Multi-Criteria Decision Making Approaches for Quality Control of Genome-Wide Association Studies

Alberto Malovini, MS\textsuperscript{1,2}, Carla Rognoni, PhD\textsuperscript{2}, Annibale Puca, MD\textsuperscript{1}, Riccardo Bellazzi, PhD\textsuperscript{2}

\textsuperscript{1}IRCCS Multimedica, Milano, Italy; \textsuperscript{2}Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy

Abstract

Experimental errors in the genotyping phases of a Genome-Wide Association Study (GWAS) can lead to false positive findings and to spurious associations. An appropriate quality control phase could minimize the effects of this kind of errors. Several filtering criteria can be used to perform quality control. Currently, no formal methods have been proposed for taking into account at the same time these criteria and the experimenter's preferences. In this paper we propose two strategies for setting appropriate genotyping rate thresholds for GWAS quality control. These two approaches are based on the Multi-Criteria Decision Making theory. We have applied our method on a real dataset composed by 734 individuals affected by Arterial Hypertension (AH) and 486 nonagenarians without history of AH. The proposed strategies appear to deal with GWAS quality control in a sound way, as they lead to rationalize and make explicit the experimenter's choices thus providing more reproducible results.

Introduction

Data quality control is a crucial aspect of Genome-Wide Association Studies (GWAS) analysis to avoid false positives and to properly carry on data interpretation. Experimental systems involving biological material are typically prone to errors: if those errors would be randomly distributed across both genotypes and phenotypes, their effect would be limited to a loss of statistical power. However errors are often non-randomly distributed. This lacks of randomness is both due to the very nature of the available experimental technologies and to the presence of several concurrent factors such as: DNA quality and preparation, specific experimental conditions, different skills of the experimenters, incorrect automated assignment (or "calling") of experimental intensity values into discrete genotype classes. Non random distribution of errors can inflate type I error rates and of course reduce the power of the study. Since most GWAS aim to identify very slight variations in allele frequencies between groups of individuals (cases and controls), even the presence of small experimental errors could dramatically affect the outcome\textsuperscript{1,2}. The ideal solution to this kind of problems would be to identify individual genotype errors and correct them one at a time, but this is almost impossible in GWAS. A feasible alternative is to apply filtering procedures in order to identify specific SNPs yielding errors in multiple individuals (markers-affecting errors), or individuals in the sample with errors across multiple SNPs (problems with the DNA sample), and simply exclude them from the analysis. Missing individual SNPs or individual genotypes are frequent, but when missing rate exceeds 5\% for SNPs or 10\% for a sample, there could be evidence of genotyping error, and a good choice would be to exclude that marker/DNA sample from the analysis and repeat the experiment\textsuperscript{3}.

Several parameters can be used to evaluate the effect of the removal of SNPs and individuals on the quality of the association study: heterozygosity rate, SNPs deviations from the Hardy Weinberg Equilibrium (HWE) in the control population or percentage of markers with significant difference in missing data rates between cases and controls. Unfortunately, the setting of these parameters is often unclear and subjective, leading to a lack of reproducibility and false positive findings. It could be therefore useful to exploit more systematic approaches in order to help researchers in these crucial steps of the analysis.

Currently, no formal methods have been proposed for choosing the most appropriate filtering criteria that take into account all of these factors. In this paper we propose to deal with this problem with a normative approach based on decision theory. We have applied two different Multi-Criteria Decision Making (MCDM) methods to the quality control steps of a GWAS for selecting the best filtering thresholds, in order to minimize the type I errors and maximize the statistical power related to the sample size.

We have tested our method on a real dataset composed by 1220 individuals (734 cases, 486 controls) genotyped using Illumina Infinium II and HumanHap 300 chips [Illumina, San Diego, CA,
Methods: Parameters for quality control

As described above, the main decisions that should be made by the data analyst for quality control in GWAS are the maximum number of missing genotyped SNPs per individual and the maximum number of missing individuals per SNPs. Once a subset of SNPs and individuals are selected, different parameters are evaluated to check the quality of the available data.

For case-control studies, markers with significant difference in missing data rates between cases and controls at a fixed significance threshold should be excluded from the analysis since they could yield false positive associations. Furthermore, SNPs that are the maximum number of missing genotyped SNPs per individual and the maximum number of missing individuals per SNPs. Once a subset of SNPs and individuals are selected, different parameters are evaluated to check the quality of the available data.

