Analysis of population–based genetic association studies applied to cancer susceptibility and prognosis

Xavier Solé, Juan Ramón González, Víctor Moreno

Abstract Along hundreds of thousands of years, genetic variation has been the keystone for human evolution and adaptation to the surrounding environment. Although this fact has supposed a great progress for the species, mutations in our DNA sequence may also lead to an increased risk of developing some diseases with an underlying genetic basis, such as cancer. Among different genetic epidemiology branches, population–based association studies are one of the tools that can help us decipher which of these mutations are involved in the appearance or progression of the disease.

This chapter aims to be a didactic but thorough review for those who are interested in genetic association studies and its analytical methodology. It will mainly focus on SNP–array analysis techniques, covering issues such as quality control, assessment of association with disease, gene–gene and gene–environment interactions, haplotype analysis and genome–wide association studies. In the last part some of the existing bioinformatics tools that perform the exposed analyses will be reviewed.

Xavier Solé
Biostatistics and Bioinformatics Unit, Catalan Institute of Oncology – IDIBELL, Av. Gran Via s/n Km 2.7, 08907 L'Hospitalet de Llobregat (Barcelona), e-mail: x.sole@iconcologia.net

Juan Ramón González
Center for Research in Environmental Epidemiology, Doctor Aiguader 88, 08003 Barcelona, e-mail: jrgonzalez@creal.cat

Víctor Moreno
Biostatistics and Bioinformatics Unit, Catalan Institute of Oncology – IDIBELL, Av. Gran Via s/n Km 2.7, 08907 L'Hospitalet de Llobregat (Barcelona) and Departament de Ciències Clíniques, Facultat de Medicina, Universitat de Barcelona, e-mail: v.moreno@iconcologia.net

1 Genetic variation and its implication in cancer

The implication of genes in cancer has long been suspected because this disease shows familial aggregation, in some instances remarkably. The study of cancer cells shows extensive genomic alterations, ranging from mutations in target genes–known as oncogenes and tumor suppressor genes–to large chromosomal aberrations. These alterations are supposed to be triggered by initial events that accumulate and confer the cancer cells proliferation advantage and escape to control of DNA damage. Alterations are acquired during the carcinogenesis process and are called somatic alterations. However, individuals that carry alterations in germ line are known to have susceptibility to develop cancer. Mutations in a few genes have already been identified as responsible for cancer syndromes like Li–Fraumeni (p53), familial breast cancer (BRCA1, BRCA2), adenomatous polyposis coli (APC) and Lynch syndrome (MLH1, MSH2, MSH6, PMS2) [32]. These mutations show high penetrance, but are rare and do not explain more than 5% of all cancers, while other 15–25% are thought to have a relevant genetic contribution.

Though families share genes and environment, and part of the familial aggregation could be related to shared lifestyles, diet and other exposures, twin studies allow estimation of the relative contribution of genes and environment. An important fraction of most frequent cancers is related to genetic factors: 42% of prostate, 35% of colorectal and 27% of breast, and similar estimates were observed for other less frequent tumors [51].

The discrepancy between heritability estimates and the proportion of cases associated to known genes raised the hypothesis that other genes should be involved in cancer aetiology, though with lower penetrance and probably high frequency. An intensive search for these susceptibility genes has been triggered when genotyping technologies have emerged that allow easy simultaneous analysis of thousands to millions of genetic markers. Also, the knowledge that recombination is not occurring at random throughout the chromosomes, but in specific regions that delimit blocks of nucleotides that are transmitted together [1] (see section 8.3.2), has helped designing strategies to extensively explore the genetic variation at a genome–wide scale in order to identify cancer susceptibility loci.

Heritable genomic variations are called polymorphisms, which occur by mutation of DNA in germinal cells and are transmitted to descendants. Most of these polymorphisms have no functional impact, either because they occur in non–coding regions or do not modify the protein product qualitative or quantitatively. Some of these polymorphisms do have a functional impact and are the base of evolution. Usually when the effect provides an advantage to the individual the polymorphism increases in frequency in the population. Conversely, deleterious mutations tend to disappear, though they can reach relatively high frequency in the population if the heterozygous status provides some advantage like sickle cell anaemia carriers, which are more resistant to malaria.

There are three types of polymorphisms at the genetic level: Single Nucleotide Polymorphism (SNP), Variable Number Tandem Repeats (VNTR) and Copy Number Variations (CNV).

