Disease Markers 27 (2009) 173–186
DOI 10.3233/DMA-2009-0654
IOS Press

Genetic polymorphisms of infectious diseases in case-control studies

Antonio G. Pacheco[a,*] and Milton O. Moraes[b]
[a]*Programa de Computação Científica, FIOCRUZ, Av. Brasil, 4365, Manguinhos, 21045-360, Rio de Janeiro, RJ, Brazil*
[b]*Laboratório de Hanseníase, Instituto Oswaldo Cruz, FIOCRUZ, Av. Brasil 4365, CEP 21040-360 Manguinhos, Rio de Janeiro, RJ, Brazil*

Abstract. In the past decade, genetic epidemiological analyses in infectious diseases have increased drastically since the publication of human genome and all the subsequent projects analyzing human diversity at molecular level. The great majority of studies use classical epidemiological designs applied to genetic data, and more than 80% of published studies use population-based case-control designs with widely spread genetic markers in human genome, like short tandem repeats (STR) or single nucleotide polymorphisms (SNP), in genes chosen by their physiological association with the disease (candidate genes). Even though genetic data is less prone to several bias issues inherent to case-control studies, some care has to be taken when designing, performing, analyzing and interpreting results from such studies. Here we discuss some basic concepts of genetics and epidemiology as a departure to evaluate and review every step that should be followed to design, conduct, analyze, interpret and present data from those studies, using particularities of infectious diseases, especially leprosy and tuberculosis as models.

Keywords: Population, SNPs, cytokines, tuberculosis, leprosy, TNF, IL-10

1. Introduction

The explosion of studies using genetic epidemiology, i.e. case controls studies with candidate genes, of infections have risen after common belief that genes greatly influence susceptibility to infectious diseases. There is a list of epidemiological evidence reinforcing this idea: it is estimated that susceptibility to infectious diseases is a phenomenon that occurs in a very variable percentage, ranging from 0.1% to 80%, of the exposed population. In diseases like leprosy it is believed that a very small fraction (0.1–1%) develops the disease, while malaria parasites exhibit a high rate of infection success (around 80% of exposed people exhibit the outcome). Other evidence is linked to the presentation of the clinical symptoms, as in most infections there is a range from mild to severe states of the disease and generally it is not common to observe a patient migrate from one form of the disease to another. Finally, twin and familial studies, especially in leprosy and tuberculosis, provide the idea of genetic inheritance [1,2] that has been consistently depicted in genome scans of families [3–5].

Among all of the genes that participate in immune response against infectious disease it is likely that cytokines and other genes associated with inflammatory and immune response play a crucial role. Indeed, efficient activation of a cellular immune response is very important in triggering a protective response against *M. tuberculosis*, *M. leprae* or other pathogens (intracellular or not) [6].

Several review papers have been published describing design and interpretation of genetic association studies [7–10], but none of them dealt specifically with infectious diseases, where some peculiarities should also be considered. One example is that as with classi-

*Corresponding author. Tel.: +55 213 8361102; Fax: +55 21 227 05141; E-mail: apacheco@fiocruz.br

ISSN 0278-0240/09/$17.00

cal case-control studies, the main weakness of genetic case-control studies relies on the correct selection of controls, an issue that is seldom discussed in the literature, in which it is very common to have controls selected among blood donor banks or other healthy volunteers, which do not take into account the exposure to the studied infection.

Other, more general problems should also be understood and addressed, as proper sample size calculations to get enough power to detect a clinically or biologically significant difference between cases and controls, but at the same time being careful not to waste precious resources with overpowered studies, although it is far more common to detect underpowered studies in the literature – for an example in tuberculosis, see Pacheco and colleagues [11].

Here we review and evaluate some premises of genetic epidemiological studies based on case-control designs using infectious diseases, specifically with leprosy and tuberculosis (TB) as a model.

2. General concepts in genetics

2.1. Genetic markers

It is obvious that many genes and genetic polymorphisms are involved in controlling signaling pathways critical to host resistance, disease susceptibility and severity. To perform a case control study using a candidate gene approach it is necessary that a gene is chosen among those that have shown consistently a biological implication with outcome of disease. For example cytokines (IFNG, IL-12, etc.) are very good candidates because there are data in humans as well as mouse models suggesting their involvement with disease outcome. To map these chosen genes it is necessary to use genetic markers. There are several types of markers in human genome like short tandem repeats (STR), variable number of tandem repeats (VNTRs) or single nucleotide polymorphisms (SNPs). They are all important in genetic studies depending on the design chosen, but since STRs and VNTRs are multiallelic, more expensive and difficult to genotype, SNPs have been used so far.

2.2. Single nucleotide polymorphisms

Single nucleotide polymorphisms are mostly biallelic point mutations, present with a frequency higher than 1% in the population, and are observed with variable densities depending on the region of the human genome studied. Very dense regions can be observed with one SNP every 50–100bp while other places at the genome have one SNP every 500–1000bp. Therefore SNPs are widely distributed along human genome being easy (and cheap) to type, creating markers that can be used in association studies. SNPs are also believed to be the true source of variability among humans, especially when they are positioned in translated stretches of the DNA altering an amino acid. Nevertheless, most SNPs are located in non-coding regions, either intergenic or intragenic. A large number of SNPs in cytokine loci have been described and studied in complex diseases like infectious and autoimmune [3,5,12].

The total number of SNPs is currently estimated to be well over 10×10^6. Recently, with denser maps of human genome, a real polymorphism and not artifacts of sequencing have been ascertained in several populations although some SNP databanks are still carrying thousands of genotypic errors. Information concerning frequency in populations with different ethnic background generally African, Asian and Caucasian are available and SNPs exhibiting a higher frequency 15–25% are generally presented in all three populations [13]. This information is currently updated in on-line websites. A SNP database, also known as dbSNP (http://www.ncbi.nih.gov/SNP); a SNP consortium (http://snp.cshl.org/) and the haplotype map international project (http://www.hapmap.org/) can be used to browse and locate SNPs that exhibit characteristics to be genotyped.

2.3. Candidate genes

How can we choose the best gene to perform a case-control study with the disease we work with? There are roughly 30.000 genes in human genome and if we did not have any information about the disease, it could be assumed that any one of these 30.000 genes and more than 5 millions of SNPs needed to be tested.

Fortunately, we can use previous biological relevant information to select and test a specific gene. Thus, candidate gene is an approach to test polymorphisms in a gene that have been previously implicated biologically with the disease.

For example, there was well-documented evidence that tumor necrosis factor (TNF) was involved in leprosy susceptibility [14] and also in the inflammatory complications along the course of the disease [15,16]. Thus it was a sound possibility to test TNF in a case-control study in leprosy, which was done in the first

genetic epidemiological population-based study in leprosy that tested the −308 SNP in the promoter region of the TNF gene [17]. The association study of the region was later replicated in other studies, although they yield conflicting results [18–20] whose implications will be discussed later.

