

ARTICLE

How to deal with the early GWAS data when imputing and combining different arrays is necessary

Hae-Won Uh^{*,1,2}, Joris Deelen^{2,3}, Marian Beekman³, Quinta Helmer¹, Fernando Rivadeneira^{2,4,5}, Jouke-Jan Hottenga⁶, Dorret I Boomsma⁶, Albert Hofman^{2,4,5}, André G Uitterlinden^{2,4,5}, PE Slagboom^{2,3}, Stefan Böhringer¹ and Jeanine J Houwing-Duistermaat¹

Genotype imputation has become an essential tool in the analysis of genome-wide association scans. This technique allows investigators to test association at ungenotyped genetic markers, and to combine results across studies that rely on different genotyping platforms. In addition, imputation is used within long-running studies to reuse genotypes produced across generations of platforms. Typically, genotypes of controls are reused and cases are genotyped on more novel platforms yielding a case–control study that is not matched for genotyping platforms. In this study, we scrutinize such a situation and validate GWAS results by actually retyping top-ranking SNPs with the Sequenom MassArray platform. We discuss the needed quality controls (QCs). In doing so, we report a considerable discrepancy between the results from imputed and retyped data when applying recommended QCs from the literature. These discrepancies appear to be caused by extrapolating differences between arrays by the process of imputation. To avoid false positive results, we recommend that more stringent QCs should be applied. We also advocate reporting the imputation quality measure (R_T^2) for the post-imputation QCs in publications.

European Journal of Human Genetics (2012) 20, 572–576; doi:10.1038/ejhg.2011.231; published online 21 December 2011

Keywords: GWAS; imputation; quality control

INTRODUCTION

Imputation-based association methods provide a powerful framework for testing ungenotyped variants for association with phenotypes. Genotype imputation is particularly useful for combining results across studies that use different genotyping platforms, because a meta-analysis of several studies with relatively modest findings can result in a number of strongly associated loci that were not previously indicated. Many successes of such meta-analysis have been reported.^{1,2}

Here, we consider the use of imputation to pool subjects genotyped with different platforms within studies. For example, when the data of control groups such as the Wellcome Trust Case Control Consortium³ are reused, the cases are typically not matched regarding genotyping platforms or arrays.⁴ Another example concerns combining expression quantitative trait loci studies with data being generated at very different time points from different platforms, thereby requiring genotype imputation.⁵ Although reusing such existing data seems to be an efficient approach, it may increase chances of observing spurious associations due to chip differences. In this paper, we discuss whether more stringent quality controls (QCs) should be applied.

In general, the following QCs are performed at the preimputation stage: minor allele frequency (MAF) $\geq 1\text{--}5\%$, Hardy–Weinberg equilibrium (HWE) P -value $> 10^{-4}\text{--}10^{-6}$, SNP call rate $\geq 90\text{--}99\%$, sample call rate $\geq 90\text{--}98\%$, and other checks such as sex mismatch and Mendelian errors. For the details of QCs in GWAS, we refer to Anderson *et al.*⁶ Imputation software such as MACH⁷ or IMPUTE⁸

can be used to impute SNPs based on the HapMap CEU-phased haplotypes. There seems to be no consensus yet on the QCs after imputation, and on reporting the quality of imputed genotypes in publications. In the tutorial of MACH an inclusion threshold r^2 of 0.3 is recommended. In addition to the preanalysis information measures, such as r^2 of MACH and *info* of IMPUTE, which are the information measures about the population allele frequency, SNPTEST⁸ provides a post-analysis information measure about the association parameter for unrelated samples. Here we propose a similar post-analysis information measure to test related samples, called R_T^2 .

As in a meta-analysis, the focus is on combining estimates of association parameters, it seems prudent to base QC on post-analysis information measures that also cover the strength of association, such as SNPTEST *info* or R_T^2 . These measures can be used to obtain homogeneity and to increase the comparability between the studies.⁹ Marchini *et al.*¹⁰ showed that based on a simulated data set of 1000 cases and 1000 controls the MACH and IMPUTE preanalysis information measures were highly correlated, and that there was a good agreement between the IMPUTE preanalysis information measure and the SNPTEST post-analysis information measure when testing an additive genetic model. In this paper we investigate whether good agreement holds for strongly associated SNPs between the pre- and postanalysis information measures, and whether the post-analysis information measures such as SNPTEST *info* and R_T^2 can have an important role as an inclusion criterion of candidate SNPs.

