Quality Control for Genome-Wide Association Studies in Humans

Arne Schillert[*], Andreas Ziegler[*]

Introduction

In their last issue in 2006, the News Staff (2006) from Science announced genome-wide association (GWA) studies to be one of the areas to watch in 2007. Indeed, the years since then have seen a surge in publications of GWA studies identifying and replicating genetic factors (Manolio and Collins (2009)). However, in addition to their apparent success, GWA studies also pose new challenges for biostatisticians, and several quality control aspects will be discussed in this paper. In the next section, we describe some differences between candidate gene and GWA studies. Subsequently, we explain the typical scenario of a GWA study. We illustrate the differences in the user-friendliness of genotype calling algorithms by comparing two of them. The last two sections are dedicated to the standard quality control (sQC) and to the reading of cluster plots (CPR), also termed signal intensity plots.

What is a genome-wide association (GWA) study?

The aim of any genetic association study is to identify associations between a phenotype on the one hand and one or more genetic markers on the other. Specifically, GWA studies use single nucleotide polymorphisms (SNPs) as genetic markers. However, while in candidate gene association studies SNPs are analyzed in candidate genes or regions, the entire genome is investigated in GWA studies. These candidates are plausible because of biological function or previous study results. The GWA study approach is primarily indirect as we do not necessarily suppose that any of the SNPs are causal. They may be in close vicinity to functional relevant variants instead and associated with the disease because of linkage disequilibrium (LD).

Because GWA studies investigate SNPs in the entire genome, a GWA study should be performed without an a priori specified hypothesis on the location. As a consequence, approaches which a priori weight the specific genetic location based on prior knowledge are not recommended (Sun et al. (2009)). In fact, many of the loci that have been identified in humans through GWA studies were not candidates. For example, the best replicated locus for myocardial infarction is located on chr. 9p21.3 (Coronary Artery Disease Consortium (2009)). The region is defined by two flanking recombination hotspots and contains the coding sequences of genes for 2 cyclin-dependent kinase inhibitors, *CDKN2A* and *CDKN2B*. However, the most strongly associated SNPs lie considerably upstream of these genes, and the nearest signal is 10 kb upstream of *CDKN2B*. Instead, these SNPs all lie within the antisense non-coding RNA at *INK4* locus (*ANRIL*) (Broadbent et al. (2008)). As a result, this region would not have been identified by a priori weighting of loci or candidate gene studies.

[*] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Str. 1, 23562 Lübeck, Germany

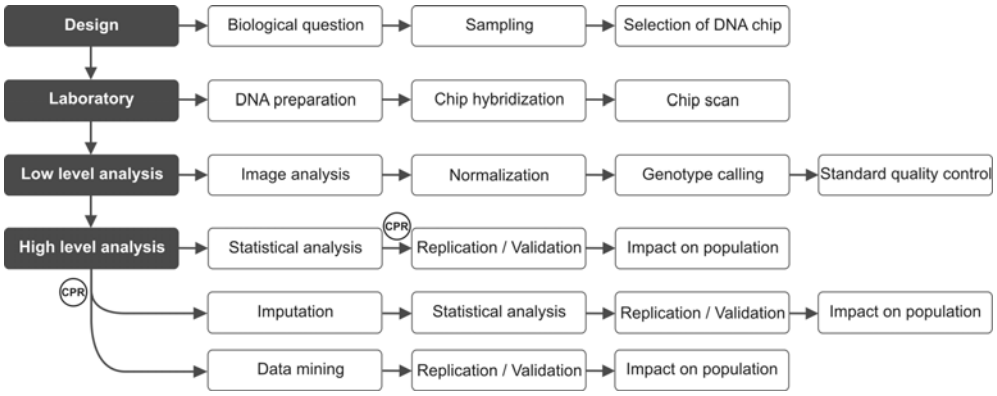

Figure 1: Succession of design, experimental and data analysis steps in a genome-wide association and subsequent studies. Adapted from Ziegler et al. (2008) and from Ziegler (2009). CPR: cluster plot reading.