3. General concepts in epidemiology

Before delving into specific issues it is important to briefly summarize general concepts of epidemiology, so that many of the issues mentioned will be clearer to the reader. In no way is this section covering in deep the concepts presented and the reader is referred to very good epidemiology books [21,22].

3.1. Association

The main objective of epidemiological studies, including genetic case-control, is to infer causality between some exposure of interest (e.g. a genetic marker) and an outcome (e.g. presence or severity of disease). To do so, we compare two (or more) groups of people, depending on the design adopted. For population-based case-control studies, two groups of persons are compared: those who are identified with the disease (cases) and those randomly selected from the source population (i.e. the population where those cases came from) and who are disease-free. Then, for both groups, the exposure of interest (in our case, a genetic marker) along with relevant co-variables is determined, in a retrospective fashion, to guarantee that the exposure occurred *before* the outcome happened.

The next step is to compare both groups to determine if such exposure is associated with the disease. This is done using some association measure, which will be the odds ratio (OR) for the great majority of case-control designs. The OR is calculated as the ratio of the odds of being exposed given that the person has the disease and the odds of being exposed given that the person does not have the disease. The odds for a group is simply the proportion of persons exposed (p) over its complement (1-p). If the odds for both groups are the same, then OR $= 1$, and no association exists between the exposure and the disease; if it is between 0 and 1, the exposure confers protection and if it is greater than 1, the exposure is a risk factor for the disease. Note that the OR ranges from zero to infinity. Thus, association of the AA genotype shown in Table 1, with an OR $=$ 2.83 means that people carrying the AA genotype have 183% more chance to develop the disease than people from the control group.

3.2. Bias

There are several issues we have to be concerned about when making such inferences, and they can be addressed in the design phase, in the analysis phase or in both phases of the study. The main objective is to avoid spurious associations that can arise, especially in non-controlled (but not restricted to them) designs, such as case-control studies.

Bias is a main concern in case-control studies, since it cannot be addressed in the analysis and depends exclusively in the design of the study. Even though in genetic epidemiology classification bias due to incorrect ascertainment of the exposure status is generally not seen as a major issue [10], errors in genotyping both cases or controls should be carefully assessed, as we discuss in more detail below.

3.3. Random error

Random variability arises from the fact that we are dealing with samples from a population and associations found (i.e. OR $\neq 1$) may be due to chance alone. In order to avoid such erroneous conclusion, statistical tests are employed to guarantee that not only an OR $\neq$ 1, but also that it is statistically different from 1. When dealing with samples, there will never be 100% of certainty that the association is not due to chance alone, but one can guarantee that with a certain confidence (generally 95%), the association is true. Random errors should be dealt with in the design phase of the study with the calculation of an adequate sample size to be compared and also in the analysis, using adequate tools to analyze the data.

3.4. P-values

P-values are derived from the classical calculation of a critical region in the sample distribution of the association measure under the null hypothesis and is translated into the probability of having that calculated OR $\neq 1$, but still the truth being that OR $= 1$, *when the experiment is repeated infinite times*. This corresponds to the type I error in a hypothesis test, or the probability of rejecting the null hypothesis when it is true. It is important to notice two things about the p-value: a) a very small p-value does not guarantee that the result is false – only that it has a small probability of being wrong, but out of the infinite different samples that could have been chosen each time you repeat the experiment, you are analyzing one and only one of

Table 1
Example of genotype and allelic frequency comparisons using a SNP for any disease cases and controls

Genotypes	Controls (fraction)	Patients (fraction)	OR (95% CI)	P-value
GG	143 (0.36)	60 (0.27)	Reference	–
GA	197 (0.49)	97 (0.43)	1.17 (0.80,1.73)	< 0.001
AA	58 (0.15)	69 (0.31)	2.83 (1.79,4.50)	
Total	398	226	–	–
Allele G	483 (0.61)	217 (0.48)	Reference	–
Allele A	313 (0.39)	235 (0.52)	1.65 (1.30,2.08)	< 0.001
A carrier	255 (0.64)	166 (0.73)	1.55 (1.08,2.22)	0.02
Pass HWE test?	Yes ($p = 0.53$)	No ($p = 0.03$)	–	–

them; b) the p-value says nothing about the strength of association (which is given by the point estimate of the OR) and it does not give us a good idea about the dispersion of the point estimate.

3.5. Confidence intervals

Confidence intervals (CIs) can also be used to establish if the ORs are statistically different from 1, only it is based on the alternative hypothesis – *when the experiment is repeated infinite times* and for a 95% confidence level the CI is calculated for that measure, 95% of such CI will not contain the value in the null hypothesis (i.e. OR $= 1$). It means that if a 95% CI does not contain the unit, we can reject the null hypothesis. One advantage in using CIs is that it gives us an idea of the dispersion of the calculated statistic, which translates to the precision of the estimate. For example a CI ranging from 1.1 and 99 with a P-value of $P = 0.048$, although statistically significant show that association is very imprecise.

3.6. Confounding

Even though this is quite a generic idea, in genetic epidemiology it is best known and studied as population stratification, when a certain characteristic is associated differently among populations with distinct genetic backgrounds. Actually, any co-variable that fall into the following criteria can be considered as a confounding variable: a) it is associated with the disease; b) it is associated with the exposure in the source-population; and c) it is not part of the biological causal chain that leads from the exposure to the disease.

One important notion that has to be considered is that confounding is not necessarily something that has to be avoided, as is bias and lack of power (discussed below), but something that has to be correctly dealt with. If one wants to eliminate the confounding effect of some co-variable, it is possible to restrict the selection for some group in the design phase (e.g. if ethnicity is a confounder, select only persons from a certain ethnic background); if one is not interested at all on the possible effects and interactions the co-variable has with the outcome, it is possible to perform a matching in the design phase and then use appropriate methods for paired data in the analysis phase of the design – this is done for example in family-based case-control studies, or when cases and controls are matched in respect to ethnicity background; and finally one can simply account for confounding in the analysis phase, when interactions will be measured and a stratified analysis can be performed either with an weighted average OR for all substrata, or individual ORs for each studied strata, even though this approach will require a larger sample size to obtain the same power.

4. Design of population-based genetic case-control studies of infectious diseases

Now that we have briefly discussed the building blocks of genetics and epidemiology, in the following sections we will put it all together and discuss more specifically the design of population-based genetic case-control studies of infectious diseases. Figure 1 summarizes the main steps in designing such studies.