¹Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; ²Netherlands Consortium for Healthy Ageing, Leiden University Medical Center, Leiden, The Netherlands; ³Section of Molecular Epidemiology, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; ⁴Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; ⁵Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands and ⁶Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands

*Correspondence: Dr H-W Uh, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands. Tel: +31 71 5269718; Fax: +31 71 5268280; E-mail: h.uh@lumc.nl

Received 16 April 2011; revised 28 October 2011; accepted 9 November 2011; published online 21 December 2011

MATERIALS AND METHODS

In 2007 we performed a GWAS for the Leiden Longevity Study (LLS)¹¹ with an affected sibling pair (ASP) and control design. One sibling from each of 420 long-lived sibling pairs was genotyped with the first generation Affymetrix Gene Chip Human Mapping 500K Array (Affy500, Perlegen Sciences, Mountain View, CA, USA). This Affy500 data set was discarded for the analysis that was eventually published.¹² To illustrate the situation in which data obtained by an early platform are combined with data generated on more recent platforms, we have here included the Affy500 data yet again. The remaining siblings were genotyped with Illumina Infinium HD Human660W-Quad BeadChips (Illumina660, San Diego, CA, USA). Using the following per-individual QC⁶ of GWA data, we excluded individuals with discordant sex information, individuals with sample call rate < 0.95, and duplicated individuals. Per-marker QC was carried out for including SNPs with the following criteria: SNP call rate > 0.95, MAF > 0.01, and HWE *P*-value > 10⁻⁴. After QC, 517K SNPs remained on the Illumina and 350K SNPs remained on the Affy500 arrays. Of these, only 60K SNPs of Affy500 overlapped with Illumina660. To reuse the genotypes we used MACH for imputation of missing 457K SNPs in Affy500 based on HapMap CEU individuals. To guarantee the quality of imputation, we set the inclusion threshold to *r*²=0.3 as recommended. For 1670 (younger unrelated) controls from the Rotterdam Study, genotypes were generated with Illumina Infinium II HumanHap 550K and HumanHap550-Duo BeadChips (Illumina550).^{12,13} Our data, therefore, differs from the usual simulation setting in the following way: the sib of each sibship genotyped with Affy500 was imputed to match the SNPs of other siblings and controls. The description of the study design and the different arrays used is given in Figure 1 and Table 1.

An additional check of the imputation accuracy was performed; 10% of the SNPs were randomly masked, and correctness of imputation was determined by comparing imputed genotypes with the masked ones. More than 99% of masked SNPs passed the default imputation threshold of *r*²=0.3, so that our data passed this additional QC. For validation of the GWAS results, the 89 top-ranking SNPs were re-genotyped with the Sequenom MassArray platform. Here, we compare imputed and measured genotypes of these top-ranking SNPs.

Methods

Score test. Modeling the LLS data needs to account for (1) ascertainment, that is, cases were long-lived sibling pairs (ASPs), and (2) the fact that one of the sibs in each pair had most markers imputed because it belonged to the Affy500 data. On the basis of the argument that the ascertainment event depends on the phenotype but is conditionally independent of the genotype given a phenotype, we use the score statistic corresponding to the retrospective likelihood for testing.

We let $X=(X_1, \dots, X_n)$ be the $n \times 1$ vector of genotype data. We code each genotype as 0, 1, or 2, corresponding to the number of minor alleles present at that locus. For n individuals, we let $Y=(Y_1, \dots, Y_n)$ be the $n \times 1$ vector of the case-control status, which is coded 0 for control subjects and 1 for case subjects. Further, \bar{Y} denotes the proportion of cases. The score statistic for testing for an additive effect of a diallelic locus on phenotype is given as $U_x=(Y-\bar{Y})X$. Under the null hypothesis of no association between genotype and disease, the score test $U_x^2/Var(U_x)$ is asymptotically distributed as χ^2 with 1 degree of freedom. To account for relatedness of cases we used the kinship coefficients matrix when computing the variance of the score statistic.¹⁴ Imputation is dealt with by accounting for loss of information due to genotype uncertainty. A detailed derivation of the score test is given in the Appendix.