What is the flow of a genome-wide association (GWA) study?

A sequence of steps has to be taken in a GWA (Figure 1; also see Ziegler et al. (2008), Ziegler (2009)). The biological question needs to be formulated first, and the study design needs to be fixed. Here, an important difference exists between human and livestock production GWA studies. While familial information is unavailable in most GWA studies in humans, extended pedigrees can usually be considered in animal or plant breeding. Another relevant aspect even at this design stage is how any positive results of the GWA study can be replicated and followed-up (Deutsche Forschungsgemeinschaft (2008)). It is especially important for rare diseases, where sampling is limited. Finally, before a GWA study can be conducted, the DNA chip needs to be chosen. According to our experience, the most important factor for researchers seems to be cost of the chip. However, coverage of the genome or quality of the chip, time and effort for quality control should at least be equally important.

In the second stage of a GWA study experiment, the DNA is hybridized on a chip after specific preparatory steps. This process is not of primary importance for data analysis at later stages, and it is therefore ignored here. Subsequent steps are similar to those of gene expression studies. After washing and staining, the array is scanned, and the raw optical images are obtained. Signal intensities are calculated from pixel values per spot, and these intensities are stored. The biostatistical analysis team typically does not receive the intensity files from the laboratory. However, the spot intensity could serve as an indicator for chip quality as for most microarray platforms (Hartmann (2005)).

After image analysis signal intensities are normalized, and genotypes are called, i.e., signal intensities are converted into genotypes. Both the normalization and the genotype calling steps are important for the overall quality of a GWA study. The normalization of signal intensities can be performed analogously to microarray gene expression experiments, and we therefore restrict our interest to the user-friendliness of the genotype calling algorithms in the next section. After genotype calling intensive sQC is performed; see below. sQC is then followed by high level statistical analyses. Standard approaches for testing for association

include the comparison of genotype frequencies between cases and controls which are calculated for each SNP separately. In general, another level of quality control, the CPR, is included after this first step of analysis. Only those SNPs with a positive finding which pass this extra quality control step are put forward to a replication study. This CPR step is discussed below in greater detail.
In addition to standard statistical analysis, meta-analyses are commonly conducted in big consortia (see, e.g., Köttgen et al. (2010)). To make SNP data comparable that have been obtained using different chips, the genotype data are imputed. The fundamental idea behind SNP imputation is simple. Many SNPs cluster in haplotype blocks, and typically it is sufficient to type just a single SNP from this block so that the genotype at the other SNPs in this block can be deduced. The imputation quality needs to be investigated, but this aspect of quality control will not be discussed in this work.
The CPR is also important for other high level statistical analysis. For example, the potential of false positive conclusions from badly typed SNPs exaggerates if data mining approaches, gene-gene interactions, or haplotypes are considered. In the final two section of this work we concentrate on sQC and automated CRP reading of GWA studies performed using the Affymetrix chip technology.

How user-friendly are the two calling algorithms CRLMM and JAPL?

In a systematic literature search we have identified 9 different genotype calling algorithms suitable for different Affymetrix chips, and we have extensively compared 5 of the algorithms (unpublished work). In this section, we compare two of these algorithms with respect to usability. Figure 2 depicts the workflow for the calling algorithm CRLMM (Corrected Robust Linear Mixed Model; Carvalho et al. (2007)). The implementation of this algorithm is characterized by its great user-friendliness. For example, data files and paths are checked for existence, and chip definition files are automatically assigned to the detected array type. Genotypes and quality measures are separated in two files, and annotation files can be easily combined with the Bioconductor package.

