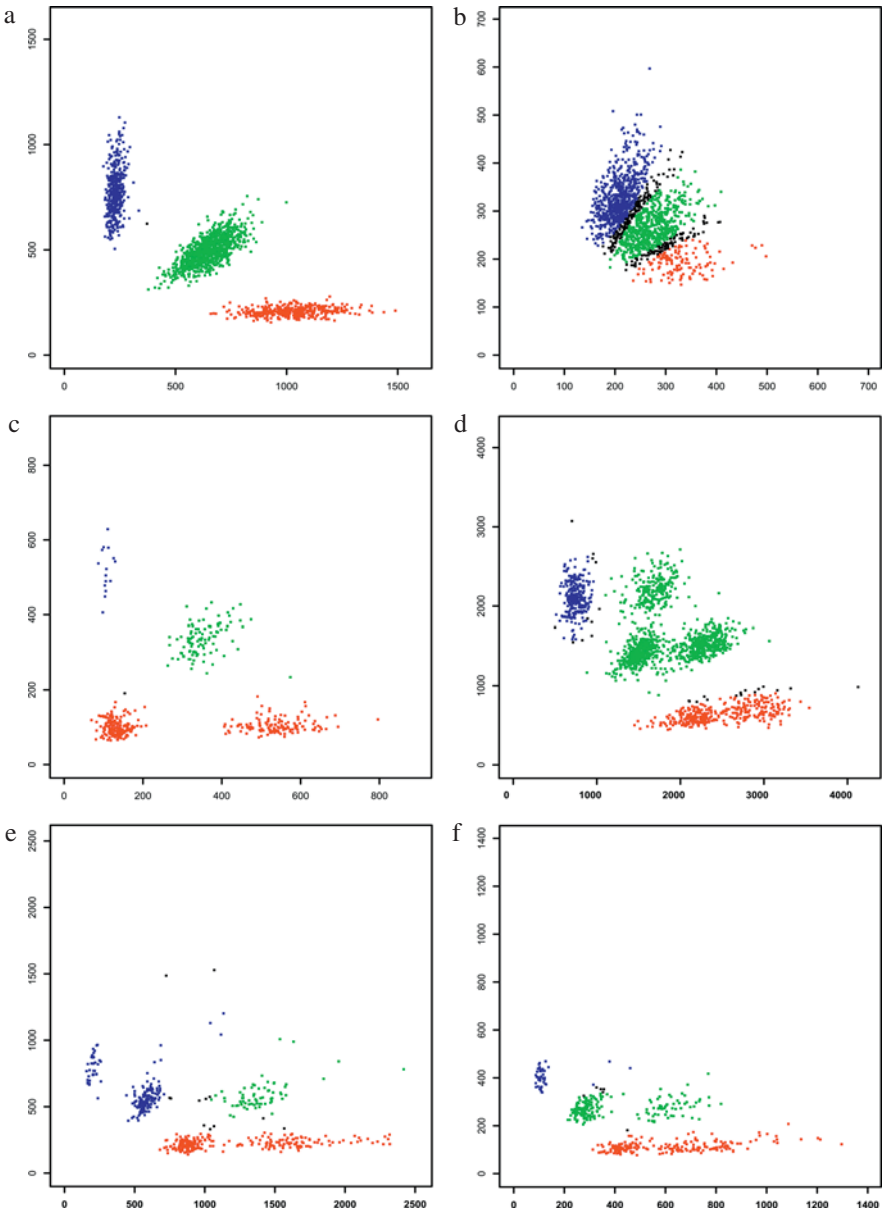
This biological knowledge is used in the genotype calling algorithms. For example, the Birdseed calling algorithm performs a multiple-chip analysis to estimate the signal intensity for each allele of each SNP, fitting probe-specific effects to increase precision. In the next step, Birdseed calls genotypes by fitting a Gaussian mixture model in the two-dimensional A-signal vs. B-signal space, using SNP-specific models to increase performance. Other algorithms have been proposed, including BRLMM (Bayesian Robust Linear Modeling using Mahalanobis distance, Rabbee and Speed, 2006), Chiamo++ (see Supplementary Material of The Wellcome Trust Case Control Consortium, 2007),

▶

Figure 3 Signal intensity plots: X-axis: signal intensity of minor allele; y-axis: signal intensity of major allele; red: homozygous A; green: heterozygous; blue: homozygous B; black: no genotype assignment. The three genotypes can be distinguished clearly in Figure a, and the genotypes have been correctly assigned. The SNP has a high genotyping quality. For the SNP displayed in Figure b, the clouds are not separate. Furthermore, the calling algorithm split the cloud of heterozygous subjects, resulting in a high proportion of missing genotypes. The SNP thus has a low quality. Quality problems for the SNPs displayed in Figures c through f were only detected using cluster plots. Figure c shows a SNP where many subjects neither have a substantial intensity for the A nor for the B allele, giving a fourth cloud. Figure d shows a SNP exhibiting more than three clusters. In fact, this SNP has homologous sequences on chromosome 1 and chromosome Y. Figures e and f give the cluster plots for two study groups and their genotype assignments at the same SNP. First, this SNP shows more than three clouds. Second, the calling algorithm assigns different genotypes to the second cloud from the left. Specifically, in Figure e they are assigned as homozygous B, while they were called heterozygous in Figure f.

GEL (genotype calling using empirical likelihood, Nicolae et al., 2006), and SNiPer-HD (SNiPer-high density, Hua et al., 2007); most of these use a Bayesian approach and utilize prior information on the position of the clouds.

One important question is whether all subjects in a study should be called together. Clayton et al. (2005) showed that there might be a differential bias resulting in a displacement of genotype clouds between cases and controls. This differential bias may originate from the preparation of DNA samples in different laboratories or on different well-plates prior to high-throughput genotyping. Also, different DNA concentrations or a degeneration of the arrays over time might lead to differential bias. While the standard approach to genotype calling currently is separate calling of cases and controls, a joint calling of all subjects adjusting for this possible differential bias is preferable, especially when selec-



tion, leading to substantially different genotype distributions between cases and controls, plays a role. To this end, Plagnol et al. (2007) recently proposed an expectation maximization algorithm to link cases and controls, or more generally samples from different DNA sources, during calling. Here, genotype frequencies are jointly calculated between cases and controls. Chiamo++ as developed for The Wellcome Trust Case Control Consortium (2007) follows a similar approach, but within a Bayesian framework.