SNPs, the most frequent polymorphisms, are changes in one nucleotide at a given genomic position. Usually one nucleotide is substituted by other, but sometimes one or a few nucleotides are deleted or inserted (Ins/Del). The results of these minor changes are diverse. If the SNP is in an exon, it may confer a change in the aminoacid chain of the resulting protein, or a truncated protein if the SNP results in an stop codon. SNPs in introns and non–coding regions may also be functional by altering splicing sites or the binding of transcription factors. The ENCODE project [11] is revealing that DNA expression is frequent in non–coding regions. SNPs in resulting RNAs might also have relevant functions.

Non–functional SNPs are scattered throughout the genome with one average distance of one SNP every 1,000 bp. The average haplotype block has a size of 20,000 bp in non–African populations and 10,000 bp in African populations. Thus, there are about 20 SNPs per haplotype block on average, but only 5–6 different haplotypes per block, because there is high redundancy. Only few SNPs per haplotype block are needed to ascertain most of the variation and identify which haplotype is carrying a causal polymorphism, if it exists. These minimum number of selected SNPs are called haplotype–tagging SNPs (htSNPs).

Variable Number Tandem repeats appear with less frequency and consist in serial repetitions of a short series of nucleotides with length variability among individuals. For example, ATATAT = $(AT)_3$, ATATATATAT = $(AT)_5$. The repeats may be mononucleotide (AAAA), dinucleotide (AT) or even larger repeats. These polymorphisms are also called microsatellites and most often are multiallelic, since the number or repeats may vary greatly. This condition increases the likelihood of heterozygosity and makes VNTRs very informative for some genetic analyses, particularly linkage. VNTRs may also have a functional effect if present in relation to coding regions. As a typical example, type 1 diabetes has been associated to a VNTR in the insulin gene. Subjects with a short number of repeats (less than 50) have double risk than subjects with more than 200 repeats [10]. More recent findings link VNTRs and predisposition to early–onset colorectal cancer [103]. Though VNTR are very informative, their genotyping usually require more elaborated and expensive methods than SNPs (usually sequencing) and for this reason these polymorphisms are less often used for linkage and association studies nowadays.

Copy Number Variations have been identified more recently as an additional source of genomic variation. These are relatively large regions spanning kilobases, sometimes covering multiple genes, that appear in multiple copies with a variable number of repetitions, in the range of 0 (deletion) to tens [69]. CNVs are a typical genomic somatic alteration in most cancers. Germ line CNVs are also being studied as a potential cancer susceptibility source [77].

2 Evolution of genetic epidemiology: from family–based to population–based association studies

Finding cancer genes is a long task that needs to answer a series of questions (see Table 1). Each question usually requires a specific study design and measures genetic information with different levels of precision. Though the methods in this chapter will focus on association, it is important to know where this design is in relation to other alternatives.

Question	**Study design**
Are genes involved in cancer?	Familial aggregation, twin studies
What is the inheritance model?	Segregation
Where are the genes?	Linkage
Which are the genes?	Association
What is the causal variant?	Fine–mapping
Which is the mechanism?	Functional studies
Interactions	GxG and GxE association

Table 1 Relevant questions and study designs in genetic epidemiology.

The first and most important question is: Are genes involved in cancer? Case–control studies showing familial aggregation of cancer provide indirect information about the potential implication of genetic factors. Having a first degree relative with cancer is a risk factor for most frequent cancers. However, this is a very crude measure that might be confounded by shared environmental exposures. Studies in migrants may also be informative. Cancer rates in second generations of migrants that are more similar to their origin than the country of residence are indicative of a genetic component. Twin studies are the most powerful to estimate heritability (i.e. the proportion of cases attributable to genetic factors). The comparison of concordance rates between monozygotic and dizygotic twins, when combined with information about shared environment, provides most valuable information [51].

The occurrence of specific cancers sometimes is a recurrent event in some families. In such situations, when a major gene is suspected to be responsible for the disease, segregation analysis of the pedigrees can provide information about the inheritance model and estimates of penetrance [6]. These studies use only phenotype information and family structure and do not require DNA markers.