Post-analysis information measures. Let the posterior probability of imputed genotypes be $\pi_i=(\pi_{i0}, \pi_{i1}, \pi_{i2})$ for subject i , and the expected dosage for the genotype counts of the i th individual be $E(X_i)=\pi_{i1}+2\pi_{i2}$. Further, let p denote the population minor allele frequency. Assuming HWE, the MACH *r*² is defined by

$$r^2 = \frac{\sum_{i=1}^n X_i^2/n - \left(\sum_{i=1}^n X_i/n\right)^2}{2\hat{p}(1-\hat{p})}, \tag{1.1}$$

so that this preanalysis information measure depends only on the allele frequency and imputed genotypes. When data are genotyped, *r*² equals one.

As in the Appendix, let K denote the genetic correlation matrix. The genotypic variance of the sample is denoted by Σ , and Σ_{loss} is the loss of information due to uncertainty. The relative efficiency measure for case-control design of Uh *et al*¹⁵ can be used as an information measure about the association parameter:

$$R_T^2 = \frac{(Y-\bar{Y})[K \circ (\Sigma - \Sigma_{loss})](Y-\bar{Y})}{(Y-\bar{Y})[K \circ \Sigma](Y-\bar{Y})}, \tag{1.2}$$

where \circ denotes the (Hadamard) term-wise product. Consequently with genotyped data $\Sigma_{loss}=0$, hence, R_T^2 equals to 1. In contrast to the preanalysis

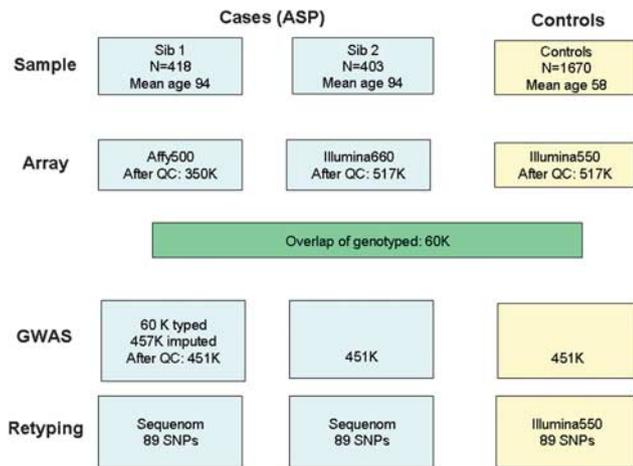


Figure 1 Study samples and arrays used. Affy500 stands for the first generation Affymetrix Gene Chip Human Mapping 500K Array, Illumina660 for Illumina Infinium HD Human660W-Quad BeadChips, and Illumina550 for Illumina Infinium II HumanHap 550K and HumanHap550-Duo BeadChips. Sib 2 and controls were all genotyped, and for Sib1 in addition to the overlapping genotyped 60K SNPs, the remaining 457K SNPs were imputed. After post-imputation QC, 451K SNPs were analyzed using ASP-control design.

Table 1 Study designs and arrays used in Figure 3

Figure 3	Study design	Sample	No. of SNPs ^a	Overlap	Imputed SNPs	QC passed and tested SNPs	Genomic control λ_{GC}
a	ASP-control	Sib 2 and control Sib 1	517K 350K	60K	457K	451K	1.16
b	Case-control	Sib 2 and control	517K	517K		517K	1.03
c	ASP-control	Sib 2 and control Sib 1	517K 350K	60K		60K	1.06
d	ASP-control	Sib 2 and control Sib 1	517K 350K	60K	97K ^b	157K ²	1.05

^aNo. of SNPs that passed QC at the pre-imputation stage.
^bNo. of SNPs with $R_T^2 \geq 0.98$.

information measure r^2 , this post-analysis information measure R_T^2 assigns more weight to associated SNPs.