4.1. Selection of cases

The most challenging step in case-control studies design is definitely the correct choice of the cases and the controls. Even though cases generally present less challenge to be chosen, the “case” status should be carefully ascertained in order to avoid classification errors, i.e. the most reliable diagnostic method should be used to maximize the assurance that a person defined as a case is a real case and that those who are not cases should be discarded as such. This means that the diagnosis method for disease ascertainment should have both sensitivity and specificity as high as possible. For example, in the case of TB, it is advisable to use culture-

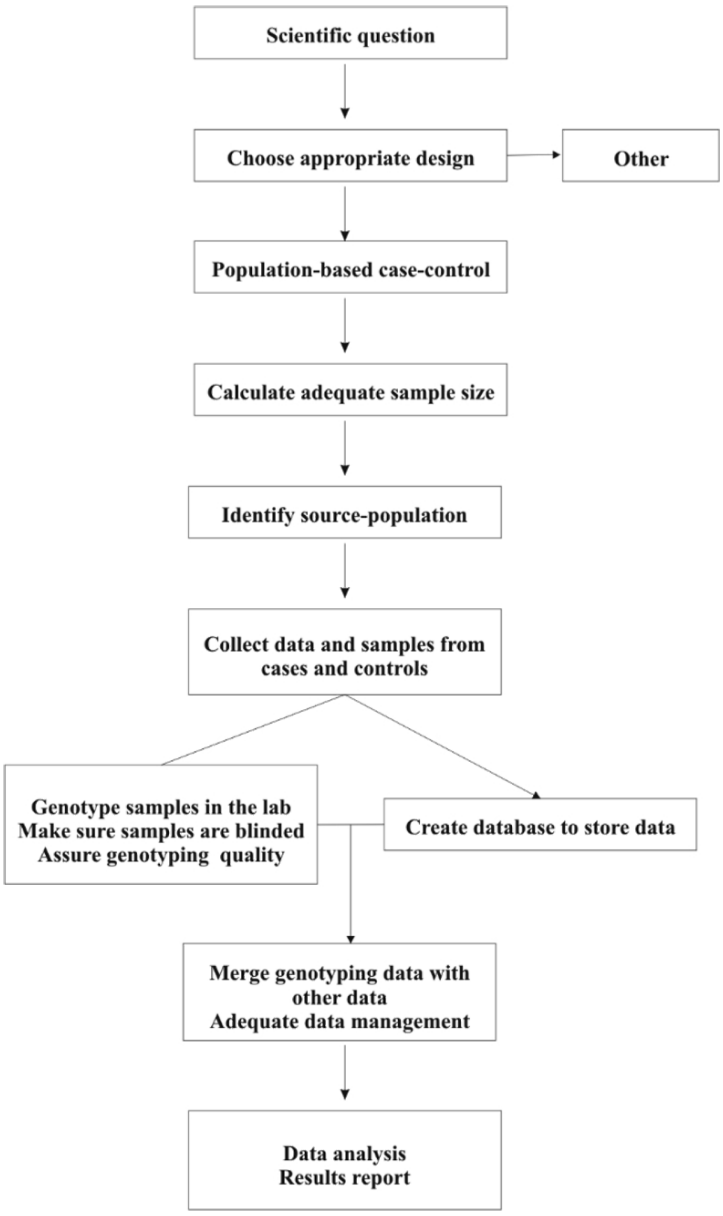

Fig. 1. Flowchart showing the main steps in designing, conducting analyzing and reporting results from a genetic case-control study.

proven cases instead of using smear microscopy tests. The reason here is that even though the latter has high sensitivity, lacks specificity since any acid-fast bacteria will be classified as TB. In leprosy it is advisable to use bacilloscopic and/or histologically classified cases.

4.2. Selection of controls

Controls are even trickier to choose, especially in the context of infectious diseases in general. As mentioned before, controls have to be chosen from the source-population where the cases came from. Even though it sounds like a straightforward task to choose controls from such a population, a closer look may reveal that the task is not that simple. For instance, it is not uncommon to find in the literature studies that compare TB or Leprosy cases with healthy blood donors. At first it may sound good, because it would be representative of the general population of a certain place, easy and cheap access to samples, etc. But notice a complication

for infectious diseases – if the question is if a certain genetic profile prevents one from being sick with TB or Leprosy, all the controls have to have been exposed (or challenged) at least once, and ideally enough time, to the infectious agent, otherwise it would be impossible to say if the controls did not get sick simply because they are less likely to be challenged with the infectious agent. The impact of such choice will be noted both on bias in the point estimate and especially on power to detect a true difference, and will be dependent on the prevalence of disease among those exposed and on the prevalence of exposure to the agent on the source population. This idea is similar to the more general problem of randomly selecting from a general population as described by Garner 2006 [23].

Even though it would be impossible to tell for sure if a person was exposed or not to a certain infectious agent, there are some ways to at least increase this probability, which would be to work with case contacts in the more general context of infectious diseases and in the specific case of TB or leprosy, household contacts. That does not guarantee that the exposure experience is the same in cases and controls, but it carries much less chance of bias and loss of power than the completely random approach.

Of course there are some other approaches that could be used to go around this problem depending on the setting the study is being conducted. For example, in the case of TB, positive Tuberculin Skin Test (TST) health professionals who work with TB patients could be used, but that does not take into account the quality and frequency of the challenge.

To illustrate this issue, we use some simulations to show how comparing controls that were not challenged may influence on both random error and bias the ORs in a hypothetical example with TB. Let us assume that we have designed a study to assess the effect of a certain SNP on TB latent infection and calculated a sample size to get adequate power (see next section) to measure a minimum difference of OR $= 2$ with baseline frequency of 10% of the allele of interest, assuming a co-dominant model and using a linear trend test. That would yield a sample size of about 300 individuals per group. Assuming that among TB household contacts the prevalence of TB latent infection is about 50% [24], Fig. 2 depicts the impact on power (A) and on the median ORs and 95%CIs after 1,000 simulations for varying fractions of non-exposure. In this context, non-exposure would represent the complement of infection prevalence in the population of healthy blood donors. As expected, this would work as a non-differential bias in respect to the exposure, because part of the controls would actually become cases if exposed to the agent, and thus have a genotype distribution similar to cases and not controls from the source population. The effect is decrease in power, coupled with bias on the ORs towards the null hypothesis (black lines). On the other hand, if we are dealing with active TB disease, the figures will change, because the expected prevalence of disease among those exposed is much lower – about 4.5% [24] and the impact on power and bias can be considered negligible (red lines).

4.3. Sample size considerations

A key feature that should be carefully handled is the calculation of a suitable sample size. On one hand, the sample size should not be so small that it will not allow tests with enough power to detect a clinically significant difference, but on the other hand, given scarce resources, it should not be bigger than necessary to provide adequate power to test the phenomenon being studied, to avoid unnecessary waste of time and money. It is not our intention to provide formulas for sample size calculations, for which the reader is referred to Biostatistics books that deal with this subject [25]. What we will discuss are some general issues that should be taken into account when calculating sample sizes.