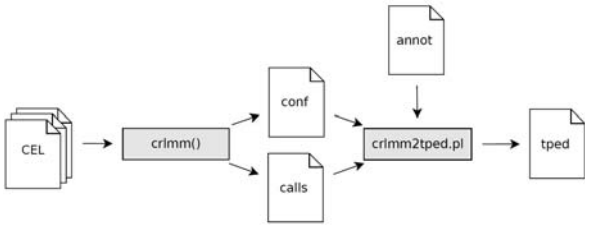

Figure 2: Flow chart for the genotype calling with CRLMM. Starting point are the cel-files. Data files are symbolized by bent corners, scripts are displayed in gray. The function crlmm() from the Bioconductor package oligo automatically reloads the required cdf files. Relevant result files are calls and conf. They contain the numerically coded genotypes and quality measures. The Perl script crlmm2tped.pl has been written by our group to combine the information of the two files. Together with the annotation file, the genotypes are converted to a tped file, which can be used by the software package PLINK (Purcell et al. (2007)).

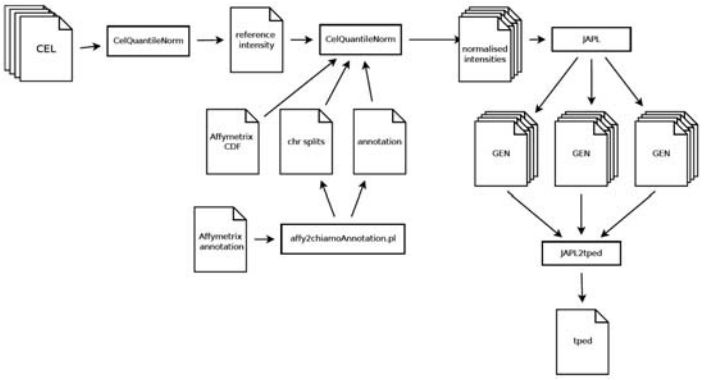

Figure 3. Flow chart for the genotype calling with JAPL. Starting point are the cel files and the Affymetrix annotation files. The path until "normalized intensities" represents preparatory steps which were performed using CelQuantileNorm. The integration of current annotation files is elaborate. The final step is the calling by JAPL. A GEN file is generated for each SNP in a chromosome specific directory. Conversion to a tped file (PLINK format) is done by a self-developed Perl script.

The workload is substantially higher for genotype calling by JAPL (Figure 3; Plagnol et al. (2007)). For example, the normalization is left to the user. Furthermore, updating annotation files is rather difficult because different symbols are used for separating columns in the required help files, and this requirement was not documented. No plausibility checks are performed concerning existence of user-defined file names. As a result, if permission to write the results file in a specific directory is denied, all calculations are performed to the very end. Finally, JAPL produces one directory per chromosome and stores one GEN file per SNP in the corresponding directory. Subsequently, the effort for data conversion is quite high.

What is standard quality control in GWA studies?

sQC is performed after genotype calling both on the subject level and on the SNP level. Standard filters on the subject level include (Ziegler (2009)):

- Call fraction which should be as high as possible;
- Cryptic relatedness, as measured, e.g., 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) analysis or similar 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 (Ziegler (2009)):

- 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. The reader should note that the performance relies on the number of samples to be called. The higher the numbers of samples, the more subjects are in the rarest genotype group. If a sufficient number of subjects from the rare genotype group are to be called, the genotype calling can detect these as separate cluster – not as outlier.
- Missing frequency (MiF), often termed 1 minus 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.
- Hardy-Weinberg equilibrium (HWE). SNPs are excluded if substantially more or fewer subjects are heterozygous at a SNP than expected. This filter should not be used if selection or other causes resulting in deviation from HWE can be expected in advance of a GWA study.

Table 1: Filters for standard quality control (sQC) of genome-wide association (GWA) studies: Travemunde criteria[α]

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 std.dev. over all samples
	Heterozygosity by gender	Within mean ± 3 std.dev. within gender group
SNP	Minor allele frequency (MAF)	≥ 1%
	Missing frequency (MiF)	≤ 2% in any study group, e.g., in both cases and controls
	MiF by gender	≤ 2% in any gender
	Hardy-Weinberg equilibrium	$p < 10^{-4}$
	Difference between control groups	$p > 10^{-4}$ in Cochran-Armitage trend test between control groups
	Gender differences among controls	$p > 10^{-4}$ in Cochran-Armitage trend test between males and females
X-Chr.	Investigate differential missingness by gender after coding all samples as females; correct gender is phenotype	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
	Gender-specific heterozygosity	No standard value available

[α]Table adopted from Ziegler (2009). SNP: single nucleotide polymorphism.

