Published in final edited form as:
Genet Epidemiol. 2009 ; 33(Suppl 1): S45–S50. doi:10.1002/gepi.20472.

Genome-Wide Association Studies: Quality Control and Population-Based Measures

Andreas Ziegler[1]
[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Germany

Abstract

Genome-wide association studies using hundreds of thousands of single-nucleotide polymorphism (SNP) markers have become a standard approach for identifying disease susceptibility genes. The change in the technology poses substantial computational and statistical challenges that have been addressed in the quality control, imputation, and population-based measure groups of the Genetic Analysis Workshop 16. The computational challenges pertain to efficient memory management and computational speed of the statistical procedures, and we discuss an approach for efficient SNP storage. Accuracy and computational speed is relevant for genotype calling, and the results from a comparison of three calling algorithms are discussed. The first statistical challenge is related to statistical quality control, and we discuss two novel quality control procedures. These low-level analyses have an effect on subsequent preparatory steps for high-level analyses, e.g., the quality of genotype imputation approaches. After the conduct of a genome-wide association study with successful replication and/or validation, measures of diagnostic accuracy including the area under the curve are investigated. The area under the curve can be constructed from summary data in some situations. Finally, we discuss how the population-attributable risk of a genetic variant that is only measured in a reference data set can be determined.

Keywords

genotype calling algorithm; data management; population attributable fraction; receiver-operating characteristic curve; signal intensity plot

INTRODUCTION

With the availability of high-throughput genotyping technologies based on hundreds of thousands of single-nucleotide polymorphisms (SNPs), genome-wide association (GWA) studies have become a standard approach for unraveling the basis of complex genetic diseases. The recent technological advances have created a series of challenges for genetic epidemiologists. Before we describe these challenges and some solutions that have been proposed at the Genetic Analysis Workshop 16 (GAW16), we describe the typical flow of a GWA and subsequent studies from the perspective of a genetic epidemiologist [Fig. 1, adapted from Ziegler et al., 2008].

Genetic epidemiological research starts at the design stage with a biological question. Having decided on the most appropriate study design, samples are collected, and the DNA chip is selected. The second stage of today's genetic epidemiological research is the laboratory stage: before chips can be hybridized, the DNA needs to be prepared. After chip

Corresponding Author: Andreas Ziegler, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Maria-Goeppert-Str. 1, 23562 Lübeck, Germany, Phone: +49 (451) 500-2780, Fax: +49 (451) 500-2999, ziegler@imbs.uni-luebeck.de.

hybridization, the chip is scanned. Now the statistical stages follow, and low-level analyses have to be performed before high-level analyses.

Low-level analysis starts with image analysis. In some software packages the investigator does not recognize that normalization of signal intensities and genotype calling are two different tasks that are performed in two steps. Based on the called genotypes and/or the signal intensities, extensive quality control is performed.

High-level statistical analysis only starts after completion of quality control. These are followed by subsequent replication and validation studies. At the end, the effect on the population is investigated. The power to identify new loci relies on the sample size, and therefore there is a need to combine data for meta-analyses across multiple studies and multiple platforms. To this end, SNPs that are not available on a specific chip are imputed, and statistical analysis, typically a meta-analysis with subsequent replication and validation, is conducted.

Finally, as in the path without imputation, population effect is investigated. Other aspects that are not depicted in Fig. 1 also play a role. For example, functional or animal studies are carried out to investigate whether the identified associations cause disease. Alternatively, family studies are used to determine whether the disease follows a particular Mendelian model.

New challenges emerge for genetic epidemiologists at different points in the conduct of a GWA and subsequent studies, and in this GAW16 group several solutions to these challenges have been proposed or evaluated. This GAW16 group only analyzed the real data, and it consists in seven different papers.