When enough information is accumulated about genetic factors as a cause of a specific cancer next question is: which are the genes? When genotyping of genetic markers became feasible, before the genome was completely sequenced, it was easier to identify regions of the genome associated to cancer and, in a second step, try to identify which gene in that region was responsible. Linkage studies explore a series of polymorphic markers carefully selected across the genome in large pedigrees of affected families. When at least three generations are genotyped, polymorphic markers can identify which alleles co–segregate with the disease and identify the regions most likely to carry the causal genes. Linkage analysis can combine the information of multiple families and is very powerful to detect signal when the penetrance is high, but since only about 400 markers are used to cover the genome, the level of resolution is in the range of megabases. After a consistent linkage signal has been detected, sometimes hundreds of genes may be in the region and this technique is not always able to improve the resolution even when increasing the number of markers because the number of subjects from the affected families is relatively small.

In order to identify the specific genes related to cancer, association studies with unrelated individuals using SNPs as genetic markers are the most powerful approach. Unrelated individuals increase the likelihood of recombination events and increase the resolution of the signal. Careful selection of SNPs, nowadays using information about haplotype blocks, can identify which genes are involved in the disease. Association studies compare the genotype frequencies of a series of SNPs between a sample of unrelated cases and a sample of controls from the same population. The possibility to include unrelated cases and controls allows the usage of large sample sizes to increase detection power. Association studies are usually focused on selected candidate genes selected belonging to regions that have shown linkage or because their known mechanism of action makes the gene possibly related to cancer. For example, typical genes studied in cancer are involved in cell cycle control, inflammation, metabolism, or DNA repair [49, 56].

Since current large–scale genotyping technology allows to simultaneously genotype millions of SNPs, currently Genome–Wide Association Studies (GWAS) are being conducted to identify susceptibility loci not necessarily related to coding regions. In fact, the first finding of these studies in prostate cancer has identified a region in 8q24 where no genes can be clearly imputed as responsible [40]. POU5F1 is the nearest expressed region, but corresponds to a pseudogene. MYC, a known oncogene that lies downstream the region, is also suspect of being involved, but the evidence is indirect [82].

Even when a gene has been clearly associated to a disease, finding the causal variant usually requires resequencing and intensive genotyping to fine–map the region. Identifying the causal variant will also need functional studies to document the mechanism of action that determines the risk.

For some genes, the genetic variation is probably not sufficient to cause cancer unless an environmental exposure is acting simultaneously. For example, polymorphisms in NAT2 have been associated to an increased risk of bladder cancer among smokers; for non–smokers the risk is not increased [34]. This is an example of gene–environment interaction that is probably relevant in many genetic determinants. Ignoring the environmental effect leads to an attenuated risk (average of smokers and non–smokers) that is difficult to detect unless the study has large sample size. Similarly to gene–environment interactions, it is likely that gene–gene interactions may exist and only carriers of multiple variants are at increased risk of developing cancer. The difficulty in detecting such interactions is that, without prior hypothesis, the search domain is huge and very large sample sizes are needed.

3 Technical issues and data quality control for SNP–array association studies

All biological experiments are subject to different sources of variability. Particularly, large–scale techniques, such as DNA microarrays, may be specially sensitive to specific experimental conditions that are not easy to keep under control [84]. Although there are some methods which try to minimize this variability, such as data normalization or experiment replication, it is not possible to remove it completely. Thus, besides being extremely careful about how all the experiments are performed and the data normalized, before going on with our analysis we will also need to check the quality of the obtained data to increase the reliability of the study results. As we previously stated, this chapter will mainly focus on SNP–array analysis techniques. Firstly, we will briefly review some the different genotyping algorithms that have been used to infer the calls from raw data. Once the genotypes are obtained, SNP–array quality control can be performed at different levels: SNP and sample (array). In the following sections we are going to briefly review some of the different calling algorithms explain, as well as explaining in detail this quality control procedure and all the steps it comprises.