Gender-specific filters include

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

These global sQC filters, termed Travemunde Criteria, are effective in removing SNPs with clustering problems. They are summarized in Table 1 which is reproduced from Ziegler (2009).

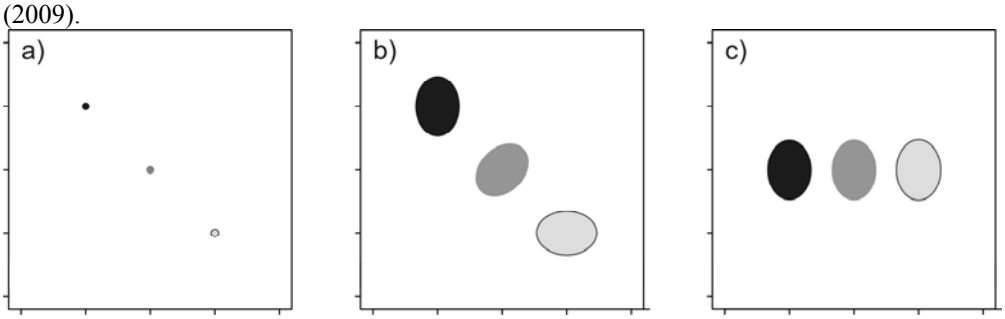

Figure 4: Signal intensity plots. a) shows the idealized situation that all subjects can be assigned uniquely without measurement error to one of the three possible genotypes. b) displays the more realistic scenario of a high-quality SNP after normalization. Signal intensities of all subjects represent 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. c) displays the signal intensities I_A and I_B of alleles *A* and *B* in the sense of a Bland–Altman plot. The x-axis shows the contrasts, i.e., the normalized difference of the signal intensities $(I_A - I_B)/(I_A + I_B)$ of both alleles, and the y-axis gives the sum of the intensities $I_A + I_B$ of both alleles. Figures a) and b) are reprinted in black and white from Ziegler et al. (2008); figure c) is reprinted in black and white from Ziegler and König (2010).

What is cluster plot reading? Can it be performed automatically?

Although the sQC approaches and the novel filters are helpful in identifying SNPs of low quality, the visual inspection of signal intensity plots is still the gold standard quality control approach. For example, Affymetrix recommends the visual analysis of all candidate SNPs in its "Best Practices" for the 6.0 chip (Affymetrix (2008, p. 257)). This recommendation is restricted to candidate SNPs since all cluster plots cannot be visually inspected because of the high workload. However, the recommendation from Affymetrix only reflects the branch to standard statistical analysis in Figure 1. Genotype imputation and data mining approaches also rely on accurate genotypes, and they do not generate single candidate SNPs but groups

of SNPs of interest. If at least one SNP in this group shows a bad clustering, the entire analysis may be flawed. It would need to be redone after elimination of the erroneous SNP. Therefore, it would be important to have an automate that is able to perform the CPR for all SNPs.

To understand CPR, it is important to know how the variables are generated. In brief, for every sample at every 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 sample is classified as homozygous, if the intensities for both alleles are equally high, the sample is assigned a heterozygous genotype. 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 4a). However, in practice signal intensities of all subjects for the three genotypes form three clouds after normalization (Figure 4b). To evaluate the quality of the genotype calls, the signal intensities for both alleles can be plotted as shown in Figure 4b or using the contrasts (Figure 4c) together with the classification. SNPs with dubious genotype assignments can be excluded.

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. To this end, they considered the smallest difference between the centres of two adjacent clouds divided by the sum of the standard deviation for these two clouds. We do not only consider the distance between clouds but also investigate the overlap of clusters (Schillert et al. (2009)). However, since genotype calling is identical to group label assignment to clusters, these approaches are special cases of internal validation methods for cluster analysis (Ziegler (2009)).