To evaluate the quality of the genotype calls, the signal intensities for both alleles can be plotted together with the classification (Figure 3), and SNPs with dubious genotype assignments can be excluded. In the presence of CNVs the number of alleles varies between subjects and some markers may exhibit more than three clusters (Figure 3) or just greater variability.

The quality of signal intensity plots is typically judged by visual inspection by two independent and experienced readers. Consequently, in most applications to date, the cluster plots are not evaluated as the first step of the quality control procedure; instead, only the signal intensities from those SNPs that survive the following quality control and which show some evidence of association in the statistical analysis are plotted.

Already, there have been some attempts to automate the time-consuming cluster plot inspection. For example, Plagnol et al. (2007) designed a measure that captures the intuition that clouds of points should be well separated for a given SNP. For this, they consider the smallest difference between the centres of two adjacent clouds divided by the sum of the standard deviation for these two clouds. We (unpublished) not only consider the distance between clouds but also investigate the overlap of clusters. For this, we estimate an ellipsoid for each cloud and allow a sensitivity parameter to vary which determines the width of the cloud. Undoubtedly, more work needs to be done in this area.

Finally we note that the Affymetrix Genome-Wide Human SNP 5.0 and 6.0 Arrays, two of the newest chips, do not contain mismatch (MM) probes. In brief, on previous chips for each probe on the array that perfectly matches its target sequence, a paired "mismatch" probe was built to measure the magnitude of cross-hybridization (for details, see, e.g., Irizarry et al., 2003). This is of importance for some calling algorithms such as GEL (Nicolae et al., 2006), which explicitly model the difference between the perfect match (PM) intensity and the MM intensity. The absence of MM probe pairs has, however, been compensated for by a design change. On the 6.0 array, the probes of a pair of alleles are located on adjacent positions on the chip and are pairwise right-most embedded. In contrast, on the Affymetrix 500K Array Set, all probes were tiled as single probes, i.e., they did not necessarily occupy adjacent positions on the chip, and they were right-most embedded but not pairwise aligned. Therefore, MM intensities gave valuable information on background intensities for the Array set. This loss in information may, however, be compensated for by the optimized selection of probe pairs for the newer chips.

3.2 Cross platform and cross technology comparisons

High reproducibility of genotypes is important. Therefore, cross platform or cross technology comparisons should be performed for interesting findings. Specifically, sequencing a DNA fragment including the SNP or the use of a different technology like TaqMan® is advisable, and concordance statistics should be reported. Although no clear guidelines exist, the reproducibility of genotypes on the same platform should be around 99%, and the cross-platform concordance should not be below 95%.

3.3 Missing frequency per SNP

SNPs are of questionable quality if their genotyping failed in many individuals, or if for many probands, the calling algorithm was not able to assign a genotype from the signal intensities. Thus, an important quality criterion is the frequency of missing genotypes for each SNP. This criterion should be investigated separately for all study groups because of differential missingness. For example, in

The Wellcome Trust Case Control Consortium (2007) study, SNPs were excluded if they had a missing frequency of $>3\%$ in either study group. Similarly, Samani et al. (2007) excluded SNPs if they revealed $>2\%$ missing in either cases or controls.

3.4 Minor allele frequency

Unfortunately, many of the genotyped SNPs on the current Affymetrix arrays are monomorphic in all populations. Thus, they do not have two variants but just one. For instance, in the German data set analysed in Samani et al. (2007), 8% of the SNPs were monomorphic. The minor allele frequency is typically used as data filter. For example, 12.4% of the SNPs, including the monomorphic ones, had an allele frequency $<1\%$ for the minor allele in the Samani et al. (2007) study. Because of low power to detect an association between the SNP and the trait of interest, it is reasonable to exclude these SNPs at the outset. Therefore, SNPs are often excluded from analysis if the minor allele frequency (MAF) is low. The filter criterion typically varies with sample size and values varying from 1% to 5% are common.

3.5 Comparison of control groups

If more than one control group is available, as in The Wellcome Trust Case Control Consortium (2007), a comparison of the genotype frequencies between the control groups is advisable. Specifically, trends of genotypes should be identical, and an equivalence test is appropriate. Unfortunately, an equivalence Cochran-Armitage trend test is not available in any of the standard packages for GWAs.

3.6 Hardy-Weinberg equilibrium

A specific population genetic principle, which describes the relationship between allele and genotype frequencies, is often utilized as quality control measure. The general statement of the *Hardy-Weinberg equilibrium* (HWE) is that the genotype and allele frequencies in a large, randomly mating population remain stable over generations, and that there is a fixed relationship between allele and genotype frequencies (Ziegler and König, 2006). As a consequence, deviations from this relationship may point to quality problems in the genotyping procedure. However, because HWE is based on a number of assumptions, deviations may also occur in the absence of genotyping error, through mechanisms such as *selection* or *assortative mating*. As selection by disease status may affect HWE, a common recommendation is to check deviation from HWE only in the control group from a case-control study, and typically SNPs are excluded from further analysis if the test for deviation from HWE in the controls yields a p-value less than 10^{-4} (see, e.g., Samani et al., 2007; The Wellcome Trust Case Control Consortium, 2007). This approach can, however, be criticized in several ways, for instance, if the entire population is in HWE but the cases deviate from HWE, then the controls should also show a deviation from HWE (Wittke-Thompson, Pluzhnikov and Cox, 2005).

3.7 Quality control on the subject level

Several quality procedures are available that utilize subject level data. The complement of the SNP-wise missing frequency is the subject-wise missing frequency, usually termed the call rate. A typical choice is to exclude probands from the analysis who have been successfully genotyped at less than 97% of the SNPs. It is important to note that this number still includes the monomorphic SNPs. Furthermore, this number is sometimes lowered because of conditions outlined in a genotyping contract. For example, it might be that individual data are considered as of sufficient quality in the laboratory if the call rate exceeds 90%.

A good quality indicator is the heterozygosity across all SNPs. For example, if the heterozygosity is too high, this may be an indicator of DNA contamination. A typical approach is to estimate the mean (m) and standard deviation (SD) of the heterozygosity across all study subjects and to exclude all subjects outside of $m \pm 3$ SD. A common alternative is to identify and exclude outliers based on the histogram of heterozygosities.

A third criterion uses the relatedness of the subjects. Specifically, for a random pair of study subjects there is a 50% probability that a randomly chosen allele is identical in state. A substantial increase in this probability is an indicator of relatedness, and such subjects should either be excluded from a case-control study or appropriate adjustments should be performed.