COMPUTATIONAL CHALLENGES – MEMORY MANAGEMENT

First, important computational challenges arise. For example, the image of a typical single Affymetrix chip requires more than 60 MB of storage. With a typical sample size of 1,000 cases and 1,000 controls, approximately 120 GB of storage are needed. After image processing, a sample still requires 37 GB of storage as an Affymetrix cel file. Only after normalization and genotype calling is the file size substantially reduced; it is approximately 3.5 GB for the typical case–control study with a total of 2,000 subjects. Of course, for simple input/output operations of the data, more than 1 GB of memory is still required. Therefore, specific data management and memory management tools have been developed for the analysis of GWA studies, including GenABEL [Aulchenko et al., 2007], GENOMIZER [Franke et al., 2006], GSCANDB [Taylor et al., 2007], OpenADAM [Yeung et al., 2008], PLINK [Purcell et al., 2007], or SNPLims [Orro et al., 2008].

The fundamental idea used in some of these specific programs is to use two bits for one SNP genotype [Aulchenko et al., 2007; Purcell, 2008] for high data compression. Specifically, a diallelic SNP-based genotype has four possible choices: 0 (AA), 1 (AB), 2 (BB), or 3 (missing), leading to 2 bits per SNP. The theoretical compression ratio therefore is 4:1 compared with a byte storage scheme (one byte for each genotype) minus some overhead.

Chen et al. [2009] investigated the performance of their own memory management tool, which uses the four SNPs per byte storage approach and compared it with the standard one SNP per byte storage approach. They applied their tool to the data from the North American Rheumatoid Arthritis Consortium (NARAC), which included 2,062 subjects and 550,000 SNPs from the Illumina Infinium HumanHap550 SNP chip [Amos et al., 2009]. In their analysis using the simple allelic χ^2 test for association, Chen et al. [2009] observed a heap memory usage of 305 MB for the compressed data storage and more than 1 GB (1074 MB)

for the uncompressed data storage. The differences between the two approaches in central processing unit (CPU) time were not pronounced for the simple allelic test. However, when haplotype blocks were to be identified, a huge discrepancy was found with CPU time of ~11 sec for compressed data storage but 169 sec for uncompressed data storage.

In conclusion, SNP data should be stored with two bits per SNP. This saves both CPU time and storage.

ACCURACY AND COMPUTATIONAL CHALLENGES – GENOTYPE CALLING

Computational speed and memory management as well as accuracy play an important role in the genotype calling stage of a study. In the last few years, many different genotype calling algorithms have been proposed for both the Affymetrix and the Illumina platforms. In their contribution, Vens et al. [2009] compared the three genotype calling algorithms BRLMM [Affymetrix, 2007], Chiamo [The Wellcome Trust Case Control Consortium, 2007], and JAPL [Plagnol et al., 2007] using Affymetrix GeneChip Human Mapping 500k Array Set data from the Framingham Heart Study (FHS) as provided for GAW16 [Cupples et al., 2009]. An important aspect of the study is that Vens et al. [2009] were not able to normalize all subjects in one run because of a memory access error when more than approximately 2,000 subjects were used in CelQuantileNorm, the normalization procedure recommended for JAPL and Chiamo. By investigating the concordance between the genotype calling algorithms, Vens et al. [2009] were able to identify previously undetected errors in strand coding. The highest number of samples with a call fraction <0.97 was observed for BRLMM, followed by Chiamo. No subject had a call fraction <0.97 when JAPL was used. Therefore, the authors conclude that JAPL would be the algorithm of choice if as many samples as possible should be retained for further analysis. This finding is in line with the conclusions of Plagnol et al. [2007], who stated that their genotype calling algorithm was specifically designed to deal with uncertain genotypes that are said to be missing by other approaches. Vens et al. [2009] also found that the highest number of SNPs was kept by Chiamo, so that this genotype calling algorithm would be the method of choice if investigators aim at keeping a high number of SNPs for further analyses after standard quality control.

When SNPs from a GWA study are represented in a deFinetti triangle [Ziegler and König, 2006], most of the SNPs group around the Hardy-Weinberg curve [Goddard et al., 2009]. BRLMM and JAPL showed excess heterozygosity, i.e., more heterozygous subjects than expected under HWE, for a larger number of SNPs than Chiamo. In contrast, Chiamo more often revealed a deficiency of heterozygotes than BRLMM and JAPL.

In summary, JAPL would be the algorithm of choice if as many samples as possible should be retained for further analysis. Chiamo would be the method of choice if investigators aim to keep a high number of SNPs for further analyses after standard quality control.