An executable C++ program for the score test and R_T^2 is available (<http://www.msbi.nl/uh>).

RESULTS

The difference between the pre- and postanalysis information measures, MACH r^2 and R_T^2 , is shown in Figure 2. Using Sib 1 and controls data, we randomly selected 1000 SNPs each from three classes of SNPs: P -values > greater than 0.05, P -values smaller than 0.001, and intermediate ones. Although for unassociated SNPs (P -value > 0.05) the two measures show good agreement, they are quite different for strongly associated SNPs (P -value < 0.001). The post-analysis measure, therefore, can be a useful tool for selecting SNPs for meta-analysis.

Quantile–quantile (Q–Q) plots in Figure 3 illustrate the GWAS results using different study designs as described in Table 1. The test statistics in all Q–Q plots were corrected by their genomic control inflation factor λ_{GC} .¹⁶ First we used combined data of ASPs (imputed Sib 1 and genotyped Sib 2) and genotyped controls. Results (Figure 3a) show deviation from first diagonal (dashed line), hence, inflation of test statistics ($\lambda_{GC}=1.16$). Next (Figure 3b), we compared genotyped Sib 2 and controls (Illumina660 for cases and Illumina550 for controls, respectively): $\lambda_{GC}=1.03$. One might conjecture that inflated test statistics in Figure 3a were caused by also considering imputed sibling data. We then investigated whether this inflation is an artifact solely from imputation, or due to combining different arrays. To determine the possibility of a chip (or batch) effect, we conducted ASP and control analysis only on genotyped overlapping 60K SNPs with Affy500 (Sib 1), Illumina660 (Sib 2), and Illumina550 (control). In Figure 3c, the genomic control inflation factor is decreased from 1.16 to 1.06 as compared with Figure 3a and increased from 1.03 to 1.06 as compared with Figure 3b. This may suggest that there is a chip-effect, which was amplified by the imputation. Figure 3d shows that by applying a very stringent extra QC ($R_T^2 > 0.98$, 60K genotyped and 97K imputed SNPs) inflation of test statistic could be dealt with ($\lambda_{GC}=1.05$). Therefore, the significantly biased results (Figure 3a) seem to be caused by the different chips from one of which is of low quality.

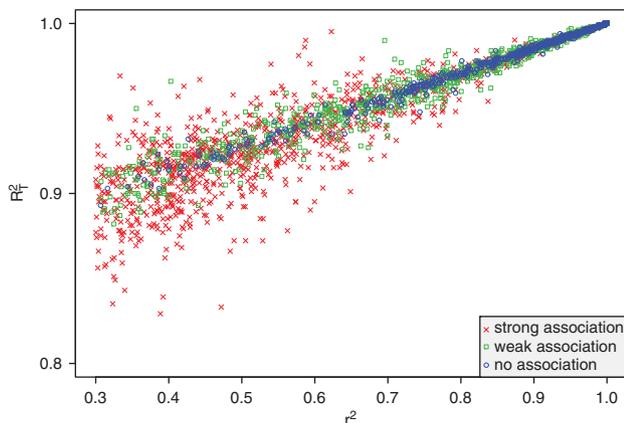


Figure 2 Comparison of the pre- and the postanalysis imputation information measure. The x axis shows the preanalysis information measure (r^2), and the y axis the post-analysis information measure (R_T^2). The blue points indicate the SNPs with no association (P -value > 0.05); there is little effect of case–control status, and two measures agree. The red ones are the SNPs that show strong association (P -value < 0.001), and the green ones are intermediate cases.

For validation, the 89 top-ranking SNPs (MACH $r^2 > 0.3$) resulting from the association analysis using the first design were retyped with the Sequenom MassArray platform. We checked the quality of genotyping (of the different platforms) as well as that of imputation. Figure 4 illustrates the comparison of minor allele frequencies (MAFs) in the long-lived siblings. In the left panel, the deviation of the points from first diagonal (dashed line) indicates the poor match of the Affy500 data and retyped sample. Meanwhile, the retyping of the Illumina660 data shows better agreement (bottom panel). Visual inspection of cluster plots of the sole exception (the red filled circle) confirmed the results of the Sequenom array.