3.8 Investigation of population stratification

Population stratification, that is, confounding by ethnicity, can lead to a substantial inflation of test statistics. It is therefore important either to demonstrate that population stratification is negligible in the data or to adequately adjust for population stratification. For example, Steffens et al. (2006) showed that there is only a minor degree of population substructure in the German population, and it is too low to be detectable from methods without using prior information of subpopulation membership (see, e.g., the structure approach of Pritchard, Stephens and Donnelly, 2000). If in addition, sampling is restricted to a small geographical region within Germany, like Bavaria, which is bounded to the South by the Alps, even greater population homogeneity can be expected (see, e.g., supplementary material of Samani et al., 2007).

If population substructure has to be taken into account, it might be possible to perform a stratified analysis using reported geographic location or ethnicity. Alternatively, one can estimate the extent of population heterogeneity. As the false positive fraction is expected to be increased when population stratification is ignored, one can compare the median over all test statistics with the theoretical median of the distribution of the test statistic under the null hypothesis (Devlin and Roeder, 1999). For illustration, a Q–Q-plot of all test statistics can be generated showing the degree of inflation of test statistics, and it has become standard to include the Q–Q-plot in the presentation of GWAs. If population substructure is present and there is no geographical information about the subjects, adjustment can be performed using genomic control (GC, Devlin and Roeder, 1999), structured association (Pritchard et al., 2000) or eigenstrat (Price et al., 2006). When possible it is sensible to restrict the Q–Q-plot and the GC approach to SNPs that are not associated with the trait of interest. Thus, there is a clear need for defining an appropriate set of GC SNPs. We note that the structure approach requires knowledge about the number of clusters in the population. Both approaches have been employed, e.g., by Sladek et al. (2007).

4 Frequentist and Bayesian Approaches to Analyses

All applied papers on GWAs start by reporting results from single marker tests of SNPs (Section 4.1). One reason for this is the computational burden in analysing hundreds of thousands of explanatory variables. Analyses using haplotypes (Section 4.2) and interactions (Section 4.3) are even more challenging and currently are therefore only used at later stages. We end by describing machine learning approaches for analysing GWAs (Section 4.4).

4.1 Single marker analyses

Although single marker tests of SNPs represent the standard first approach to GWAs, the statistical test used for this primary analysis varies substantially (a detailed description of the different test statistics can be found in Ziegler and König, 2006). For example, Klein et al. (2005) and McPherson et al. (2007) tested for allelic association in their case-control analysis. As nicely explained by Sasieni

(1997), this test is only valid if HWE holds. Other authors therefore prefer to use the 1 degree of freedom Cochran-Armitage trend test, see, e.g., Samani et al. (2007) and The Wellcome Trust Case Control Consortium (2007). This test has reasonable power under all genetic models. An alternative is to use a two degree of freedom test or to test for the three standard genetic models, i.e., dominant, additive and recessive, to take the maximum of these and adjust for the three tests. This adjustment can be done by means of permutations (Freidlin et al., 2002) as utilized, e.g., by Sladek et al. (2007), or by using a conditional test, in which the correlation of the three test statistics is taken into account (Hothorn and Hothorn, unpublished). The latter approach might be preferable because the computational demands of permutation procedures can become excessive in GWAs.

Having fixed the primary analysis, the next important question is what can be declared significant in GWAs. This question is not novel, and the first suggestions trace back to Risch and Merikangas (1996), when GWAs were only remotely conceivable. A simple answer to this question would be to assume that the SNPs on the selected GWA panel are independent of each other. Then, if an array with 500 000 SNPs has been selected, the local significance level is 10^{-7} for a global significance level of 5% when the Bonferroni correction is applied. This simple approach has been used, e.g., by Klein et al. (2005) but it is conservative as it inappropriately assumes independence of SNPs on a SNP array. Therefore, some researchers have proposed using weaker local significance levels, like 10^{-6}, 10^{-5} or even 10^{-4} (Arking et al., 2006), in order to account for LD. A possible method, suggested by Zondervan and Cardon (2007), is to assume a certain constant LD and so adjust the effective number of tests. Another tractable way to estimate the effective number of tests is to re-sample from available empirical data under the null hypothesis. Based on the ENCODE data, The International HapMap Consortium proposed a local significance level of 5.5×10^{-8} (The International HapMap Consortium, 2005). A similar conclusion was reached using a permutation approach based on the data from The Wellcome Trust Case Control Consortium (Gusnanto and Dudbridge, 2007).

These suggestions focus on the evaluation of a single SNP only. However, if one SNP shows an association with the disease of interest and if this SNP is in strong LD with another SNP on the array, the other SNP should also be associated with the disease. Thus, it might be plausible to claim that a chromosomal region is interesting if two or more SNPs in modest to strong LD have nominal p-values below an even weaker threshold. Some research groups go even further and avoid formal significance thresholds at all. Instead, they order the results according to p-values and explore the SNPs with the lowest p-values first (Helgadottir et al., 2007).

This illustrates a fundamental difference between GWAs and, for instance, clinical trials with regard to the importance of false positive results: In clinical trials, a false positive finding may lead to patients being treated with an ineffective or even detrimental therapy; in GWAs, a false positive result merely increases the cost of following up the initial findings. More generally, a phase III clinical trial represents an endpoint and has direct clinical consequences, whereas a GWA is one element in the chain of genetic research.

Furthermore, there is the usual duality between type I and type II errors. And it is important to note that with the current sample sizes in use, i.e., approximately 1000 cases and 1000 controls per GWA, using a local type I error level of 10^{-6} and assuming a risk allele frequency of 0.2, the power is approximately 4% to detect an odds ratio (OR) of 1.3, which is a realistic assumption for complex diseases (see, e.g., Samani et al., 2007).

One might argue that improved procedures allowing for multiple testing of a large number of correlated tests might solve this problem. Lin (2005) developed a Monte-Carlo approach, modified by Dudbridge (2006), to approximating the joint distribution of test statistics along the entire genome. Although these approaches can be substantially more powerful than standard Bonferroni or Holm type methods, the computational effort is high. To reduce this, a number of approaches such as importance sampling (Kimmel and Shamir, 2006) could be used. Alternatively, one might perform a few permutations and then fit an extreme value distribution to the results for extrapolating the tail of the distribution. At the same time, the correlation of the test statistics may be taken into account by conditioning on the correlation structure (Dudbridge and Koeleman, 2004). Another approach would be to use joint

tests of groups of SNPs to avoid a multiple-testing penalty for individual SNP tests. For example, one could jointly test SNPs of groups of genes (Goeman et al., 2004) or form the sum or the product of a statistic over sets of SNPs (Hoh and Ott, 2003). However, these approaches have the disadvantage that it does not permit identifying the most promising individual SNPs (Balding, 2006).