Formally, the validity of a genotype calling can be measured as follows (Ziegler (2009); 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. A low variance is an indicator of closeness.
- *Connectedness* assesses how well the partitioning groups subjects together with their nearest neighbours.
- *Separability* indicates how distinct two genotype groups are, and one compares the distance between two 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 (e.g., Lovmar et al. (2005), Plagnol et al. (2007), Schillert et al. (2009)).
- *Cluster stability* investigates how sensitive is a method to perturbation of the data, i.e., how sensitive the genotypes are with respect to small changes in the signal intensity (Teo et al. (2008)). The approach is impractical for application because genotypes need to be called anew after adding the perturbation.

The usefulness of these approaches for large sample sizes and GWA studies is just being studied in detail. The first results are promising (see, e.g., Schillert et al. (2009)) but the automated evaluation of cluster plots needs further improvement.

Acknowledgements

The work presented in this paper was funded by the German Ministry of Education and Research (grant: 01EZ0874) and the German Research Foundation (grant: ZI 591/17-1).

References

CelQuantileNorm. http://www.wtccc.org.uk/info/software.shtml; [19.04.2010].

Affymetrix (2008). *Affymetrix® Genotyping Console 3.0 User Manual.* http://www.affymetrix.com /support/downloads/manuals/genotyping_console_manual.pdf; [31.10.2009].

Broadbent, H. M., Peden, J. F., Lorkowski, S., *et al.* (2008). *Hum. Mol. Genet.*, 17:806-814.

Carvalho, B., Bengtsson, H., Speed, T. P., *et al.* (2007). *Biostatistics*, 8:485-499.

Coronary Artery Disease Consortium, Samani, N. J., Deloukas, P., *et al.* (2009). *Arterioscler. Thromb. Vasc. Biol.*, 29:774-780.

Deutsche Forschungsgemeinschaft (2008). *Genome-Wide Association Studies (GWAS). A Checklist of Methodological and Conceptual Requirements.* http://www.dfg.de/download/pdf/foerderung /grundlagen_dfg_foerderung/informationen_fachwissenschaften/lebenswissenschaften/checkliste_gwa_studien.pdf; [18.04.2010].

Handl, J., Knowles, J., and Kell, D. B. (2005). *Bioinformatics*, 21:3201-3212.

Hartmann, O. (2005). *Methods Inform. Med.*, 44:408-413.

Kim, M., and Ramakrishna, R. S. (2005). *Pattern Recognit. Lett.*, 26:2353-2363.

Köttgen, A., Pattaro, C., Böger, C. A., *et al.* (2010). *Nat. Genet.*, 42:376-384.

Lovmar, L., Ahlford, A., Jonsson, M., *et al.* (2005). *BMC Genomics*, 6:35.

Manolio, T. A., and Collins, F. S. (2009). *Annu. Rev. Med.*, 60:443-456.

Plagnol, V., Cooper, J. D., Todd, J. A., *et al.* (2007). *PLoS Genet.*, 3:e74.

Purcell, S., Neale, B., Todd-Brown, K., *et al.* (2007). *Am. J. Hum. Genet.*, 81:559-575.

Schillert, A., Schwarz, D. F., Vens, M., *et al.* (2009). *BMC Proc.*, 3:S58.

Sun, J., Jia, P., Fanous, A. H., *et al.* (2009). *Bioinformatics*, 25:2595-6602.

Teo, Y. Y., Small, K. S., Clark, T. G., *et al.* (2008). *Ann. Hum. Genet.*, 72:368-374.

The News Staff (2006). *Science*, 314:1850-1855.

Ziegler, A. (2009). *Genet. Epidemiol.*, 33:S45-S50.

Ziegler, A., and König, I. R. (2010). *A Statistical Approach to Genetic Epidemiology: Concepts and Applications.* Second ed. Wiley-VCH, Weinheim.

Ziegler, A., König, I. R., and Thompson, J. R. (2008). *Biom. J.*, 50:8-28.