DISCUSSION

Our study illustrates that imputation, whereas combining different arrays in GWAS using data from the earliest platforms without sufficiently stringent QCs may produce false positive associations. A simple remedy to better quality is to choose a stricter threshold for inclusion at the pre- and postimputation stages. For preimputation QCs we refer to Anderson *et al.*⁶

In addition to the preanalysis measures such as r^2 of MACH and *info* of IMPUTE, which are the relative information measures only depending on the population allele frequency and imputation accuracy, we proposed an additional post-analysis measure R_T^2 .

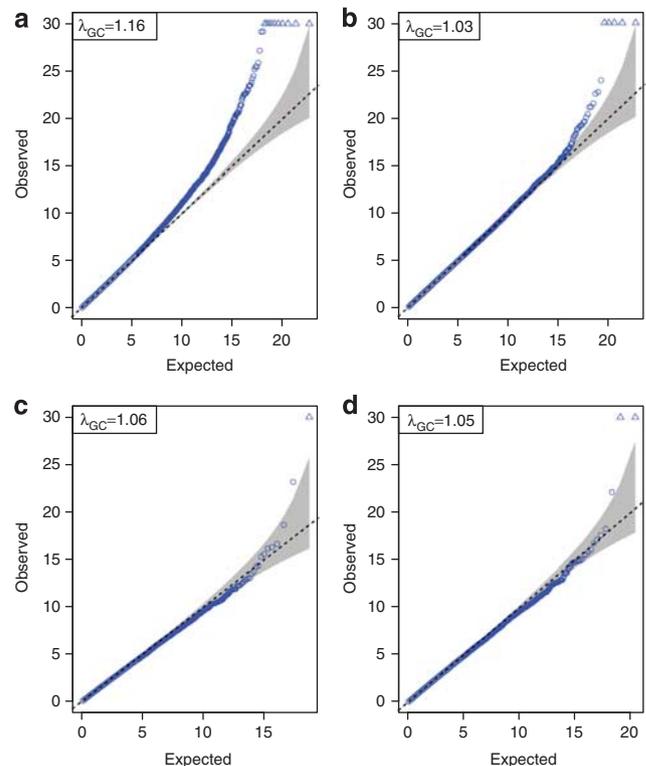


Figure 3 Quantile–quantile plots obtained from LLS GWAS analyses. The triangles indicate the SNPs at which the test statistic exceeds 30 (corresponding P -value < 5×10^{-8}). The 95% concentration bands (shaded gray) are included. (a) ASP–control design: combined data of imputed Affy500 (Sib 1), typed Illumina660 (Sib 2), and typed Illumina550 (control). Deviation from the dashed line indicates inflation of test statistics. (b) Case–control design: genotyped with Illumina660 (Sib 2) and Illumina550 (control). (c) ASP–control design: 60K overlap using combined typed data of Affy500 (Sib 1), Illumina660 (Sib 2), and Illumina550 (control). (d) ASP–control design: as in (a), but only SNPs with $R_T^2 > 0.98$. Details are provided in Table 1.