One might also argue that the false discovery rate (FDR), which allows a controlled proportion of positive results to be false (Benjamini and Hochberg, 1995; Dudbridge, Gusnanto and Koeleman, 2006) would be more appropriate than the family-wise error rate (for a detailed discussion of these terms, see, e.g., Bretz, Landgrebe and Brunner, 2005). The FDR is often applied in gene expression studies where a fairly large proportion of tested null hypotheses are false. In GWAs, however, many more hypotheses and, even more important, many more false hypotheses are tested. Therefore, we cannot expect a remarkable difference between the FDR and the family-wise error rate in GWAs (Benjamini and Hochberg, 1995). Furthermore, the FDR is difficult to interpret for single SNPs or genes because only an expected value is controlled for.

These two approaches thus do not overcome the fundamental problem that formal statistical testing often is not the primary aim in a GWA. Instead, researchers are more interested in summarizing the available information. Furthermore, when a test is performed, it is of greater interest to know whether a SNP is worthy of further investigation.

The unsatisfactory nature of traditional approaches to multiple testing and to defining genome-wide significance has led some researchers to suggest the use of Bayesian alternatives. Bayes theorem tells us that

$$P(H_0 \mid D) = \frac{P(D \mid H_0)\,\pi_0}{P(D \mid H_0)\,\pi_0 + P(D \mid H_1)\,(1-\pi_0)}\,,$$

where H_0 and H_1 are alternative hypotheses, D represents the data and π_0 is the prior probability of H_0. Writing the corresponding formula for $P(H_1 \mid D)$ and taking the ratio leads to

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_0)}{P(D \mid H_1)}\,\frac{\pi_0}{1-\pi_0}\,.$$

The ratio $P(D \mid H_0)/P(D \mid H_1)$ is called the Bayes factor and measures the impact of the data on the support for H_0 in preference to H_1. Its use in genetic epidemiology, as an alternative to p-values, has been advocated by The Wellcome Trust Case Control Consortium (2007), and by Wakefield (2007). The interpretation of a Bayes factor is not dependent on whether other hypotheses are evaluated and so does not need an adjustment for multiple comparisons. The main drawbacks of Bayes factors are that they can be difficult to calculate when the hypotheses are complex, there is no generally accepted threshold for declaring preference for one hypothesis over another, and the Bayes factor can be very sensitive to the choice of priors on the parameters of the hypotheses (Kass and Raftery, 1995).

Wacholder et al. (2004) used Bayes theorem in the situation in which it is known that the data are significant at the $\alpha\%$ test level and there is a single point alternative H_1. In those circumstances,

$$P(H_0 \mid D) = \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)\,(1-\pi_0)}\,,$$

where $1-\beta = P(D \mid H_1)$ is the power of the test and the corresponding Bayes factor is $(1-\beta)/\alpha$. Wacholder et al. called $P(H_0 \mid D)$ the False Positive Report Probability (FPRP). This simple but important formula illustrates how the interpretation of data depends not only of the significance level, but also on the power of the study and on our prior assessment of the probability that H_0 holds. It is the fact that the p-value only captures part of the story that is at the root of many of its problems.

The FPRP applies when the only information that we have is that the test statistic is significant at the $\alpha\%$ level and as such it might be very helpful when planning a study. However, once the data are collected we need to condition on the exact p-value and not merely on the knowledge that we have significance at the $\alpha\%$ level (Thomas and Clayton, 2004). Once the data are collected, $P(D \mid H_1)$ and

$P(D \mid H_0)$ are no longer tail areas but instead are values of probability density and the formula for calculating $P(H_0 \mid D)$ needs to be changed accordingly. This point was noted by Wakefield (2007) who described the resulting calculation as a Bayesian False Discovery Probability (BFDP) and by Samani et al. (2007) who called it a modified FPRP. As Bayes factors and posterior probabilities can be time-consuming to evaluate when the hypotheses depend on several parameters, it may not be practical to perform a full Bayesian analysis of a GWA. A sensible alternative is to perform simpler Bayesian calculations using more tightly specified hypotheses and only for SNPs that appear interesting to follow this with sensitivity analyses and calculations for the fuller hypotheses. This was the approach described in the supplementary material of Samani et al. (2007).

A plausible estimate for the prior odds of true association at any specified locus might be of the order of 100 000 : 1 against, for example, on the basis of 1 000 000 'independent' regions of the genome and an expectation of 10 detectable genes involved in the condition. It is thus possible to perform a Bayesian analysis using a uniform a priori distribution. Other plausible estimates might vary from this by an order of magnitude in either direction. Then, assuming a power of 0.5 and a significance threshold of 5×10^{-7}, the posterior odds in favour of a 'hit' being a true association would be 10 : 1 (The Wellcome Trust Case Control Consortium, 2007, Supplementary Material).

4.2 Haplotype analyses

As described above, the main statistical analyses of a GWA are univariate tests of single SNPs. Based on these results various in-depth analyses may then be performed.

A standard approach is the *haplotype* analysis of SNPs in a chromosomal region of interest. There are several reasons why haplotypes are interesting (for overviews see, e.g., Clark, 2004; Ziegler and König, 2006), and a thorough discussion has been given by, e.g., Clayton, Chapman and Cooper (2004). First, haplotypes are biologically relevant. For example, Fitze et al. (2003) investigated influences of, and interactions between, mutations in the *RET* proto-oncogene (OMIM *164761) and polymorphisms on the development of Morbus Hirschsprung (OMIM #142623). In patients with a more severe and a milder form, they found that the $-5\,\mathrm{G} > \mathrm{A}$ polymorphism modulated the phenotypic influences of mutations in the RET gene. Because the -5A variant was associated with the milder form when it was on the same chromosome as the mutation, it seems that the -5A allele has a protective effect, and the RET haplotype seems to influence disease severity. Second, one may say that variations in populations are inherently structured in haplotypes so that these represent the natural unit for analyses. Most importantly, even in GWAs we cannot assume that the disease causing variant is included on the array. Therefore, haplotype analyses can have greater power than single marker analyses (Allen and Satten, 2007).