STATISTICAL CHALLENGES – QUALITY CONTROL

After genotype calling, standard quality control is performed on the subject level as well as on the SNP level. Standard filters on the subject level include

- Call fraction which should be as high as possible;
- Cryptic relatedness, as measured by identity by state (IBS) between pairs of subjects. If the IBS is too high, subjects might be closely related;
- Ethnic origin, as determined by principal component (PC), multidimensional scaling (MDS) or non-metric multidimensional scaling (NMDS) analysis. Study

populations should be as homogeneous as possible, and subjects with a different ethnic background should be excluded from analyses;

- Excess or deficiency of heterozygosity. If the heterozygosity on a chip is too high, the DNA might be contaminated. If it is too low, hybridization might have failed.

Standard filters on the SNP level include

- Minor allele frequency (MAF). Most genotype calling algorithms tend to perform poorly for SNPs with low MAF, and the power of a study is low for detecting associations to SNPs with a low MAF.
- Missing frequency (MiF), often termed 1−[SNP call rate]. It indicates how well the clusters of a SNP are separated. For case-control studies, the MiF should be investigated separately in cases and in controls because differential missingness between cases and controls can result in spurious associations [Clayton et al., 2005].
- Hardy-Weinberg equilibrium (HWE). SNPs are excluded if substantially more or fewer subjects are heterozygous at a SNP than expected (excess heterozygosity or heterozygote deficiency).

These global filters are effective in removing SNPs with clustering problems. They reduce a large number of highly significant erroneous associations and lower the genomic control lambda so that quantile-quantile plots do not show more outliers than expected under the null [Ling et al., 2009]. However, these filters are not able to identify all SNPs of bad quality. Therefore, Ling et al. [2009] have introduced sex-specific filters which should be added to the standard quality control procedures. The first three are for X-chromosomal markers (X), and the last four for autosomal markers:

- X: proportion of male heterozygote calls;
- X: absolute difference in the call fractions for males and females;
- X: code all samples as females, use the correct sex as phenotype and investigate whether the proportion of missing data is associated with sex;
- Absolute difference in call fractions for males and females;
- Proportion of heterozygotes in males and females in all samples;
- Missing data by sex;
- Test of allelic association by sex among controls.

The last test, which is carried out in the control group only, is especially meaningful because sex-based confounding is likely to cause some small differences in allele and genotype frequencies.

The traditional standard quality control filters, termed Travemunde Criteria, are summarized with the additional standard quality control filters in Table I. The name Travemunde Criteria comes from a consortium meeting in Travemunde held in 2007 that was used for the work of Samani et al. [2007] and subsequent papers.

Although the standard quality control approaches and the novel filters are helpful in identifying SNPs of low quality, the visual inspection of signal intensity plots is still the ultimate quality control approach when an association has been identified [Ziegler et al., 2008]. For example, Affymetrix states in its “Best Practices” for the analysis of data from GWA studies “Visually analyze all candidate SNPs” [Affymetrix, 2008, p. 257]. The recommendation to inspect only candidate SNPs is probably a consequence of the fact that

systematic visual inspection of all cluster plots is impossible in a high-throughput setting because of the high workload. For example, we currently require approximately 2 h for the independent visual inspection of 100 cluster plots by two experienced readers, and readers are fatigued after a short period.

Nevertheless, the inspection of all cluster plots, i.e., on the genome-wide level, is of interest. For example, for genotype imputation, which often is the basis for meta-analyses of GWA studies, only SNPs of high quality should be used [de Bakker et al., 2008]. Furthermore, when machine learning approaches or genome-wide haplotype analyses are used for GWA data [Trégouët et al., 2009; Ziegler et al., 2007], all SNPs should be quality assured.

Therefore, approaches would be helpful that allow the automated inspection of cluster plots, and this task is comparable to measuring the internal validity of the clustering in cluster analysis [Halkidi et al., 2002a; Halkidi et al., 2002b; Handl et al., 2005]. Intuitively, the genotype calling performs well for a specific SNP if neighboring points in a signal intensity plot that are similar are assigned the same genotype and points that are dissimilar are assigned to different genotypes. Furthermore, a good SNP will have small distances within a genotype group and large distances between different genotypes.