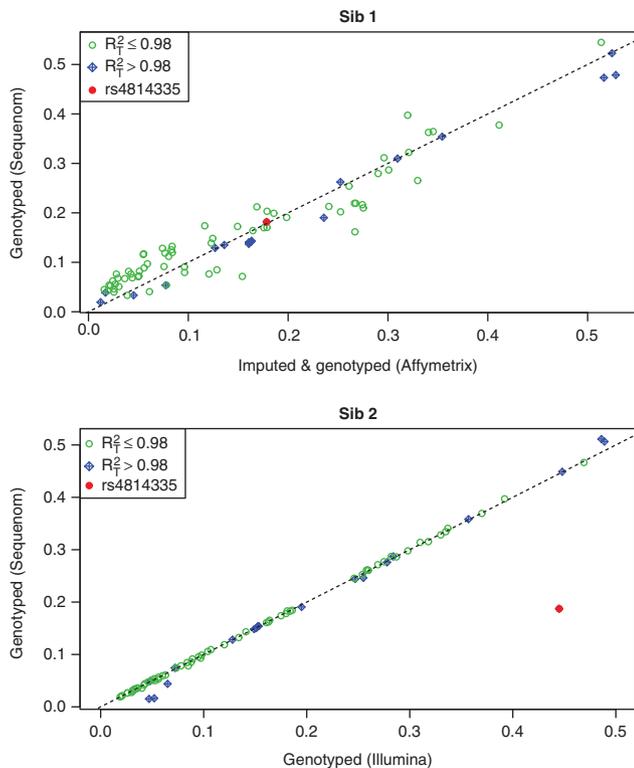


Figure 4 Comparison of the MAF between GWAS and replication data. Top: x axis shows MAF of imputed Sib 1 data using Affy500, and y axis MAF of the same SNPs replicated with Sequenom. The green colored did not pass the threshold $R_T^2 > 0.98$. Bottom: x axis shows MAF of (genotyped) Sib 2 data using Illumina660, and y axis MAF of the same SNPs replicated with Sequenom. The red-filled circle in both panels indicates the same SNP.

Our measure is an information measure that assesses the above information but also includes strength of association. When testing independent samples, this is equivalent to the information measure of SNPTEST. For a recessive or dominant model, Marchini *et al*¹⁰ showed that the post-analysis measures are quite different from the preanalysis information measure r^2 . For strongly associated SNPs under an additive model we showed that R_T^2 and r^2 could be quite different (Figure 2). For example, meta-analyses aim to combine estimates of association parameters, which argues for the use of post-analysis QC measures such as R_T^2 and SNPTEST *info*. In situations such as ours, filtering on R_T^2 leads to a reduction in heterogeneity between studies, making the studies more comparable and meta-analysis more powerful. To interpret the results of meta-analysis properly, it also is important to report the difference between the studies, such as the quality of both genotyping and imputation.

All information measures need to be carefully considered in further analysis. In our study, by re-genotyping strongly associated SNPs, we found that an extremely tight inclusion threshold of our imputation quality measure R_T^2 greater than 0.98 was needed to achieve reliable results as shown in Figures 3 and 4; only 18 from the 89 top-ranking SNPs passed the post-analysis QC. These plots suggest that false positive findings are caused by imputation based on arrays of inferior quality, when cases and controls are not matched for genotyping platforms. Actually, in our GWAS for longevity we discarded the Affy500 data set because of the small number of reliable SNPs. It should be noted that 97K imputed SNPs remained in the analysis even for this stringent cutoff (Table 1). We also retyped the Affy500 cases with the Illumina 660K platform and recently published our GWAS.¹²

In Figure 3c one may ask whether the Q–Q plot using only 60K overlapping SNPs is comparable to Q–Q plots using larger number of SNPs. We compared the distribution of association P -values using 60K cases and controls and 350K cases and controls, and both distributions were quite similar (data not shown).

The results presented here, were based on an early scan data with a small sample size. When combining modern arrays within studies, less bias may be expected due to better genotyping quality. On the other hand, the enormous sample size of pooled studies may amplify even the small individual effects, for example, due to platform effects, population strata, or genotyping batch effects, resulting in false positive findings, as heterogeneity between studies is amplified by imputation. Imputation of genotypes while combining different data sets can be a very powerful method, and has identified susceptibility loci using early scan data.^{17,18} However, our findings stress that when combining newer data sets with early scan data rigorous QCs should be applied to ensure reproducible findings including pre- and post-analysis stages. Moreover, we recommend that post-analysis QC measures should be reported in publications as they give the most direct insight into influence of imputation on association.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We acknowledge R van der Breggen, N Lakenberg, D Kremer, and HED Suchiman for their efforts in genotyping by Sequenom MassArray. This work is supported by a grant from the Netherlands Organization for Scientific Research (NWO 917.66.334). We thank all the participants of the Leiden Longevity Study and the Rotterdam Study. This study was supported by a grant from the Innovation-Oriented Research Program on Genomics (SenterNovem IGE05007), the Centre for Medical Systems Biology, and the Netherlands Consortium for Healthy Ageing (Grant 050–060-810), all in the framework of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NWO), and BBMRI-NL (Biobanking and Biomolecular Resources Research Infrastructure). The generation and management of GWAS genotype data for the Rotterdam study is supported by the Netherlands Organization for Scientific Research NWO Investments (No. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2) and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Project No. 050-060-810; we thank P Arp, M Jhamai, M Verkerk, L Herrera, and M Peters for their help in creating the GWAS database. The Rotterdam Study is funded by the Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly, the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam.