Because haplotypes are rarely directly observed, comparisons of haplotype frequencies between cases and controls need to be based on estimates of the frequencies. In practice, a number of algorithms are available for this (Epstein and Satten, 2003; Schaid et al., 2002; Stram et al., 2003; Tregouet et al., 2002), and usually a likelihood ratio or a score test is employed for comparing cases and controls. Here, one has to be careful because the data will become sparse with an increasing number of different haplotypes so that the standard approximation to the χ^2 distribution might be inappropriate. Therefore, many investigators prefer using a permutation approach (Schaid et al., 2002). Another limitation is that many haplotype estimation algorithms rely on the assumption of HWE, a point discussed in detail by, e.g., Schaid (2004); see Wellek (2004) for an equivalence test for HWE. Extensions of this basic approach to haplotype analysis are typically based on the framework of generalized linear models and allow for the inclusion of further variables such as environmental covariates and interactions between environmental effects and haplotypes (Epstein and Satten, 2003; Kwee et al., 2007; Schaid et al., 2002; Spinka, Carroll and Chatterjee, 2005; Tregouet and Garelle, 2007).

The haplotype approaches are based on the assumption that, in the vicinity of a predisposing functional variant, haplotypes carrying this variant (case haplotypes) are more related than haplotypes not

carrying the variant (random haplotypes). Therefore, the expectation is that the case haplotypes share significantly longer stretches of DNA around the variant (te Meerman, van der Meulen and Sandkuijl, 1995), and this is the basic idea of haplotypes sharing approaches and cladistic trees (Beckmann et al., 2005; de Andrade and Allen, 2007). Although the power of statistical methods based on haplotype sharing depends strongly on the recombinational and mutational history of the underlying population, it can be dramatically higher than single locus tests (Allen and Satten, 2007). Furthermore, it might be possible to better localize the functionally relevant gene or even a functionally variant than with single locus analyses (for an interesting application, see, e.g., Foerster et al., 2005).

Many of the above-mentioned approaches are not tailored for an analysis of the entire array or sliding windows of haplotypes because of the computational burden, and computational feasability is an important area of current research. For example, Lin et al. (2004) proposed exhaustive testing of haplotype associations over all possible windows of segments, using a computationally efficient permutation procedure to assess the significance of these correlated tests.

Another area of active research based on haplotypes that has received great attention in the past few months are imputation approaches allowing to test untyped variants by coupling typed SNPs with external information from databases describing LD patterns across the genome (Abecasis, 2007; Epstein, Allen and Satten, 2007; Marchini et al., 2007; Servin and Stephens, 2007). These approaches are important for pooled analyses of multiple GWAs for identifying common alleles with small effects. Although it is possible to use SNPs that tag the variant of interest well as proxies for these analyses, the analysis of all SNPs available in a database might be less cumbersome if these have been imputed.

4.3 Gene-gene and gene-environment interactions

Empirical evidence supports the idea that the interplay of genes and environmental factors affects many common diseases (Kraft et al., 2007). A classical example of *epistasis*, a special case of gene-gene *interaction*, is the coat colour in the Labrador retriever (Templeton, Stewart and Fletcher, 1977). These dogs are typically black, brown or yellow, and this is determined by the tyrosinase related protein 1 (TYRP1; usually termed the brown locus) and the melanocortin 1 receptor gene (MC1R; usually termed the extension locus). Dogs that are homozygous ee at the extension locus are yellow irrespective of their genotype at the brown locus. If, in contrast, the Labrador carries at least one E allele, coat colour depends on the genotype at the brown locus. Specifically, if a dog is homozygous bb at the brown locus, its coat colour is brown, otherwise, it is black. An example of epistasis in humans that has received considerable attention and that has been questioned is the interplay of the DLG5 and CARD15 genes in Crohns disease (Stoll et al., 2004).

An interesting example of gene-environment interaction, possibly without a marginal gene effect is the interplay of exposure to pesticides and the C3435T polymorphism of the MDR1 gene with regard to Parkinsons disease (Drozdzik et al., 2003; Furuno et al., 2002).

These examples demonstrate the possible importance of interactions. However, its investigation in GWAs poses a challenge with regard to both multiple testing and computational burden. For instance, if we consider 300 000 SNPs, the number of SNP pairs to be analyzed is close to 100 billion, and this figure does not even include higher order interactions. Interestingly, though, Marchini et al. (2005) showed that the exhaustive testing of all possible pairwise gene-gene interactions is a computationally feasible undertaking in GWAs given a large computer cluster. Nevertheless, a number of open issues remain. First, there usually is no a priori hypothesis about the specific interaction model, leaving us with four degrees of freedom for the main effects and the same number of degrees of freedom for the interactions. And in contrast to single SNP analyses, it is unclear how the interaction should be modelled with fewer degrees of freedom to maintain good power under a wide variety of genetic interaction models. Second, even if there is a way to adequately model gene-gene interactions, the power to detect these is likely to be fairly low.

When studying gene-environment interactions, the complexity of the model is determined by the measurement scale of the environmental factor. In contrast to gene-gene interactions, the situation can be simplified by using the case-only study design. As pointed out by many authors, this needs, however, to be carried out with care (see, e.g., Schmidt and Schaid, 1999). One caveat is that subtle undetected gene-environment correlations can greatly bias case-only interaction estimates (Kraft et al., 2007).

All these examples underline the importance of studying gene-gene and gene-environment interactions but they also demonstrate that no clear recommendation on how to handle these in GWAs can be given as yet and that further work is required in this area.

4.4 Machine learning approaches for analysing genome-wide association studies

In addition to the frequentist and Bayesian approaches for analysing GWAs discussed above, a number of machine learning procedures are available which may be adapted for this scenario (König et al., 2008; Ziegler et al., 2007). A number of attempts in this direction have been made, and, to give an example, at the latest Genetic Analysis Workshop in 2006, several of these including random forests (Breiman, 2001), logistic trees with unbiased selection (LOTUS, Chan and Loh, 2004), and logic regression (Kooperberg et al., 2001) were applied to the analysis of simulated and real GWA data sets. As primary advantages, these essentially non-parametric methods can better handle the situation of many SNPs and additional variables being assessed in relatively small samples. They are often less prone to overfitting and less computationally intensive than conventional tests, although not all current available implementations keep this promise. These methods claim to be able to handle higher-order interactions as well as pairwise effects. Depending on the method employed, however, this might again prove to be problematic. For instance, the detection of a SNP-SNP interaction using random forests requires that these two SNPs have been selected jointly in a sufficient number of trees. To give a simple example, if we grow a random forest with $\sqrt{500\,000} \approx 700$ SNPs per tree, the probability that two specific SNPs occur in the same tree approximates 2×10^{-6}, so that at least 500 000 trees have to be grown in the forest to make sure that this combination is expected to be represented at least once. This makes it computationally challenging.