Neutral genetic variants in a large random-mating population are expected to display HWE, under which assumption expected genotype frequencies satisfy $E(MM)=p^2$, $E(Mm)=2pq$, $E(mm)=q^2$, where $p$ and $q$ are the frequency of $M$ and $m$ alleles in the population, respectively. Genotyping errors can shift the SNPs observed frequencies from the expected proportions, and therefore testing for deviations from the HWE in the control population represents a standard approach to detect genotyping errors. Such test can be performed using a Pearson goodness-of-fit statistic with one degree of freedom (d.f) under the null hypothesis of HWE. Some problems are related to the application of this quality-filter: (1) biological factors can lead SNPs to deviate from HWE potentially related to true causal association, which will be missed if the marker fails the test; (2) the significance threshold of the statistic is often unclear.

Graphical displays such as plots and histograms could represent useful tools to assess the quality of the data and to identify potential low-quality samples and markers. One graphical inspection approach involves plotting the missing rate for each individual against the fraction of markers that are heterozygous. This graphical approach can highlight low quality/cross-contaminated DNA samples showing extreme heterozygosity rate and it could be useful for the choice of an appropriate genotyping rate threshold. It could be also useful to observe the relationship between missing rate statistic for each individual and the genomic inflation factor value $\lambda$ (computed by dividing the median of the test statistics by 0.456, the expected median of a $\chi^2$ distribution with 1 d.f.) 6: an index of population stratification that could be also inflated by the presence of low-quality DNA samples.

Methods: Multi-criteria decision making

The choice of the best genotyping rate thresholds for individuals and SNPs can be considered as a MCDM problem, that implies to identify and choose alternatives based on the values and preferences of the decision maker. One paradigm of the MCDM is the Multiple Attribute Decision Making (MADM) approach that requires that alternatives are described by a set of attributes or criteria (e.g. in our case the parameters described above).

MADM problems are assumed to have a finite number of alternatives and solving the problem means assigning a score to each alternative and then sorting and ranking them. Applications of MCDM require that preferences associated with the alternatives are independent of each other from one criteria to another one. This means that the preference scores for all alternatives on a criteria are assigned with no knowledge about the alternatives preference scores on any of the other criteria.

Once identified the criteria relevant for the problem, the next step is to rank the alternatives against each criteria from best to worst, from most preferred to least preferred: the decision maker is asked to assign a score measure to each alternative against each criteria. A value scale from 0 to 100 can be used, where the best preference is 100 and the worst is 0. The assigned score measures are then represented in the evaluation matrix ($S$). Considering $n$ alternatives and $m$ criteria, $S$ is a $m \times n$ matrix and the $S_{ij}$ element represents the score of alternative $Ai$ according to criteria $Cj$. We assumed that a higher score value means a better performance since any goal of minimization can be easily transformed into a maximization goal. This method provides the basic framework for collecting and organizing the information. In this case values are quantitative but MCDM methods allow also the use of qualitative measures.

Usually criteria are measured in different units, so the evaluation matrix has to be normalized. There are different procedures for data standardization but the most frequently used are equations:

1) $a_{ij}=S_{ij}/\max(S_{ij})$

2) $a_{ij}=(S_{ij}-\min(S_{ij}))/\{(\max(S_{ij})-\min(S_{ij}))\}$

For all MADM models is required that the choice of the best alternative is made taking into account the relative importance of the criteria, so a weight is assigned to each criteria. Criteria weights can be assigned directly by the decision-maker or through a weighting methodology.
In our case a method based on the Analytic Hierarchy Process (AHP) seemed the most appropriate. The basic idea is to convert subjective assessments of relative importance to a set of overall weights\textsuperscript{10}. The AHP is based on pairwise comparisons among criteria. For each pair of criteria, the decision maker is required to answer to a pairwise comparison question asking the relative importance of the two. A nine-point scale can be used to express the intensity of the preference (1= Equal importance or preference; 3=Moderate importance or preference of one over another; 5=Strong or essential importance or preference; 7=Very strong or demonstrated importance or preference; 9=Extreme importance or preference). If it is judged that criteria \( C_i \) is more important of criteria \( C_j \), then the reciprocal of the relevant index is assigned (\( c_{ij} \)). The \( c_{ij} \) elements are represented into a comparison matrix \( C \) of size \( m \times m \).

The matrix \( C \) is used to compute the weights for each criteria that are represented in the form of a vector \( w=(w_1, w_2, \ldots, w_m) \). There are different approaches to obtain it but the least squares method is simple and straightforward. This method consists of calculating the geometric mean of each row of the matrix \( C \), calculate the sum of the geometric means and normalize each of the geometric means by dividing by the sum just computed\textsuperscript{11}.