- Li Y, Willer C, Sanna S, Abecasis G: Genotype imputation. *Annu Rev Genomics Hum Genet* 2009; **10**: 387–406.
- Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.
- The Wellcome Trust Case Control Consortium: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
- ANZ genes: Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosome 12 and 20. *Nat Genet* 2009; **41**: 824–828.
- Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE: Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet* 2010; **86**: 581–591.
- Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT: Data quality control in genetic case-control association studies. *Nat Protoc* 2010; **5**: 1564–1573.

- 7 Li Y, Abecasis G: Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet* 2006; **579**: 2290.
- 8 Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet* 2007; **39**: 906–913.
- 9 Cantor RM, Lange K, Sinsheimer JS: Prioritizing GWAS results: a review of statistical methods and recommendations for their approach. *Am J Hum Genet* 2010; **86**: 6–22.
- 10 Marchini J, Howie B: Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**: 499–511.
- 11 Westendorp RG, van Heemst D, Rozing MP *et al*: Nonagenarian siblings and their offspring display lower risk for mortality and morbidity than sporadic nonagenarians: the Leiden Longevity Study. *J Am Geriatr Soc* 2009; **59**: 1634–1637.
- 12 Deelen J, Beekman M, Uh HW *et al*: Genome-wide association study identifies a single major locus contributing to survival into old age; the *APOE* locus revisited. *Ageing Cell* 2011; **10**: 686–698.
- 13 Hofman A, Breteler MM, Van Duijn CM *et al*: The Rotterdam Study: 2010 objectives and design update. *Eur J Epidemiol* 2009; **24**: 553–572.
- 14 Uh HW, Wijk HJ, Houwing-Duistermaat JJ: Testing for genetic association taking into account phenotypic information of relatives. *BMC Proc* 2009; **5**(Suppl 7): S123.
- 15 Uh H-W, Houwing-Duistermaat JJ, Putter H, van Houwelingen HC: Assessment of global phase uncertainty in case-control studies. *BMC Genet* 2009; **10**: 54.
- 16 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
- 17 Stuart PE, Nair RP, Ellinghaus E *et al*: Genome-wide association analysis identifies three psoriasis susceptibility loci. *Nat Genet* 2010; **42**: 1000–1004.
- 18 Ellinor PT, Lunetta KL, Clazer NL *et al*: Common variants in *KCNN3* are associated with lone atrial fibrillation. *Nat Genet* 2010; **42**: 240–244.



This work is licensed under the Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 Unported Licence. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>

APPENDIX

We first address the ascertainment of the independent cases. Let $Y=(Y_1, \dots, Y_n)$ be the phenotype, $X=(X_1, \dots, X_n)$ denotes genotype dosage 0, 1, or 2. Further, \bar{Y} is the mean of Y in the whole sample, or the proportion of cases in case-control studies. As the ascertainment event S depends on the phenotype but is conditionally independent of the genotype given Y , $P(X|Y, S)=P(X|Y)$. Therefore, the retrospective likelihood based on $P(X|Y)$ is appropriate under selection. On the basis of retrospective likelihood, the score statistic for testing for an additive effect of a genotyped locus on phenotype is as follows. The score is,

$$U_X = (Y - \bar{Y})^T X, \quad (1)$$

and the variance of U_X

$$\text{Var } U_X = (Y - \bar{Y})^T (Y - \bar{Y}) \sigma_X^2, \quad (2)$$

where σ_X^2 is the genotypic variance. Under HWE assumption, σ_X^2 can be estimated by $2\hat{p}(1 - \hat{p})$ with the MAF estimate \hat{p} . Under H_0 , the test statistic $U_X^2/\text{Var } U_X$ is asymptotically distributed as χ^2 with 1 degree of freedom.