One further interesting potential of machine learning algorithms pertains to the aspired aim of the study. As described above, the typical aim of a GWA is to detect differences in genotype frequencies between cases and controls, and thus to identify possible candidate SNPs in interesting genetic regions. Another aim, though, may be to predict the disease status or, more generally, the phenotype of a proband as correctly as possible based on a set of SNPs (Ziegler et al., 2007). This can only indirectly be achieved by classical methodology, but usually is the explicit target of machine learning procedures. In conclusion, it seems that for GWAs, machine learning methods may be useful additional tools for data mining, as well as for parts of gene identification and disease prediction.

5 Accumulating the Evidence in Multistage Procedures

The use of multistage procedures has always been common practice in genetic epidemiological studies (Böddeker and Ziegler, 2001). Using a GWA as the first step in a multistage procedure is attractive for a number of reasons. The first motivation, which is typical of classical biometrical applications, is to decrease the required sample sizes or, more generally, the cost of the study. To this end, several suggestions have been made for adapting formal group sequential approaches to the specific situation of GWAs (Lin, 2006; Müller, Pahl and Schäfer, 2007; Satagopan and Elston, 2003; Satagopan, Venkatraman and Begg, 2004; Wang et al., 2006). In which sense are these classical designs adapted? The proposed designs differ in three aspects from the typical group sequential clinical trials,. Firstly, the primary aim in clinical trials usually is to stop a trial early in case of significance. The aim in GWAs,

however, typically is to extract those SNPs in the first stage with promising effects and avoiding expenditure on SNPs that have no prospect of reaching significance (König and Ziegler, 2003). Secondly, stopping early does not result in the termination of the entire study, but merely in a focus on those promising SNPs. These are then genotyped in additional samples of cases and controls, and test statistics are combined from both stages. The final specific feature of sequential designs for GWAs is that minimizing the sample size cannot be equated with minimizing the costs. Indeed, even minimizing the number of genotypings in the entire study does not minimize the costs, because genotyping in the first stage using an available chip is much less expensive than later genotyping of fewer selected SNPs.

These aspects are considered, so that a cost efficient two-stage design can be specified. To give an example, using the design suggested by Wang et al. (2006), upon specification of the power (90%), the overall significance level (one-sided 10^{-7}), the variant allele frequency (0.2), the odds ratio (1.35), the cost per genotype in the first stage (0.2 cents) and in the second stage (3.5 cents), the nominal significance levels for both stages is derived (0.00373 and 1.7×10^{-7}) as well as the required sample sizes for both stages (1920 and 4458). Other proposed designs additionally take into account the cost of phenotyping (Müller et al., 2007) or the dependency structure between SNPs (Lin, 2006).

To date, we are not aware of any GWA having actually implemented a formal group sequential design despite the obvious potential to save costs. To speculate, this may be explained by the difficulty of specifying the effect sizes and allele frequencies ahead of the study. What is more, a universal effect and frequency needs to be defined, but in practice, the values are likely to differ remarkably between the SNPs. Also, the research team may be reluctant to base the decision to carry a SNP forward merely on the value of the single SNP test statistic or *p*-value instead of taking an entire region into account.

A second reason for a multistage procedure is that one may wish to follow up a result in more detail by increasing the marker density in interesting regions. In practice, this is either done in a separate analysis (e.g., Nam et al., 2007) or in concert with a group sequential design as described above (Wang et al., 2006).

Finally, it has always been emphasized that results from genetic association studies need to be confirmed in independent samples to rule out systematic effects like genotyping error or *population stratification*. This pertains even more to GWAs, where there is rarely a strict control of the type I error and thus random positive findings also need to be excluded. Therefore, after a first GWA, improving the credibility of a result by confirming it in additional studies or analyses is desirable. Here, the simplest case with many examples is that SNPs having shown association in a GWA are genotyped in an independent sample either within a second GWA (Samani et al., 2007) or in a small-scale confirmation study (e.g., Herbert et al., 2006; Hunter et al., 2007). Overall, it seems to have become standard practice to include confirmation results from independent samples in the publication of the original GWA (also see NCI-NHGRI Working Group on Replication in Association Studies, 2007). When a number of studies have been performed using the same SNPs, their results can be combined in meta-analyses. These provide increased power, help in identifying false positives due to biases in individual studies and highlight heterogeneity, which may itself be interesting (Ioannidis, Patsopoulos and Evangelou, 2007). To do meta-analyses without publication bias requires access to the full set of results from each GWA, and this provides a strong argument for appropriate data sharing policies of GWAs data. Researchers are encouraged to discuss this point with ethics committees, data protection agencies and institutional review boards (IRBs) when a study is initiated. This issue might be even more complicated when additional information, such as family structure and segregation of a disease in families is collected. This is in contrast to the policy by the *National Institutes of Health* (NIH) (specifically, see notice numbers NOT-OD-08-013 and NOT-OD-07-088) and supports the position of the *International Genetic Epidemiology Society* that original study investigators should be allowed to maintain the data in non-Federal databases to minimize re-identification and other risks for misuse. Even the authors of this article do not agree in whether data should automatically be released publicly.

6 The Future: Whole Genome Sequencing

The next challenges already knock at our doors. Whole-genome sequencing, i.e., the sequencing of the equivalent of an entire human genome for $ 1000 has been announced as a goal for the genetics community, and new sequencing technologies suggest that reaching this goal is a matter of when, rather than if (see the *Nature Genetics* "Question of the Year"). This will replace the SNP typing, thus the indirect approach for the identification of disease genes and genetic risk factors. Specifically, it will be economically feasible to screen entire genes, which are likely to harbour functionally relevant DNA variants, or even chromosomes allowing studying effects of non-coding genetic regions (Ropers, 2007). As a consequence, biostatisticians have to be prepared for the analysis of several millions of independent variables at a time because the length of the entire human genome is $\sim 3.2 \times 10^9$.

At the same time, massive data from other sources will be brought in. For example, the first genome-wide association study of global gene expression has recently been published including ~400 000 SNPs and ~55 000 transcripts representing ~20 000 genes (Dixon et al., 2007). Thus, a combination of genomic data with transcriptomic data and proteomic data together with the metabolic response to drugs will be an area of research requiring our expertise.

We are convinced that the work at these cutting edges of interdisciplinary research will be fascinating, and we look forward to the prosperous future of biostatistics.