Sample size depends on several factors: desired power, probability of type I error, the magnitude of the difference to be tested and the variance of the variable studied in the source-population. Power refers to the ability of a statistical test to find an association (i.e. reject the null hypothesis, in our case, reject that OR $=$ 1) when the association exists (i.e. when the alternative hypothesis is true, OR $\neq 1$) this is actually the complement of the so-called probability of the type II error, or β error, so it is also referred to as $1 - \beta$. In general practice, it is desirable that a test has at least 80% power to detect the difference one wants to test. The probability of the type I error, also known as α is taken to be 0.05 and the magnitude of the difference to be tested in our case can be translated to the minimum OR one expects to consider clinically (or biologically) relevant and the variance in our case depends basically on the baseline proportion of the reference allele among controls (p_0) and the OR itself (which actually can be translated into the proportion among cases – p_1).

One important issue to be carefully considered is the choice of the minimum OR that represents a clinically or biologically relevant difference between cases and controls. In theory it is possible to get a sample that is

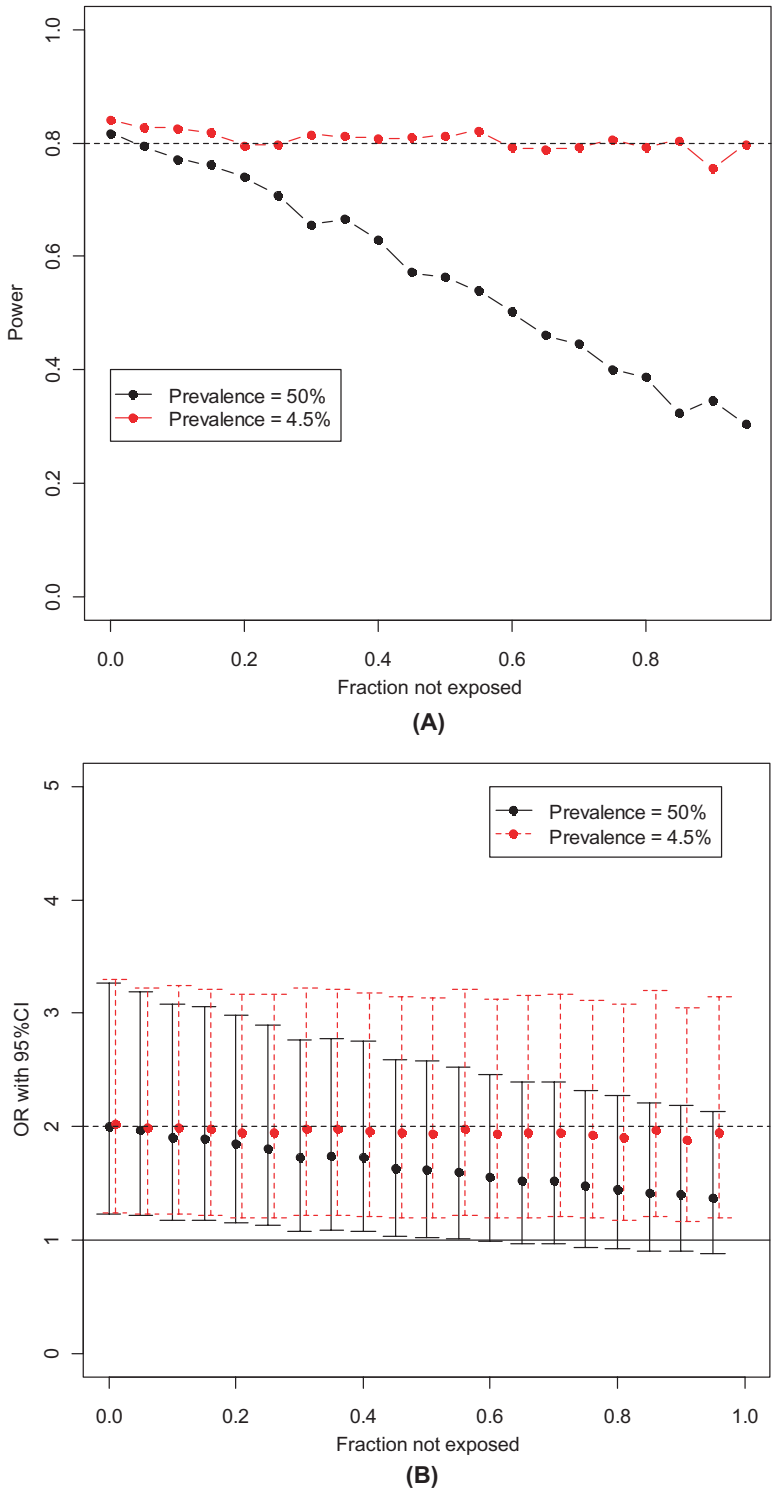

Fig. 2. Effect of non-exposure to infectious agent with high prevalence (latent TB infection, black lines) and low prevalence (active TB disease, red lines) of the infection among those exposed for the case of an odds ratio (OR) of 2 and allele frequency of 0.1 among controls. (A) Effect on power. (B) Effect on ORs and 95% confidence intervals. Filled circles: 50% prevalence; open circles: 4.5% prevalence.

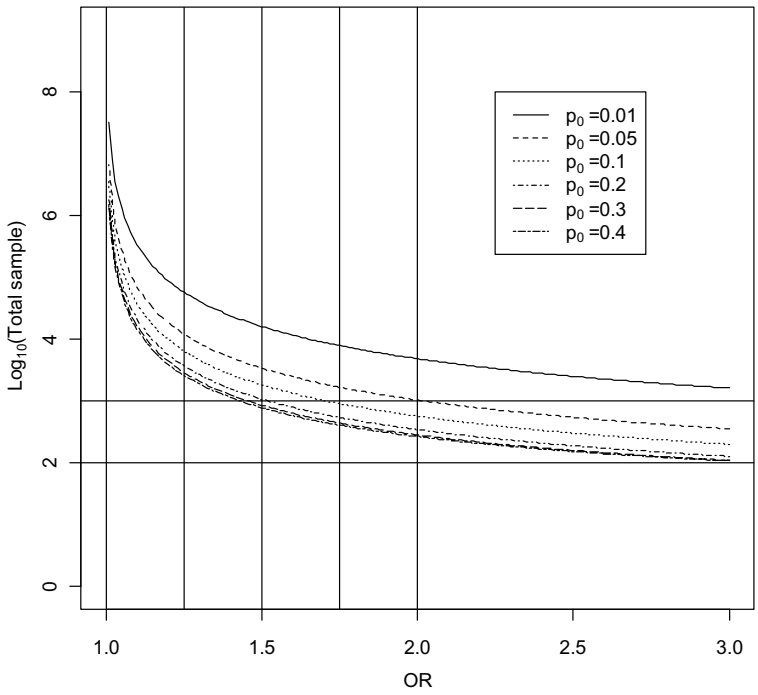

Fig. 3. Necessary sample sizes in $\log_{10}$ to achieve 80% power on a simple binomial comparison for varying odds ratios (x axis) and allele frequency among the control group (p_0).