When using multiplex cases from the same pedigree, we need to take into account correlations. We define the correlation matrix K for n subjects as follows:

$$K = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{12} & 1 & \dots & \rho_{1n} \\ \vdots & \dots & \dots & \vdots \\ \rho_{1n} & \rho_{2n} & \dots & 1 \end{pmatrix}$$

The off-diagonal entries, ρ_{ij} s, are twice the kinship coefficient between individuals i and j ($i \neq j$). Then, the expression of the denominator of the score statistic is replaced by

$$\text{Var } U_X = (Y - \bar{Y})^T K (Y - \bar{Y}) \sigma_X^2. \quad (1)$$

To deal with imputed genotypes, the uncertainty caused by imputation needs to be considered. On the basis of the statistical theory for missing data, the genotype data can be partitioned into two parts

$$X_{\text{comp}} = [X_{\text{obs}}, X_{\text{mis}}]. \quad (2)$$

The log likelihoods for the complete data (l_{comp}) and observed (incomplete) data (l_{obs}) are given by

$$l_{\text{comp}}(\theta) = \log P(X_{\text{obs}}, X_{\text{mis}}|\theta),$$

$$l_{\text{obs}}(\theta) = \log \int P(X_{\text{obs}}, X_{\text{mis}}|\theta) d, X_{\text{mis}}$$

Let $U(\theta)$ be the complete data score $\partial l_{\text{comp}}/\partial \theta$, and $I(\theta)$ the complete data information $-\partial^2 l_{\text{comp}}/\partial \theta^2$, respectively.

Instead of observing X , for imputed genotypes the posterior probability $\pi_i=(\pi_{i0}, \pi_{i1}, \pi_{i2})$ is given for subject $i=1, \dots, n$. Let the expected dosage for the genotype counts of the i th individual be $\bar{X}_i=EX_i=\pi_{i1}+2\pi_{i2}$. Then we replace the genotype counts X by

$$U_{\bar{X}} = (Y - \bar{Y})^T \bar{X}$$

in the score statistic (1).

Let $\Sigma = \sigma_X^2 \mathbf{1} \mathbf{1}^T$ be $n \times n$ matrix with the genotypic variance σ_X^2 where $\mathbf{1}$ represents a vector of ones of length n . And, the $n \times n$ matrix Σ_{loss} denotes the loss of information.

Then, the score and information for the observed data likelihood are given by

$$U_{\text{obs}}(\theta) = EX_{\text{mis}}|X_{\text{obs}} U(\theta),$$

$$I_{\text{obs}}(\theta) = EX_{\text{mis}}|X_{\text{obs}} I(\theta) - \text{Var}_{X_{\text{mis}}|X_{\text{obs}}} U(\theta) = \sum - \sum_{\text{loss}}$$

Here, the term $\text{Var}_{X_{\text{mis}}|X_{\text{obs}}}(\cdot)$ represents the loss of information due to imputation uncertainty. The elements of Σ_{loss} are defined by the outer product of the square root of individual loss l_i ,

$$l_i = \pi_{i1}(1 - \pi_{i1}) + 4\pi_{i2}(1 - \pi_{i2}) - 4\pi_{i1}\pi_{i2}$$

Thus, on the diagonal we have $\Sigma_{\text{loss};ii}=l_i$ and off the diagonal we have

$$\sum_{\text{loss};ij} = \sqrt{l_i l_j}$$

for $i, j=1, \dots, n$. Then the variance of the score statistic can be expressed as

$$\text{Var}_{X_{\text{obs}}} U_{\bar{X}} = n^{-1} (Y - \bar{Y})^T \left[K \circ \left(\sum - \sum_{\text{loss}} \right) \right] (Y - \bar{Y}),$$

where \circ denotes the (Hadamard) term-wise product.

References

- Uh HW, Wijk HJ, Houwing-Duistermaat JJ: Testing for genetic association taking into account phenotypic information of relatives. *BMC Proc* 2009; (Suppl 7): S123.
- Louis TA: Finding the observed information matrix when using the EM algorithm. *J R Stat Soc* 1982; **44**: 226–233.