Acknowledgements Grants from the European Union for funding the Cardiogenics Consortium and from the Bundesministerium für Bildung und Forschung are gratefully acknowledged.

URLs

HapMap project: http://www.hapmap.org

International Genetic Epidemiology Society Position Statement: http://www.geneticepi.org/elsi/PositionStatement_IGES_GWAS_RFI-11-17-06.doc

Nature Genetics "Question of the Year": http://www.nature.com/ng/qoty/index.html

National Institutes of Health Policy for Sharing GWAs Data: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-013.html, /NOT-OD-07-088.html

OMIM: Online Mendelian Inheritance in Man. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), http://www.ncbi.nlm.nih.gov/omim/

References

Abecasis, G. R. (2007). Turning a flood of data into a deluge: "in silico" genotyping for genome-wide association scans. *Genetic Epidemiology* **31**, 653.

Affymetrix (2006). GeneChip® Mapping 500 K Assay Manual. Affymetrix, Santa Clara (CA). http://www.affymetrix.com/support/downloads/manuals/500k_assay_manual.pdf

Allen, A. S. and Satten, G. A. (2007). Statistical models for haplotype sharing in case-parent trio data. *Human Heredity* **64**, 35–44.

Arking, D. E., Pfeufer, A., Post, W., Kao, W. H., Newton-Cheh, C., Ikeda, M., West, K. et al. (2006). A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nature Genetics* **38**, 644–651.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**, 781–791.

Beckmann, L., Ziegler, A., Duggal, P., and Bailey-Wilson, J. E. (2005). Haplotypes and haplotype-tagging single-nucleotide polymorphism: Presentation Group 8 of Genetic Analysis Workshop 14. *Genetic Epidemiology* **29**, S59–71.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Böddeker, I. R. and Ziegler, A. (2001). Sequential designs for genetic epidemiological linkage or association studies. A review of the literature. *Biometrical Journal* **43**, 501–525.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.
Bretz, F., Landgrebe, J., and Brunner, E. (2005). Multiplicity issues in microarray experiments. *Methods of Information in Medicine* **44**, 431–437.
Chan, K. and Loh, W. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computional and Graphical Statistics* **13**, 826–852.
Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology* **27**, 321–333.
Clayton, D., Chapman, J., and Cooper, J. (2004). Use of unphased multilocus genotype data in indirect association studies. *Genetic Epidemiology* **27**, 415–428.
Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J. et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* **37**, 1243–1246.
Colhoun, H. M., McKeigue, P. M., and Davey Smith, G. (2003). Problems of reporting genetic associations with complex outcomes. *The Lancet* **361**, 865–872.
Cordell, H. J. (2002). Epistasis: what it means, what it doesnt mean, and statistical methods to detect it in humans. *Human Molecular Genetics* **11**, 2463–2468.
de Andrade, M. and Allen, A. S. (2007). Summary of contributions to GAW 15 group 13: candidate gene association. *Genetic Epidemiology* **31**, 110–117.
Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997–1004.
Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., Taylor, J. et al. (2007). A genome-wide association study of global gene expression. *Nature Genetics* **39**, 1202–1207.
Drozdzik, M., Bialecka, M., Mysliwiec, K., Honczarenko, K., Stankiewicz, J., and Sych, Z. (2003). Polymorphism in the P-glycoprotein drug transporter MDR1 gene: a possible link between environmental and genetic factors in Parkinsons disease. *Pharmacogenetics* **13**, 259–263.
Dudbridge, F. (2006). A note on permutation tests in multistage association scans. *American Journal of Human Genetics* **78**, 1094–1095; author reply 1096.
Dudbridge, F., Gusnanto, A., and Koeleman, B. P. (2006). Detecting multiple associations in genome-wide studies. *Human Genomics* **2**, 310–317.
Dudbridge, F. and Koeleman, P. C. (2004). Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *American Journal of Human Genetics* **75**, 424–435.
Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., Struewing, J. P. et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093.
Epstein, M. P., Allen, A. S., and Satten, G. A. (2007). Efficient and flexible testing of untyped variants in case-control studies [abstract 30]. *Annual Meeting of The American Society of Human Genetics, October 25, 2007*, San Diego (CA), 40. http://www.ashg.org/genetics/ashg06s/index.shtm
Epstein, M. P. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* **73**, 1316–1329.
Fitze, G., Appelt, H., König, I. R., Görgens, H., Stein, U., Walther, W., Gossen, M. et al. (2003). Functional haplotypes of the RET proto-oncogene promoter are associated with Hirschsprung disease (HSCR). *Human Molecular Genetics* **12**, 3207–3214.
Foerster, J., Nolte, I., Junge, J., Bruinenberg, M., Schweiger, S., Spaar, K., van der Steege, G. et al. (2005). Haplotype sharing analysis identifies a retroviral dUTPase as candidate susceptibility gene for psoriasis. *Journal of Investigative Dermatology* **124**, 99–102.
Freidlin, B., Zheng, G., Li, Z. H., and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Human Heredity* **53**, 146–152.
Furuno, T., Landi, M. T., Ceroni, M., Caporaso, N., Bernucci, I., Nappi, G., Martignoni, E. et al. (2002). Expression polymorphism of the blood-brain barrier component P-glycoprotein (MDR1) in relation to Parkinsons disease. *Pharmacogenetics* **12**, 529–534.
Goeman, J. J., de Kort, F., van de Geer, S. A., and van Houwelingen, J. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93–99.
Gusnanto, A. and Dudbridge, F. (2007). Estimating significance thresholds for genomewide association scans. *Genetic Epidemiology* **31**, 629.
Hartmann, O. (2005). Quality control for microarray experiments. *Methods of Information in Medicine* **44**, 408–413.
Hattersley, A. T. and McCarthy, M. I. (2005). What makes a good genetic association study? *The Lancet* **366**, 1315–1323.

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., Jonasdottir, A. et al. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491–1493.

Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A., Illig, T., Wichmann, H. E. et al. (2006). A common genetic variant is associated with adult and childhood obesity. *Science* **312**, 279–283.

Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics* **4**, 701–709.

Hua, J., Craig, D. W., Brun, M., Webster, J., Zismann, V., Tembe, W., Joshipura, K. et al. (2007). SNiPer-HD: Improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics* **23**, 57–63.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S. et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics* **39**, 870–874.

Ioannidis, J. P., Patsopoulos, N. A., and Evangelou, E. (2007). Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE* **2**, e841.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–794.