Formally, the validity of a genotype calling can be measured as follows [for reviews see Handl et al., 2005; Kim and Ramakrishna, 2005]:

- *Compactness* measures closeness of genotypes. This concept is related to the intra-cluster variation, and therefore a typical example for such a measure is the variance. Of course, the variance also indicates how different the subjects within a genotype group are. However, a low value of variance is an indicator of closeness.
- *Connectedness* attempts to assess how well partitioning groups subjects together with their nearest neighbors. Representatives of such measures count violations of nearest neighbor relationships.
- *Separability* indicates how distinct two genotype groups are, and therefore the distance is compared between two different clusters.
- *Combinations* of the above criteria: A number of approaches combine measures of the above types, and several measures assess both intra-cluster homogeneity and inter-cluster separation, and an example for such a measure has been given by Plagnol et al. [2007] in the context of GWA studies. Another example is the non-linear combination of both measures using the silhouette width as discussed by Lovmar et al. [2005] in the context of genotype quality.
- *Cluster stability* is a special form of internal cluster validation. It measures how sensitive a method is to perturbation of the data, i.e., how sensitive the genotypes are with respect to small changes in the signal intensity. Measures of this type repeatedly re-sample or perturb the original data set, and re-cluster the resulting data. The consistency of the corresponding results provides an estimate of the significance of the clusters obtained from the original data set. In the context of genotype quality, it has been discussed by Teo et al. [2008]. The major disadvantage of this approach is its CPU time. Specifically, genotypes need to be called anew after adding the perturbation. Currently, for a study with 1,500 chips the required time is approximately one working day on a Dual Quad Core with 32 GB RAM. Of course, the analyses should be performed repeatedly for averaging the random effects.

The usefulness of these approaches for large sample sizes and GWA studies has not been studied in detail. Therefore, the proposal of Schillert et al. [2009] can be considered a first

step in this direction. They introduced an automated cluster plot analysis (ACPA) approach, and their method falls in the group of connectedness. In the method, the Mahalanobis distance is considered from the center of a cluster to all samples within the cluster. Next, a cluster boundary is defined by distending the ellipses of the cluster using a factor that depends on the interquartile range. Finally, the number of samples from the other clusters falling in the boundary of the cluster under consideration is calculated. If the number of subjects falling in the boundary of a different cluster exceeds a certain limit, the SNP is said to have unreliable clustering. They assessed the performance of ACPA with the decision made by two independent readers based on the BRLMM calls for 1,000 randomly selected SNPs from the FHS. Sensitivity – correct detection of low quality SNPs – was 88% and specificity – correct detection of high quality SNPs – was 86%. By varying the width of the boundary, Schillert et al. [2009] were able to increase the specificity to 99% with a sensitivity of 50%.

In summary, standard quality control, including the novel filters proposed by Ling et al. [2009], is an absolute requirement before high-level analyses. The automated evaluation of cluster plots should be further improved.

ACCURACY AND COMPUTATIONAL CHALLENGES – GENOTYPE IMPUTATION

There is a growing need to work with complete genotypic data, e.g., for machine learning approaches, and to combine genotype data across multiple studies that have been obtained from different platforms. The analysis of missing data has a long tradition in statistics, and it is important to be aware of the different missing data mechanisms and potential pitfalls for the statistical analysis [D'Agostino, 2007; Gail, 1991; Laird, 1988]. While traditional statistical approaches for dealing with missing data use data from the study of interest only, several approaches have been proposed in the context of GWA studies recently that make use of external data sources [Li and Abecasis, 2006; Marchini et al., 2007; Nicolae, 2006; Servin and Stephens, 2007]. A disadvantage of the available publications is that the statistical assumptions underlying the employed methods are rarely formulated. Several studies were performed at GAW16 that compared the performance of several genotype imputation packages in terms of accuracy, speed, and user-friendliness [for a review see Thomas, 2009].