large enough to detect very small differences (e.g. a 1% increase in risk, which corresponds to an OR of 1.01), but in practice it would quickly become impossible to get such a sample. Figure 3 illustrates the behavior of sample sizes needed to detect various ORs (x axis) for 6 different baseline proportions in a simple binomial comparison (e.g. comparing a genotype of interest in a dominant model), in a balanced design (equal number of cases and controls). Note that the y axis is in $\log_{10}$ scale, so in the scenario of $p_0 = 0.01$ and to detect a 1% risk, we would need a sample of over 32 million people (!). For reasonable differences like 50% (OR $= 1.5$) or 100% (OR $= 2$) increases, sample sizes of 1,000 people would be enough, but not for very rare alleles. In this regard, whenever choosing for the SNPs to study in a candidate(s) gene(s) it is likely to choose for SNPs with minor allele frequency higher than 10%.

Another issue that has to be taken into account is multiple comparisons. The type I error will be fixed in 5% for a *single comparison*. This means that if the study involves more than one SNP to be studied in the same sample, this fact has to be taken into account, as the effect of multiple comparisons is the inflation of this error. For example, if 20 SNPs are being compared, then for a 5% type I error, it is expected that at least one of them (5% of 20) will be associated with the disease due to chance alone. Several methods have been proposed to account for that problem, including Bonferroni correction and false discovery rates [26]. These approaches should be taken into account when calculating adequate sample sizes, which will generally increase the number of people needed to achieve the same type I error.

Even though sample sizes for simple models are fairly easy to calculate through algebraic results, and are implemented in common statistical software, as would be the case of simple proportion comparisons, as we showed, sometimes algebraic results do not exist or are just intractable mathematically. For those cases, simulations would help find a suitable sample size, which is done by the construction of power curves and evaluation of different scenarios, given some guesstimation for prior parameters. To illustrate how simulations work, let us stick to our simple example and compare power curves for varying numbers of total sample sizes to compare proportions in two groups with the allele frequency among controls of 10% and an expected OR of 2. Figure 4 shows the comparison of an algebraic calculation (black lines) and 1,000 simulations for each sample size. Note that they are not exactly the same, but fairly similar. In this simple example, a simulation algorithm with 4 lines of code would be enough to

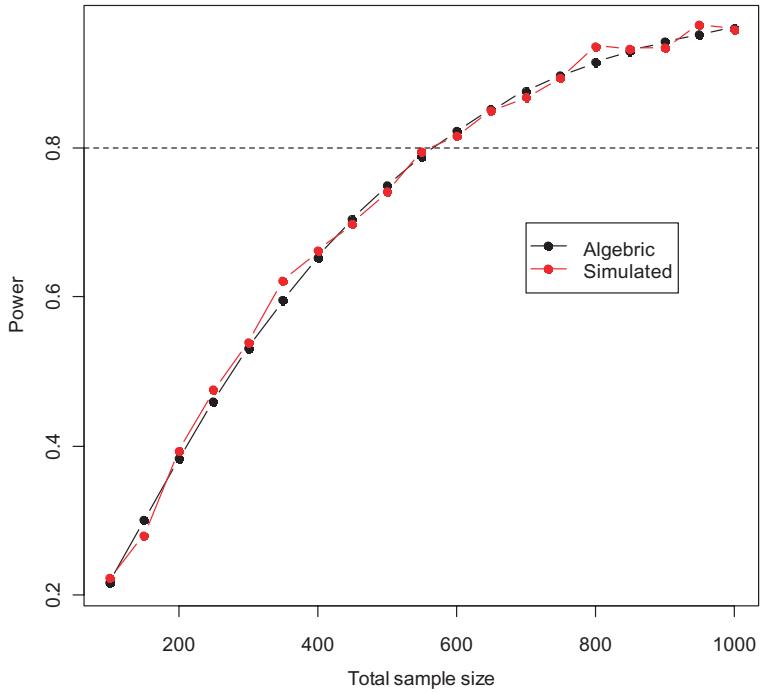

Fig. 4. Comparison of power curves calculated from algebraic (black lines) and 1,000 simulated samples (red lines) for the simple case of a binomial comparison for an odds ratio of 2 and prevalence of the allele among controls of 10%.

do the job, but for more complex examples, it would demand some effort to correctly simulate the desired scenarios.

5. Genotyping SNPs and quality assurance

In order to genotype SNPs it is used a method to specify the region that is generally a PCR and then another conjugated method to sort out the variation presented in a specific position. To do so, a digestion with a restriction enzyme, hybridization or even biochemical and molecular methods to separate or weight the DNA strand carrying the SNP (measuring by high pressure liquid chromatography (HPLC), a mass spectrometer or a sequencer). In the most of the papers published that analyses a small number of SNPs, techniques based on PCR with post analysis using RFLPs (restriction fragment length polymorphism), SSP (single strand polymorphism) or sequencing are the most common, although these tests are becoming obsolete and replaced by methods based on PCR coupled with fluorescent detection (real time PCR). Old-fashioned methods like PCR-RFLP and PCR-SSP have a very low cost-benefit ratio since it is laborious, expensive, and error-prone. On the other hand, these methods are still easy to perform and to fit in a low tech lab that cannot afford the cost of sequencing machines or real-time PCRs or the new generation of genotyping platforms. A summary of common methods for genotyping is depicted in Table 2. Based on this table, one would assume that performing of genome-wide association studies (GWAS) using platforms genotyping with millions of SNPs would be cheaper. It is actually true, but the cost of the assay per sample that enrolls millions of SNPs/assay is far more expensive than a simple PCR coupled with sequencing or even a real time PCR. It is also needed a nice infra-structure to set up a genotyping platform, which also takes into account to choose the strategy to perform a case-control study for any disease.