Kimmel, G. and Shamir, R. (2006). A fast method for computing high-significance disease association in large population-based studies. *American Journal of Human Genetics* **79**, 481–492.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A. K. et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389.

König, I. R., Malley, J. D., Pajevic, S., Weimar, C., Diener, H.-C., Ziegler, A., and on behalf of the German Stroke Study Collaborators (2008). Patient-centered yes/no prognosis using learning machines. *International Journal of Data Mining and Bioinformatics*, in press.

König, I. R. and Ziegler, A. (2003). Group sequential study designs in genetic-epidemiological case-control studies. *Human Heredity* **56**, 63–72.

Kooperberg, C., Ruczinski, I., LeBlanc, M. L., and Hsu, L. (2001). Sequence analysis using logic regression. *Genetic Epidemiology* **21 Suppl 1**, S626–631.

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., and Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity* **63**, 111–119.

Kwee, L. C., Epstein, M. P., Manatunga, A. K., Duncan, R., Allen, A. S., and Satten, G. A. (2007). Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. *Genetic Epidemiology* **31**, 75–90.

Lin, D. Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **6**, 781–787.

Lin, D. Y. (2006). Evaluating statistical significance in two-stage genomewide association studies. *American Journal of Human Genetics* **78**, 505–509.

Lin, S., Chakravarti, A., and Cutler, D. J. (2004). Exhaustive allelic transmission disequilibrium tests as a new approach to genomewide association studies. *Nature Genetics* **36**, 1181–1188.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413–417.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913.

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., Hinds, D. A. et al. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491.

Müller, H. H., Pahl, R. and Schäfer, H. (2007). Including sampling and phenotyping costs into the optimization of two stage designs for genome wide association studies. *Genetic Epidemiology* **31**, 844–852

Nam, R. K., Zhang, W. W., Loblaw, D. A., Klotz, L. H., Trachtenberg, J., Jewett, M. A., Stanimirovic, A. et al. (2007). A genome-wide association screen identifies regions on chromosomes 1q25 and 7p21 as risk loci for sporadic prostate cancer. *Prostate Cancer and Prostatic Disorders*,

NCI-NHGRI Working Group on Replication in Association Studies (2007). Replicating genotype–phenotype associations. What constitutes replication of a genotype–phenotype association, and how best can it be achieved? *Nature* **447**, 655–660.

Nicolae, D. L., Wu, X., Miyake, K., and Cox, N. J. (2006). GEL: a novel genotype calling algorithm using empirical likelihood. *Bioinformatics* **22**, 1942–1947.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Rabbee, N. and Speed, T. P. (2006). A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22**, 7–12.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Ropers, H. H. (2007). New perspectives for the elucidation of genetic disorders. *American Journal of Human Genetics* **81**, 199–207.

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R. J. et al. (2007). Genome-wide association analysis of coronary artery disease. *The New England Journal of Medicine* **357**, 443–453.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

Satagopan, J. M. and Elston, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genetic Epidemiology* **25**, 149–157.

Satagopan, J. M., Venkatraman, E. S., and Begg, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60**, 589–597.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology* **27**, 348–364.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425–434.

Schmidt, S. and Schaid, D. J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology* **150**, 878–885.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* **3**, e114.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885.

Spinka, C., Carroll, R. J., and Chatterjee, N. (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* **29**, 108–127.

Steffens, M., Lamina, C., Illig, T., Bettecken, T., Vogler, R., Entz, P., Suk, E.-K. et al. (2006). SNP-based analysis of genetic substructure in the German population. *Human Heredity* **62**, 20–29.

Stoll, M., Corneliussen, B., Costello, C. M., Waetzig, G. H., Mellgard, B., Koch, W. A., Rosenstiel, P. et al. (2004). Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nature Genetics* **36**, 476–480.

Stram, D. O., Leigh Pearce, C., Bretsky, P., Freedman, M., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N. et al. (2003). Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity* **55**, 179–190.

te Meerman, G. J., van der Meulen, M. A., and Sandkuijl, L. A. (1995). Perspectives of identity by descent (IBD) mapping in founder populations. *Clinical and Experimental Allergy* **25 Suppl 2**, 97–102.

Templeton, J. W., Stewart, A. P., and Fletcher, W. S. (1977). Coat color genetics in the Labrador retriever. *Journal of Heredity* **68**, 134–136.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

The News Staff (2006). Breakthrough of the year. The runners-up. *Science* **314**, 1850–1855.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. *Nature* **447**, 661–678.

Thomas, D. C. and Clayton, D. G. (2004). Betting odds and genetic associations. *Journal of the National Cancer Institute* **96**, 421–423.

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S. et al. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genetics* **39**, 984–988.

Tregouet, D. A., Barbaux, S., Escolano, S., Tahri, N., Golmard, J. L., Tiret, L., and Cambien, F. (2002). Specific haplotypes of the P-selectin gene are associated with myocardial infarction. *Human Molecular Genetics* **11**, 2015–2023.

Tregouet, D. A. and Garelle, V. (2007). A new JAVA interface implementation of THESIAS: testing haplotype effects in association studies. *Bioinformatics* **23**, 1038–1039.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute* **96**, 434–442.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics* **81**, 208–227.

Wang, H., Thomas, D. C., Peer, I., and Stram, D. O. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genetic Epidemiology* **30**, 356–368.

Watkins, H. and Farrall, M. (2006). Genetic susceptibility to coronary artery disease: from promise to progress. *Nature Reviews Genetics* **7**, 163–173.

Wellek, S. (2004). Tests for establishing compatibility of an observed genotype distribution with Hardy–Weinberg equilibrium in the case of a biallelic locus. *Biometrics* **60**, 694–703.

Wittke-Thompson, J. K., Pluzhnikov, A., and Cox, N. J. (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 967–986.

Zanke, B. W., Greenwood, C. M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S. M., Prendergast, J. et al. (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature Genetics* **39**, 989–994.

Ziegler, A., DeStefano, A. L., König, I. R., and on behalf of Group 6 (2007). Data mining, neural nets, trees – Problems 2 and 3 of Genetic Analysis Workshop 15. *Genetic Epidemiology* **31**, 51–60.

Ziegler, A. and König, I. R. (2006). *A Statistical Approach to Genetic Epidemiology. Concepts and Applications*. Wiley-VCH, Weinheim.

Zondervan, K. T. and Cardon, L. R. (2007). Designing candidate gene and genome-wide case-control association studies. *Nature Protocols* **2**, 2492–2501.