CHALLENGES FOR PUBLIC HEALTH – MEASURES OF DIAGNOSTIC ACCURACY

When a series of disease-associated SNPs have been identified, replicated, and possibly validated [for a detailed discussion of the terminology, see Igl et al., 2009], standard measures of diagnostic accuracy for a quantitative diagnostic test are investigated. These include the area under the curve (AUC), which can be constructed even if only summary data are available [Lu and Elston, 2008]. Jeffries and Zheng [2009] compared the Lu-Elston approach with the standard logistic regression method when individual-level data are available. They observed that the Lu-Elston method is valuable when only summary statistics can be used. However, the conventional logistic regression is preferable when full data sets are available, because it allows model selection using standard likelihood theory. Furthermore, to provide useful information without a complete data set, the Lu-Elston method is subject to two constraints. First, to be included in the model, continuous covariates are converted to factors with a few levels. Second, unless considerable information regarding pairwise LD is available, the SNPs are modeled as independent. This

means that multilocus genotype probabilities have to be obtained by the product of single-SNP genotype probabilities.

A different scenario for population-based measures has been considered by Hadley and Strachan [2009]. They showed that the population attributable risk (PAR), i.e., the proportion of cases attributable to a variant, at the untyped functionally relevant SNP can be estimated from the allele frequency p and the allelic relative risk RR at an observed SNP as follows. The (PAR) is often called attributable risk for short. In a first step, a parameter φ_{obs} is estimated at the typed – observed – SNP as $\varphi_{obs}=\frac{p(RR-1)}{1+p(RR-1)}$. In the second step, φ_{true} at the functionally relevant position is estimated via $\varphi_{true} = \varphi_{obs}/\mathscr{D}$ where $\mathscr{D}$ is the usual normalized Lewontin's measure of linkage disequilibrium (LD). When the functionally relevant SNP is not typed in a specific study, the $\mathscr{D}$ estimated from an external data source is used. Finally, the PAR at the untyped SNP is obtained as $PAR = \varphi_{true}(2 - \varphi_{true})$.

When a set of genotyped SNPs $k = 1,\ldots,K$ is available that are in LD with the functionally relevant variant, Hadley and Strachan [2009] proposed to calculate φ_{true} as a weighted average across typed SNPs. The weights should be inversely proportional to the variance. Using the delta method, Hadley and Strachan [2009] derived the approximate variance of φ_{true} and showed that inverse-variance weights proportional to the r^2 LD measure are appropriate so that $\varphi_{true}=(\sum_{k=1}^{K} r_k^2\varphi_{obs,k})/(\sum_{k=1}^{K} r_k^2)$, where r_k^2 is the coefficient of determination between the k^{th} typed SNP and the unmeasured functionally relevant variant.

In summary, the conventional logistic regression is preferable for constructing an AUC over the Lu-Elston approach when full datasets are available. Using a simple transformation, the PAR can be estimated at an untyped SNP from genotyped SNPs when information about the LD is available.

Acknowledgments

The author is grateful to all participants of GAW16 Group 8. The group discussions would not have been a success without the participating senior colleagues Joan E. Bailey-Wilson, Heather Cordell, Charles C. Gu, and Yan Sun. The author is also grateful to the authors of the seven *BMC Proceedings* papers summarized in this work: Xiang Chen, David Hadley, Neal Jeffries, Hua Ling, Arne Schillert, Daniel F. Schwarz, and Maren Vens. This work was supported by the German Ministry of Education and Science, grant 01 EZ 0874, and the German Research Foundation, grant ZI 591/17-1. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

REFERENCES

Affymetrix. BRLMM: An improved genotype calling method for the GeneChip® Mapping 500K Array Set. Santa Clara, CA: Affymetrix; 2007.

Affymetrix. Affymetrix® Genotyping Console 3.0 user manual. Santa Clara, CA: Affymetrix; 2008.