As we mentioned before, genetic case-control studies suffer less with classification biases than classic epidemiologic studies. One major source of such bias in genetics emerges from genotyping errors in the laboratory [27]. Of course, the main measure to avoid this kind of bias is to use sound protocols for genotyping and avoiding errors during the experiment. If possible it is necessary to assure the quality of your DNA samples. The DNA obtained for genotyping must exhibit high quality that can be measured spectrophotometrically in separate samples or in plates or even by robust

Table 2
Common genotyping methods and platforms

Type	Method	Routine (handling)	scale SNPs/day	Cost in US$	Application (Companies)
Low scale	PCR-RFLP and PCR-SSP	1 SNP/sample (manual)	300–500	$0.50–2.00/SNP	Case-controls (in house)
Low scale	Allelic discrimination (real time PCR)	1 SNP/sample (semi-automatic)	500–1000	$2.00/SNP	Case-controls (Applied Biosystems)
Moderate scale	Sequencing	1–10 SNPs/sample (semi-automatic)	500–1000	$1–2.00/SNP	Case-controls (Applied Biosystems, BioRad)
Moderate scale	Mass spectrometry	1–50 SNPs sample (semi-automatic)	1000	$0.10–1/SNP	Case-controls (Sequenom)
Large scale	Bead arrays	100–10^5 (semi-automatic)	Hundreds of thousands	$0.01-/SNP	GWAS (Illumina)
Large and ultra-large scale	Microarray geneChips	10^6 (semi-automatic)	Millions	$0.001/SNP	GWAS (Affymetrix)

real-time/conventional PCR reactions. Nevertheless, genotyping errors are common and it is likely to use control samples (for example DNA of a known genotype test by high quality sequencing) to validate your method. Most common errors are observed because of the presence of unknown SNP in the neighboring nucleotides of a target SNP can affect primer or probe complementary hybridization leading to another calling to that sample. Also problems with annealing in methods that rely on hybridization can occur and efficiency on several commercial fluorescent probes varies significantly. Each method chosen can undergo miscalling of the SNP, i.e. and error in the genotyping (such as GG is called as GA in TNF -308 position) that obviously create a bias for analysis. For example PCR-RFLPs that are not carefully set up can generate an inflated number of heterozygotes when a partial digestion with the restriction enzyme is observed. Several scenarios can be drawn based on miscalling of a genotype.

One important issue that has to be accounted for in the laboratory is that the person responsible for calling the true genotype for patients and controls samples should be blinded in respect to their status, in order to avoid classification bias.

6. Data abstraction and management

This topic is mentioned here just to emphasize that one important source or systematic error is poor care in properly abstraction and storage of data. In general genetic data is generated in the lab while epidemiological data is generated somewhere else (e.g. in the clinic the patient is attending to) and at some point these data have to be merged into a single database to allow proper analysis of data. The details of how this should be done is out of the scope of this paper, but one word of caution is to always use appropriate software to handle databases and to avoid as much as possible word processors and electronic sheets to store data.

7. Data analysis and presentation

It is also important to take into account is how to report the findings of a genetic association study to allow reader to assess for themselves the merits and limitations of the study and evaluate if the data can be used for other studies (e.g. meta-analyses). Recently, the STREGA (Strengthening the Reporting of Genetic Association Studies) statement [28] has been published with recommendations on how data should be reported in scientific papers. This statement is an extension of a more general one, STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) [29, 30]. We strongly recommend that readers refer to those documents and adhere to their recommendations when writing their manuscripts.

Here we highlighted some of the items also mentioned in STREGA. One important step is to choose appropriate statistical software that is able to perform all the necessary tests. Even though there are several mainstream packages that can do the job, one very good choice is the R environment [31] that has increasingly been used for the analysis of genetic data.

7.1. Description

The first step is to describe the data at hand. For genetic data, genotype and allele frequencies (Table 1) along with a description of Hardy-Weinberg equilibrium status among cases and controls should always be provided (see below), along with appropriate statistical tests and if more than one locus is studied, description of linkage disequilibrium among them is also warranted. Other co-variables (if present) should also be described in respect to the disease status.

7.2. HWE tests for genotyping error ascertainment

Since genotyping errors can pose serious problems when data is analyzed, especially if there is reason to believe that the errors are differential among cases and controls, it is customary to use tests for departure from Hardy-Weinberg equilibrium (HWE) to assure genotyping quality. Even though the usefulness of such procedure is an area of current intense debate [32–36], it is our impression that one should test and report results about HWE departures, especially among controls, as the general population is assumed to be randomly distributed. The main point is that HWE tests should be applied and their results reported clearly, as recommended by STREGA [28].

It is very important to know if the population tested exhibits some sort of cryptic stratification due to uneven admixture, migration or other genetic effects. Sometimes, this problem has great effect over results since cases and/or controls can be recruited from different "structured subpopulations" generating SNP frequencies that will be significantly different irrespective of the loci that are depicted to be tested. One way to prevent this issue is the use of genomic controls [37]. Briefly, several randomly selected SNPs are used to normalize the differences between cases and controls. Other methods try to use genetic ancestry based on "control databases" derived from genome-wide studies in different populations to match samples from cases and controls [38,39]. In general, important departures from HWE can be observed when different subpopulations are clustered in the same group and it is highly recommended that populations presenting several SNPs deviating from HWE be carefully reassigned or some genetic ancestry screening in the group be performed. Obviously, the best way to avoid these problems is to carefully select controls and patients.

7.3. Simple comparisons

Traditionally descriptions can be accompanied by simple comparisons of frequencies, using chi-square tests or Fisher's exact tests where appropriate, but recently simple comparisons using univariable logistic models have been increasingly used, so that a table could have a 'raw' OR and an 'adjusted' OR for the multivariable model.

7.4. Logistic regression

Logistic regression [40] is the model of choice to be used when working with ORs, because the calculated coefficients can easily be converted to ORs by simply taking their exponential. It belongs to a general class of generalized linear models (GLM) [41] with a logit link function, with residuals being expected to behave as a binomial distribution. The results of a logistic model are depicted in Table 1 along with genotype and allele descriptions.

In this case, we present the models in three ways: (i) making no assumption about the inheritance mode – here, a nominal genotype variable entered in the model, in the form of two dummy variables, one for heterozygous group (GA) and another one for the homozygous group of interest (AA). Even though it seems like the GG homozygous group is not in the model, it is actually represented by entering zero in both dummies. Even though both ORs and 95%CIs are shown, only a single p-value is reported, because they are not truly two variables, but rather one variable with 3 possible values represented as two variables; (ii) assuming codominance and that the allelic dose has a linear trend – this is also known as the Cochran-Armitage test for trend in respect to allele A. The single OR reported for this allele assumes that the OR of having 2 A alleles in respect to those that have 1 A allele is the same as the OR of those that have 1 A allele in respect to those who have no A allele; (iii) assuming dominant inheritance, where the presence of the allele A determines the disease outcome – this approach does not seem to be supported by the inheritance-free model and indeed underestimates the effect of the A allele.

Multivariable logistic models are very popular in epidemiology and are handy to include co-variables one would like to control for in the analysis phase, in case confounding is suspected. Co-variables can be nominal, ordinal or continuous and the models allow for very good interpretability of their coefficients and ORs.

Finally, the ease of extensibility of GLMs can become handy when one wants to explore more complex situations, as we will see in the next section.