The Simple Multi-Attribute Rating Technique (SMART) is the simplest method among MADM. The ranking value \( x_j \) of the alternative \( A_j \) is obtained by the following formula under an additive model:

\[
x_j = \left( \sum_{i=1}^{m} w_i \alpha_i + \sum_{i=k}^{m} w_i (1-\alpha_i) \right) / \sum_{i=1}^{m} w_i \quad j=1,\ldots,n
\]

The best alternative is the one with the highest ranking value.

The second approach we have tested represents an alternative MCDM strategy that implements a different procedure for criteria weights assignment, based on direct elicitation of user preferences.

We will refer to this method as D-MCDM. In this case the decision maker is not required to answer to a pairwise comparison question about the relative importance between two criteria but he has to assign weights directly to each of them in a 0 to 10 scale.

The score for each alternative is obtained by the following formula, where \( k \) represents the number of criteria that have to be maximized and \( m-k \) is the number of criteria that should be minimized:

Results

Our medical staff recruited 734 patients with high blood pressure with age ranging between 35-55 years. The control population is represented by 486 nonagenarians collected during the course of the last few years. After approval of the ethical committee and under informed consent collected following the Italian law, blood has been drawn from every patient participating to the study. DNA has been extracted and anamnestic, clinical and laboratory data have been collected. All samples were assayed with the Illumina Infinium II and HumanHap 300 chips containing 318,237 PhaseII Hap map tagging SNPs. Data were exported by Illumina Bead Studio Software with an average genotyping rate of 97%.

Using gPLINK\textsuperscript{12} we simulated to apply 10 different individual missing rate filters: from 10% (remove individuals with genotyping rate <90%) to 1% (remove individuals with genotyping rate <99%). For each individual missing/genotyping filtering threshold, we applied 10 independent SNP missing rate/genotyping rate thresholds: from 10% (remove markers with genotyping rate <90%) to 1% (remove markers with genotyping rate <99%). For every combination of individual and SNP missing/genotyping rate filters we have computed the following parameters:

- percentage of individuals (C1) and SNPs (C2) that pass the genotyping filters
- percentage of SNPs excluded because of significant difference in missing data rates between cases and controls (p<0.05) on the basis on the Fisher exact test, as implemented in gPLINK (C3)
- percentage of SNPs with MAF <1% (C4)
- percentage of SNPs deviating from the HWE in the control population (p<0.001) (Pearson goodness-of-fit statistic with one d.f.)\textsuperscript{5}, (C5)
- heterozygosity rate standard deviation per individual (C6)
- genomic inflation factor (\( \lambda \)) computed by taking the median of the distribution of the chi-square statistic from results of the Armitage Trend Test performed over the set of markers from the study, and dividing this
value by 0.456, the median of the corresponding \( \chi^2 \) distribution\(^6\)(C7).

We applied the two MCDM methods considering a pair of individual and SNP genotyping rates as an alternative. The alternatives represent all the possible combinations of these two parameters (10x10). The parameters C1-C7 have been considered the criteria of our study. Based on the 7 criteria and the 100 alternatives, we built an evaluation matrix \( S \) of size 7x100 (a portion of the evaluation matrix \( S \) is represented in Table 1).

In this work, the matrix \( S \) has been standardized using equation 1. Then we built the comparison matrix \( C \) (Table 2). Since the statistical power of a study relies on the number of analyzed samples, we gave high relative importance to this parameter, trying to keep the percentage of analyzed individuals high. In order to take into account in the study as many markers as possible, we maximized a second parameter: the percentage of SNPs that pass the genotyping filter. Otherwise we chose to minimize the values of the remaining criteria, since they may be index of low-quality data. The matrix \( C \) has been used to compute the weights for each criteria as previously described. The obtained vector \( w \) is \((0.33, 0.07, 0.07, 0.07, 0.07, 0.33, 0.07)\). By applying the SMART formula we calculated the scores corresponding to each alternative: the highest score corresponds to the alternative “individual genotyping rate >95% and SNP genotyping rate >96%”, while the lowest score corresponds to the alternative “individual genotyping rate >90% and SNP genotyping rate >99%” (Figure 1A).

Then, we performed the same analysis using the D-MCDM approach, setting the same relative importance assumptions for each criteria and performing standardization by equation 2. The highest score obtained corresponds to the alternative “individual genotyping rate >95% and SNP genotyping rate >97%”, while the lowest score corresponds to the alternative “individual genotyping rate >90% and SNP genotyping rate >99%” (Figure 1B).

Finally, in order to perform a comparison of the results, we plotted the different alternatives against the corresponding normalized scores obtained by applying the two previously described strategies for individual genotyping rate 95% and 99% (Figure 2).