3.1 Introduction to genotype calling algorithms

The *call* of a specific SNP for a single sample is essentially its genotype, that is, the combination of its two corresponding alleles. Since most usually we will be working with two–allele SNPs (also called biallelic), for a given SNP with alleles A and B there will be three possible calls: two homozygous (AA and BB) and one heterozygous (AB–or equivalently BA–). These calls are automatically obtained using algorithms that process raw intensities coming from the scanned image of the microarray. Usually, for a given SNP and sample we will have two intensity values, each one corresponding to one of the two alleles. Some array platforms, however, have also probes which are strand–specific (sense and antisense), finally leading to 4 intensity values.
Over the last few years, genotyping algorithms have evolved in accordance with the size of the available arrays. The embryo technology of the SNP arrays, Affymetrix Variation Detection Arrays (VDAs), contained about 1500 SNPs. An algorithm called ABACUS (Adaptive Background Genotype Calling Scheme) was then designed to extract the calls [23]. As well as showing a certain trend to drop heterozygous calls, this method was clearly unsuitable when the first SNP arrays appeared (e.g. Affymetrix 10K). Thus, ABACUS was soon replaced by newer algorithms such as MPAM (Modified Partitioning Around Medoids) [53]. This algorithm was based on the robust classification method called PAM (Partitioning Around Medoids), but is was modified to penalize small between–group distances, since PAM tends to split large clusters into two different groups to minimize the total sum of distances

of all the observations to their corresponding nearest medoid. MPAM worked well for SNPs that had enough data in each of the three genotypes, but not as well for SNPs with one missing or very small genotype or when the number of arrays to be analyzed was small.

With the advent of 100K arrays, a lot of SNPs with low minor allele frequency (see section 3.2.3) were included in the new platform, making the performance of MPAM decreased remarkably. Thus, it was replaced by the newer DM (Dynamic Model) algorithm [26], in which four Gaussian models were fitted for the probe intensities of each SNP (one for each genotype and one for the null values), and then a genotype call was assigned to each sample depending on its likelihood. The DM algorithm had a main limitation: it was a single–array algorithm, that is, it could not take profit of aggregating data sets to better assess how each SNP behaved. Furthermore, it seemed to poorly classify heterozygous samples when compared to MPAM.

Arguing that neither MPAM nor the DM algorithm were using currently available genotypic information, and only about a year after the publication of the DM method, [68] proposed a new algorithm, called Robust Linear Model based on Mahalanobis distance classification (RLMM). This method had two main advantages over the formerly designed DM algorithm: firstly, it was a multi–chip algorithm, thus allowed to assess both probe effects and allele signals for each SNP. Secondly, genotypes were estimated by means of a multiple–sample classification, that is, using information of other SNP to better define the properties of the three groups corresponding to the three possible genotypes. To combine intensities across probes and arrays and produce allele–based summaries it used the robust multi–chip average method [46]. This method took advantage of the large amount of publicly available information on genotype calls (i.e. HapMap) to define regions for each genotype group, thus improving the accuracy of the classification. Nevertheless, although this remarkable increase in accuracy, RLMM still had some problems in dealing with the inter–study or inter–laboratory variability, which may be caused by sample preparation procedure, among other reasons.

Affymetrix soon adopted RLMM as the standard methodology to analyze 100K and 500K SNP arrays. The method was slightly modified with the addition of a Bayesian approach which yielded differences in the clustering space transformation and in the estimation of both cluster centers and variances. Although it was its main aim, the resulting algorithm, known as BRLMM (Bayesian RLMM) [2], still did not seem to handle accurately the inter–study variability.

To solve this issue, Carvalho et al. proposed another modified version of the RLMM, known as CRLMM (Corrected Robust Linear Model with Maximum Likelihood Classification) [14]. Essentially, it uses an adapted version of RMA preprocessing method for SNPs (called SNP–RMA), which is designed to remove most of the study/laboratory effect. In much the same way as BRLMM does, it also uses Bayesian approach to inform lowly populated clusters. Recently, CRLMM designers have added a new recalibration step to the algorithm which further increases its accuracy level [52]. The new version also incorporates a new quality metric to assess call confidences at the SNP level, which may be very useful to filter out poor–quality SNPs . This method seems to perform better than all the previous algorithms explained, and even better than Birdseed [48], which is the recently designed algorithm by Affymetrix and the Broad Institute for the 6.0 generation of SNP arrays, so it may be a good choice if we need to decide which calling method we are going to use.

Finally, we must point out that although all the calling algorithms we have mentioned in this section have been basically designed to be applied to Affymetrix SNP arrays, the underlying basis of the analysis can also be suitable for arrays made by other manufacturers, such as Illumina's BeadChip technology.

3.2 SNP–level quality control

This is the first level of quality control. Once we have the SNP calls, it is important to check their quality one–by–one in order to detect uninformative or poor–quality SNPs, so that they can be permanently removed from all the samples contained in our dataset.

SNP–level quality control can be divided in different parts, each one of them checking different quality issues. SNPs that meet *all* the requirements are the ones that will be kept for further analysis.