8. Haplotypes

Haplotypes are a combination of 2 or more polymorphisms (SNPs in this case) within a single chromosome in an individual. It has become a very popular approach to study the association of haplotypes with diseases when two or more SNPs are studied. The advantages of using haplotypes instead of individual SNPs are due to: (i) less information contained in a SNP alone than in a combination of SNPs that are in LD; (ii) if we are to combine the information of several SNPs separately, correction for multiple comparisons (e.g. Bonferroni, FDR) would have to be employed, lowering considerably the power to detect associations.

Whenever it is impossible to genotype the cases parents, the determination of the haplotype phase is not always possible. If more than one locus is heterozygous and it is not possible to genotype the individual's parents, there is no way of knowing for sure in which chromosome the combination of alleles lie (even though there are some lab techniques that would allow that directly). One way to go around this problem is to use statistical imputation of missing data, which will assign a probability of an individual having certain haplotypes (depending on the combinations involved, an individual can have several different pairs of them), given the individual's haplotype and the known phase haplotype distribution in the population studied.

Recently, a method that combines Generalized Linear Models (GLM) with the Expectation-Maximization (EM) algorithm have been proposed [42], in which a score statistic is used to infer the association of haplotypes of unknown phase and diseases, and with the possibility to include covariates in the model. This approach combines the imputation of missing data (EM algorithm) with logistic regression models, when the outcome is binary (the method is also extended to other outcome types, through GLM). Even though the method is able to quickly calculate the statistic to infer associations, it does not calculate the maximum likelihood estimate for the model parameters, which does not allow the calculation of the strength of association involved.

In a more recent paper the same group proposed and implemented new methods to calculate the MLE for the GLM [43], incorporating the uncertainty of the EM phase of the algorithm, allowing the calculation of an Odds Ratio OR) when the outcome is binary.

9. Meta-analysis

Meta-analysis is a powerful tool when correctly used and can provide a consensus answer using available data from different sources that study the same phenomenon, treating each study as a cluster. Of course, the results will be more accurate as more data is available and this underscores the need for publication of even non-significant results of well-designed studies to avoid publication bias in meta-analysis.

The details of performing meta-analysis are out of the scope of this paper, but some insight on its interpretation along with some issues is warranted. The main feature is to identify an association of interest and perform a thorough literature search on the subject. The more spread and unrestricted the search, the better the chance of not having publication bias, in which papers that describe statistically significant results are more likely to be accepted for publication than studies that do not find significant associations. This would happen even if published studies are not so well-designed, but suffer from the so-called "winner's curse" [44,45].

Of course, one would like to restrict some of the information, especially based on quality of data collection, conformation to study guidelines and of course, enough data available to be pooled in the publication (even though in some cases, data can be requested directly from the authors for that end). The quality of genotyping is again warranted in meta-analysis, and even though there is evidence some evidence that the use of data with departures from HWE does not influence the results [34] we still think it is a good policy to exclude studies that report loci not in HWE [11].

The interpretation of the results is the same as described for the logistic regression and ORs above, only now there will be a summary OR that in general is visualized together with the individual ORs for each study included in the meta-analysis (generally in a figure known as forest graph). Currently it is usual to report both the fixed effects summary OR and the random effects summary OR, emphasizing the former (which is more powerful) if the random variation across studies is not significant and the latter otherwise.

References

[1] B. Beiguelman, Some remarks on the genetics of leprosy resistance, *Acta Geneticae Medicae Et Gemellologiae* **17** (1968), 584–594.

[2] A.V. Hill, The immunogenetics of human infectious diseases, *Annual Review of Immunology* **16** (1998), 593–617.