Amos CI, Chen WV, Seldin MF, Remmers E, Taylor KE, Criswell LA, Lee AT, Plenge RM, Kastner DL, Gregersen PK. Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. BMC Proc. 2009; 3 Suppl 7:S2. [PubMed: 20018009]

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: An R library for genome-wide association analysis. Bioinformatics. 2007; 23:1294–1296. [PubMed: 17384015]

Chen X, Zhang M, Wang M, Zhu W, Cho K, Zhang H. Memory management in genome-wide association studies. BMC Proc. 2009; 3 Suppl 7:S54. [PubMed: 20018047]

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet. 2005; 37:1243–1246. [PubMed: 16228001]

Cupples LA, Heard-Costa N, Lee M, Atwood LD. for the Framingham Heart Study Investigators. Genetic Analysis Workshop 16 Problem 2: The Framingham Heart Study data. BMC Proc. 2009; 3 Suppl 7:S3. [PubMed: 20018020]

D'Agostino RB Jr. Overview of missing data techniques. Methods Mol Biol. 2007; 404:339–352. [PubMed: 18450058]

de Bakker PI, Ferreira MA, 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. [PubMed: 18852200]

Franke A, Wollstein A, Teuber M, Wittig M, Lu T, Hoffmann K, Nurnberg P, Krawczak M, Schreiber S, Hampe J. GENOMIZER: An integrated analysis system for genome-wide association data. Hum Mutat. 2006; 27:583–588. [PubMed: 16652332]

Gail MH. A bibliography and comments on the use of statistical models in epidemiology in the 1980s. Stat Med. 1991; 10:1819–1885. [PubMed: 1805315]

Goddard KA, Ziegler A, Wellek S. Adapting the logical basis of tests for Hardy-Weinberg equilibrium to the real needs of association studies in human and medical genetics. Genet Epidemiol. 2009 (Epub ahead of print).

Hadley D, Strachan DP. Inference of disease associations with unmeasured genetic variants by combining results from genome-wide association studies with linkage disequilibrium patterns in a reference data set. BMC Proc. 2009; 3 Suppl 7:S55. [PubMed: 20018048]

Halkidi M, Batistakis Y, Vazirgiannis M. Cluster validity methods: Part I. Sigmod Rec. 2002a; 31:40–45.

Halkidi M, Batistakis Y, Vazirgiannis M. Clustering validity checking methods: Part II. Sigmod Rec. 2002b; 31:19–27.

Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics. 2005; 21:3201–3212. [PubMed: 15914541]

Igl B-W, König IR, Ziegler A. What do we mean by "replication" and "validation" in genome-wide association studies? Hum Hered. 2009; 67:66–68. [PubMed: 18931511]

Jeffries N, Zheng G. Evaluation of an optimal receiver operating characteristic procedure. BMC Proc. 2009; 3 Suppl 7:S56. [PubMed: 20018049]

Kim M, Ramakrishna RS. New indices for cluster validity assessment. Pattern Recognit Lett. 2005; 26:2353–2363.

Laird NM. Missing data in longitudinal studies. Stat Med. 1988; 7:305–315. [PubMed: 3353609]

Li Y, Abecasis GR. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet. 2006; S79:2290. [abstract].

Ling H, Hetrick K, Bailey-Wilson JE, Pugh EW. Application of sex-specific single-nucleotide polymorphism filters in genome-wide association data. BMC Proc. 2009; 3 Suppl 7:S57. [PubMed: 20018050]

Lovmar L, Ahlford A, Jonsson M, Syvanen AC. Silhouette scores for assessment of SNP genotype clusters. BMC Genomics. 2005; 6:35. [PubMed: 15760469]

Lu Q, Elston RC. Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet. 2008; 82:641–651. [PubMed: 18319073]

Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007; 39:906–913. [PubMed: 17572673]

Nicolae DL. Testing untyped alleles (TUNA)-applications to genome-wide association studies. Genet Epidemiol. 2006; 30:718–727. [PubMed: 16986160]

Orro A, Guffanti G, Salvi E, Macciardi F, Milanesi L. SNPLims: A data management system for genome wide association studies. BMC Bioinformatics. 2008; 9 Suppl 2:S13. [PubMed: 18387201]

Plagnol V, Cooper JD, Todd JA, Clayton DG. A method to address differential bias in genotyping in large-scale association studies. PLoS Genet. 2007; 3:e74. [PubMed: 17511519]

Purcell, S. PLINK (v1.05). A whole-genome association toolset. Boston, MA: Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital. 2008. http://pngu.mgh.harvard.edu/~purcell/plink/

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, König IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H. WTCCC and the Cardiogenics Consortium. Genome-wide association analysis of coronary artery disease. N Engl J Med. 2007; 357:443–453. [PubMed: 17634449]