### Table 1. Evaluation matrix \( S \)

| Criteria | Alternatives |
|----------|--------------|
|          | CR90,SNP90   | CR90,SNP91   |
| C1       | 100          | 99           |
| C2       | 7            | 7            |
| C3       | 96           | 96           |
| C4       | 35           | 40           |
| C5       | 90           | 90           |
| C6       | 100          | 100          |
| C7       | 53           | 53           |

### Table 2. Comparison matrix \( C \)

| C1 | C2 | C3 | C4 | C5 | C6 | C7 |
|----|----|----|----|----|----|----|
| 1  | 5  | 5  | 5  | 5  | 1  | 5  |
| 1/5| 1  | 1  | 1  | 1  | 1/5| 1  |
| 1/5| 1  | 1  | 1  | 1  | 1/5| 1  |
| 1/5| 1  | 1  | 1  | 1  | 1/5| 1  |
| 1  | 5  | 5  | 5  | 5  | 1  | 5  |
| 1/5| 1  | 1  | 1  | 1  | 1/5| 1  |

### Figure 1. SMART (A) and D-MCDM (B) score profiles (“CR” and “SNP” indicate individual and SNP genotyping rate, respectively)

### Discussion

The results we obtained by using the two strategies are very similar for individual genotyping rates <95%, with comparable score profiles. For individual genotyping rates >96% the interpretation of the two profiles is more complex: the profiles derived from the D-MCDM seem to decrease more rapidly than the ones generated by the SMART approach. D-MCMD penalizes more extreme choices, in particular high values of the genotyping rates: by applying very stringent samples genotyping filters we risk to lose
useful data, with a decrease of the sample size and, therefore, statistical power (Figure 1). This result is due to the elicitation process of the criteria weights, which is done independently for each criteria. The process may lead to sharper shapes of the score function. On the contrary, SMART is able to take into account correlations between criteria and therefore is related to smoother score functions (Figure 2).

![Figure 2](image)

**Figure 2.** SMART and D-MCDM score profiles for individual genotyping rate 95% (A) and 99% (B)

However, SMART is a more “expensive” model in terms of time required for the elicitation process, since it requires pairwise comparison for the relative weights assignment. D-MCDM, instead, only requires to give a weight to each criteria, which can be easily obtained by a direct score elicitation strategy, such as ranking–based ones⁹.

Conclusions

The strategies presented in this paper can be considered as instruments to perform a quality control dealing with GWAS. In particular they lead to set appropriate thresholds, rationalizing and making explicit the experimenter’s choices with the aim to provide more reproducible results. As shown in this paper, the experimenter is asked to provide a vector (D-MCDM) or a matrix of weights (SMART); then, after computing the evaluation matrix $S$, the method provides a principled way to take decisions and an instrument to understand if decisions may change when case different weighting strategies are chosen.

Finally, our methodology can be easily applied to other datasets where researchers can define their own criteria and weights. By making them explicit, the results obtained can be fully reproducible by other research groups, even starting from the row data.

Acknowledgments

This study is part of the FIRB project ITALBIONET, Rete Nazionale di Bioinformatica. A. Malovini was supported by the FIRB project “Internazionalizzazione Italia-Israele”. We are grateful to the Multimedica staff (Valeria Novelli, Roberta Roncarati and Chiara Viviani Anselmi) for data collection.

References

1. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet. 2005; 37:1243–1246.
2. Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet. 2006;38(6):659-62. Epub 2006 May 2.
3. Welcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 Jun 7;447(7145):661-78.
4. Weir BS. Genetic Data Analysis II. Sinauer Associates Inc: Sunderland, Massachusetts, USA. 1996;112-133.
5. Wigginton JE, Cutler DJ and Abecasis GR. A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am J Hum Genet. 2005;76:887-93.
6. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic based association studies. Theor. Popul. Biol. 2001;60: 155–166.
7. Gal T, Stewart TJ, Hanne T. Multicriteria Decision Making - Advances in MCDM Models, Algorithms, Theory and Applications Series: International Series in Operations Research & Management. Science. 1999;21:560 p.
8. Valiris G, Chytas P, Glykas M. Making decisions using the balanced scorecard and the simple multi-attribute rating technique. 2005.
9. Nijkamp P, Rietveld P, Voogd H. Multicriteria evaluation in physical planning. North Holland. Amsterdam. 1990.
10. Saaty TL. The Analytic Hierarchy Process. McGraw-Hill, New York. 1980.
11. Saaty TL, Vargas L. Comparison of eigenvalue, logarithmic least squares and least squares methods in estimating ratios. Mathematical Modelling. 1984; 5:309-324.
12. [http://pngu.mgh.harvard.edu/~purcell/plink](http://pngu.mgh.harvard.edu/~purcell/plink)