[3] A. Alter, A. Alcaïs, L. Abel and E. Schurr, Leprosy as a genetic model for susceptibility to common infectious diseases, *Human Genetics* **123** (2008), 227–235.
[4] S. Marquet and E. Schurr, Genetics of susceptibility to infectious diseases: tuberculosis and leprosy as examples, *Drug Metabolism and Disposition: The Biological Fate of Chemicals* **29** (2001), 479–483.
[5] M.O. Moraes, C.C. Cardoso, P.R. Vanderborght and A.G. Pacheco, Genetics of host response in leprosy, *Leprosy Review* **77** (2006), 189–202.
[6] M. Moraes, J. McNicholl, T. Huizinga and T. Ottenhoff, Cytokine Genes I: IL10, IL6, IL4, and the IL1 Family, in: *Genetic Susceptibility to Infectious Diseases*, R.A. Kaslow, J. McNicholl and A.V.S. Hill, eds, Oxford University Press, USA 2008, pp. 208–226.
[7] J. Attia, J.P.A. Ioannidis, A. Thakkinstian, M. McEvoy, R.J. Scott, C. Minelli, J. Thompson, C. Infante-Rivard and G. Guyatt, How to use an article about genetic association: B: Are the results of the study valid? *JAMA: The Journal of the American Medical Association* **301** (2009), 191–197.
[8] J. Attia, J.P.A. Ioannidis, A. Thakkinstian, M. McEvoy, R.J. Scott, C. Minelli, J. Thompson, C. Infante-Rivard and G. Guyatt, How to use an article about genetic association: A: Background concepts, *JAMA: The Journal of the American Medical Association* **301** (2009), 74–81.
[9] J. Attia, J.P.A. Ioannidis, A. Thakkinstian, M. McEvoy, R.J. Scott, C. Minelli, J. Thompson, C. Infante-Rivard and G. Guyatt, How to use an article about genetic association: C: What are the results and will they help me in caring for my patients? *JAMA: The Journal of the American Medical Association* **301** (2009), 304–308.
[10] H.J. Cordell and D.G. Clayton, Genetic association studies, *Lancet* **366** (2005), 1121–1131.
[11] A.G. Pacheco, C.C. Cardoso and M.O. Moraes, IFNG +874T/A, IL10 −1082G/A and TNF -308G/A polymorphisms in association with tuberculosis susceptibility: a meta-analysis study, *Human Genetics* **123** (2008), 477–484.
[12] J.P. Bayley, T.H.M. Ottenhoff and C.L. Verweij, Is there a future for TNF promoter polymorphisms? *Genes and Immunity* **5** (2004), 315–329.
[13] D.A. Hinds, L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer and D.R. Cox, Whole-genome patterns of common DNA variation in three human populations, *Science* (*New York, NY*) **307** (2005), 1072–1079.
[14] E.N. Sarno, G.E. Grau, L.M. Vieira and J.A. Nery, Serum levels of tumour necrosis factor-alpha and interleukin-1 beta during leprosy reactional states, *Clinical and Experimental Immunology* **84** (1991), 103–108.
[15] E.P. Sampaio, M.O. Moraes, J.A. Nery, A.R. Santos, H.C. Matos and E.N. Sarno, Pentoxifylline decreases in vivo and in vitro tumour necrosis factor-alpha (TNF-alpha) production in lepromatous leprosy patients with erythema nodosum leprosum (ENL), *Clinical and Experimental Immunology* **111** (1998), 300–308.
[16] M.O. Moraes, E.N. Sarno, R.M. Teles, A.S. Almeida, B.C. Saraiva, J.A. Nery and E.P. Sampaio, Anti-inflammatory drugs block cytokine mRNA accumulation in the skin and improve the clinical condition of reactional leprosy patients, *The Journal of Investigative Dermatology* **115** (2000), 935–941.
[17] S. Roy, W. McGuire, C.G. Mascie-Taylor, B. Saha, S.K. Hazra, A.V. Hill and D. Kwiatkowski, Tumor necrosis factor promoter polymorphism and susceptibility to lepromatous leprosy, *The Journal of Infectious Diseases* **176** (1997), 530–532.
[18] J. Fitness, S. Floyd, D.K. Warndorff, L. Sichali, L. Mwaungulu, A.C. Crampin, P.E.M. Fine and A.V.S. Hill, Large-scale candidate gene study of leprosy susceptibility in the Karonga district of northern Malawi, *The American Journal of Tropical Medicine and Hygiene* **71** (2004), 330–340.
[19] D.S.A. Franceschi, P.S. Mazini, C.C.C. Rudnick, A.M. Sell, L.T. Tsuneto, M.L. Ribas, P.R. Peixoto and J.E.L. Visentainer, Influence of TNF and IL10 gene polymorphisms in the immunopathogenesis of leprosy in the south of Brazil, *International Journal of Infectious Diseases: IJID: Official Publication of the International Society for Infectious Diseases* (2008), in press.
[20] A.R. Santos, P.N. Suffys, P.R. Vanderborght, M.O. Moraes, L.M.M. Vieira, P.H. Cabello, A.M. Bakker, H.J. Matos, T.W.J. Huizinga, T.H.M. Ottenhoff, E.P. Sampaio and E.N. Sarno, Role of tumor necrosis factor-alpha and interleukin-10 promoter gene polymorphisms in leprosy, *The Journal of Infectious Diseases* **186** (2002), 1687–1691.
[21] K.J. Rothman and S. Greenland, *Modern epidemiology*, Lippincott-Raven, Philadelphia, PA, 1998.
[22] M. Szklo and F.J. Nieto, *Epidemiology: Beyond the Basics*, Jones and Bartlett Publishers, 2006.
[23] C. Garner, The use of random controls in genetic association studies, *Human Heredity* **61** (2006), 22–26.
[24] J. Morrison, M. Pai and P.C. Hopewell, Tuberculosis and latent tuberculosis infection in close contacts of people with pulmonary tuberculosis in low-income and middle-income countries: a systematic review and meta-analysis, *The Lancet Infectious Diseases* **8** (2008), 359–368.
[25] B. Rosner, *Fundamentals of Biostatistics*, Duxbury Press, 2005.
[26] J.P. Shaffer, Multiple Hypothesis Testing, *Annual Review of Psychology* **46** (1995), 561–584.
[27] F. Pompanon, A. Bonin, E. Bellemain and P. Taberlet, Genotyping errors: causes, consequences and solutions, *Nature Reviews. Genetics* **6** (2005), 847–859.
[28] J. Little, J.P.T. Higgins, J.P.A. Ioannidis, D. Moher, F. Gagnon, E. von Elm, M.J. Khoury, B. Cohen, G. Davey-Smith, J. Grimshaw, P. Scheet, M. Gwinn, R.E. Williamson, G.Y. Zou, K. Hutchings, C.Y. Johnson, V. Tait, M. Wiens, J. Golding, C. van Duijn, J. McLaughlin, A. Paterson, G. Wells, I. Fortier, M. Freedman, M. Zecevic, R. King, C. Infante-Rivard, A. Stewart and N. Birkett, STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement, *PLoS Medicine* **6** (2009), e22.
[29] J.P. Vandenbroucke, E. von Elm, D.G. Altman, P.C. Gøtzsche, C.D. Mulrow, S.J. Pocock, C. Poole, J.J. Schlesselman and M. Egger, Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration, *PLoS Medicine* **4** (2007), e297.
[30] E. von Elm, D.G. Altman, M. Egger, S.J. Pocock, P.C. Gøtzsche and J.P. Vandenbroucke, The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies, *PLoS Medicine* **4** (2007), e296.
[31] R. Development Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2008.
[32] D.G. Cox and P. Kraft, Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error, *Human Heredity* **61** (2006), 10–14.
[33] S.M. Leal, Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium, *Genetic Epidemiology* **29** (2005), 204–214.

[34] C. Minelli, J.R. Thompson, K.R. Abrams, A. Thakkinstian and J. Attia, How should we use information about HWE in the meta-analyses of genetic association studies? *International Journal of Epidemiology* **37** (2008), 136–146.
[35] T.A. Trikalinos, G. Salanti, M.J. Khoury and J.P.A. Ioannidis, Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations, *American Journal of Epidemiology* **163** (2006), 300–309.
[36] G.Y. Zou and A. Donner, The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note, *Annals of Human Genetics* **70** (2006), 923–933.
[37] B. Devlin and K. Roeder, Genomic control for association studies, *Biometrics* **55** (1999), 997–1004.
[38] D. Luca, S. Ringquist, L. Klei, A.B. Lee, C. Gieger, H.E. Wichmann, S. Schreiber, M. Krawczak, Y. Lu, A. Styche, B. Devlin, K. Roeder and M. Trucco, On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants, *American Journal of Human Genetics* **82** (2008), 453–463.
[39] K. Roeder and D. Luca, Searching for disease susceptibility variants in structured populations, *Genomics* **93** (2009), 1–4.
[40] D.W. Hosmer and S. Lemeshow, *Applied logistic regression*, Wiley-Interscience Publication, 2000.
[41] P. McCullagh and J.A. Nelder, *Generalized Linear Models, Second Edition*, Chapman & Hall/CRC, 1989.
[42] D.J. Schaid, C.M. Rowland, D.E. Tines, R.M. Jacobson and G.A. Poland, Score tests for association between traits and haplotypes when linkage phase is ambiguous, *Am J Hum Genet* **70** (2002), 425–434.
[43] S.L. Lake, H. Lyon, K. Tantisira, E.K. Silverman, S.T. Weiss, N.M. Laird and D.J. Schaid, Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous, *Hum Hered* **55** (2003), 56–65.
[44] J.P.A. Ioannidis, Why most discovered true associations are inflated, *Epidemiology* (*Cambridge, Mass.*) **19** (2008), 640–648.
[45] N.S. Young, J.P.A. Ioannidis and O. Al-Ubaydli, Why current publication practices may distort science, *PLoS Medicine* **5** (2008), e201.