Schillert A, Schwarz DF, Vens M, Szymczak S, König IR, Ziegler A. ACPA: Automated cluster plot analysis of genotype data. BMC Proc. 2009; 3 Suppl 7:S58. [PubMed: 20018051]

Servin B, Stephens M. Imputation-based analysis of association studies: Candidate regions and quantitative traits. PLoS Genet. 2007; 3:e114. [PubMed: 17676998]

Taylor M, Valdar W, Kumar A, Flint J, Mott R. Management, presentation and interpretation of genome scans using GSCANDB. Bioinformatics. 2007; 23:1545–1549. [PubMed: 17400728]

Teo YY, Small KS, Clark TG, Kwiatkowski DP. Perturbation analysis: A simple method for filtering SNPs with erroneous genotyping in genome-wide association studies. Ann Hum Genet. 2008; 72:368–374. [PubMed: 18261185]

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]

Thomas DC. Genome-wide association studies for discrete traits. Genet Epidemiol. 2009 this volume.

Trégouët D-A, König IR, Erdmann J, Munteanu A, Braund PS, Hall AS, Götz A, Linsel-Nitschke P, Perret C, DeSuremain M, Meitinger T, Wright BJ, Preuss M, Balmforth AJ, Ball SG, Meisinger C, Germain C, Evans A, Arveiler D, Luc G, Ruidavets JB, Morrison C, van der Harst P, Schreiber S, Neureuther K, Schäfer A, Bugert P, El Mokhtari NE, Schrezenmeir J, Stark K, Rubin D, Wichmann HE, Hengstenberg C, Ouwehand W, Ziegler A, Tiret L, Thompson JR, Cambien F, Schunkert H, Samani NJ. Wellcome Trust Case Control Consortium; Cardiogenics Consortium. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat Genet. 2009; 41:283–285. [PubMed: 19198611]

Vens M, Schillert A, König IR, Ziegler A. Look who is calling: A comparison of genotype calling algorithms. BMC Proc. 2009; 3 Suppl 7:S59. [PubMed: 20018052]

Yeung JM, Sham PC, Chan AS, Cherny SS. OpenADAM: An open source genome-wide association data management system for Affymetrix SNP arrays. BMC Genomics. 2008; 9:636. [PubMed: 19117518]

Ziegler A, DeStefano AL, König IR. on behalf of Group 6. Data mining, neural nets, trees – Problems 2 and 3 of Genetic Analysis Workshop 15. Genet Epidemiol. 2007; 31 Suppl 1:S51–S60. [PubMed: 18046765]

Ziegler, A.; König, IR. A statistical approach to genetic epidemiology: Concepts and applications. Weinheim: Wiley-VCH; 2006.

Ziegler A, König IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biom J. 2008; 50:8–28. [PubMed: 18217698]

Figure 1.
Succession of design, experimental and data analysis steps in a genome-wide association and subsequent studies. Adapted from Ziegler et al. [2008].

Table I

Filters for standard quality control (sQC) of genome-wide association (GWA) studies: Travemunde Criteria.

Level	Filter criterion	Standard value for filter
Subject	Call fraction	≥ 97%[a]
	Cryptic relatedness	Study specific
	Ethnic origin	Study specific; visual inspection of principal components
	Heterozygosity	Within mean ± 3 SD over all samples
	Heterozygosity by sex	Within mean ± 3 SD within sex group
SNP	Minor allele frequency (MAF)	≥ 1%
	Missing frequency (MiF)	≤ 2% in any study group, e.g., in both cases and controls
	MiF by sex	≤ 2% in any sex
	Hardy-Weinberg equilibrium	$p < 10^{-4}$
	Difference between control groups	$p > 10^{-4}$ in Cochran-Armitage trend test between control groups
	Sex differences among controls	$p > 10^{-4}$ in Cochran-Armitage trend test between males and females
X-chromosomal SNPs only	Code all samples as females, use the correct sex as phenotype and investigate whether the proportion of missing data is associated with sex	No standard value available
	Proportion of male heterozygote calls	No standard value available
	Absolute difference in the call fractions for males and females	No standard value available
	Sex-specific heterozygosity	No standard value available