3.2.1 Percentage of present calls

As we stated in section 3, defective hybridizations or incorrect analytical processes may result in poor quality data and could hamper obtaining reliable genotype calls. In the case of SNP–level percentage of present calls, difficulties may arise mainly from improper functioning of genotype calling algorithms. Some of them, such as BRLMM, may introduce some systematic bias in the missing values they report [43]. As a consequence, this bias may not randomly affect all three different genotypes of a SNP, but only some of them. Having the calls for all SNPs and samples, we can then assess the missing rate for each SNP across all the hybridizations. Since missing call values will potentially be related to low levels of genotyping quality, we should discard from our study those SNPs with a poor call rate. Although rather subjective, an 80% of present calls is usually considered as the minimum threshold applied to filter out potential low quality or highly biased SNPs.

3.2.2 Hardy–Weinberg equilibrium (HWE)

Given a SNP and its allele frequencies for a specific population, the Hardy–Weinberg principle determines what the expected genotype frequencies should be, assuming that the different alleles are transmitted independently from one generation to another and with no selective pressure over them. Therefore, if we have a SNP with two alleles, A and B, with population frequencies p and q (or equivalently

1-p) respectively, the expected genotype probabilities are:

$$
\begin{aligned}
f_{AA} &= p^2\\
f_{AB} &= 2 \cdot p \cdot q\\
f_{BB} &= q^2
\end{aligned}
$$

To assess if one SNP follows the Hardy–Weinberg law we can use the Pearson's goodness–of–fit Chi–square test statistic, χ^2, with one degree of freedom (in case we have a biallelic polymorphism). The null hypothesis is that the SNP is indeed under HWE, so we will reject SNPs with p–values smaller than a specific significance level. The Chi–square statistic, however, may have a poor performance when we have small genotype counts, so in that case it will be better to use a Fisher's Exact test instead [38, 100].

The fact that a SNP does not follow the Hardy–Weinberg law may be due to different reasons:

- Small population size.
- The allele–calling algorithm is underperforming for one of the genotypes (i.e. it fails to correctly call heterozygotes).
- The SNP is mapping to multiple genomic locations.
- The genotyped individuals are not independent (i.e. because of inbreeding).
- There has been a positive selection of a certain allele (i.e. an allele associated to longevity).
- If we use a significance level of a 5%, we may find by chance different observed frequencies from the ones we expect. This would happen for a 5% of the SNPs for which we are evaluating HWE. Theoretically, we would need to perform p–value adjustment (see section 6.4) to solve this issue. Nevertheless, what is usually done in this context is to set a more restrictive threshold for significance for HWE tests, but not as restrictive as it would be using standard p–value correction methods. Researchers have widely accepted 0.001 as a suitable boundary, being 0.0001 in the case of GWAS, where more tests are performed. Even if after setting this more astringent threshold we still find SNPs with genotype frequencies under no HWE then one of the other issues on this list may be the reason, so we will need to evaluate our data in detail to find what is causing this genotypic imbalance.

In the typical case–control study, HWE may be only evaluated in control populations, which is where it should hold true. If we do not identify clearly the reason of the disequilibrium, it may be necessary to remove the affected SNPs. In case we want to keep them, association results for those SNPs need to be checked carefully, as there may be some influence from one of the items mentioned above. As a guidance, we can also look at HWE among the cases, since a SNP with no equilibrium might be potentially related with the disease.

3.2.3 Minor Allele Frequency (MAF)

The *minor allele frequency*, or MAF, of a SNP is its lowest allele frequency. There is a huge variation in the MAF among different SNPs, from a very low percentage (e.g. less than 1%) to almost a 50%, meaning that in fact there is no minor allele. Although the MAF is not a quality measure by itself, it might be useful to filter SNPs according to it for subsequent analysis. It must be taken into account that, if a certain SNP has a very low MAF, we will have very little statistical power to detect its potential association with the disease (see section 6.3). Furthermore, these SNPs are more difficult to genotype reliably. Therefore, removing those SNPs, from which a priori it will be hard to get any useful information, might increase the quality of our data and will slightly reduce the number of hypothesis tested, so we will be a bit less restrictive in the step of p–value correction for multiple hypothesis testing.

3.2.4 Genotype calling and exploration of signal intensity plots

As we have seen in section 3.1, SNP allele intensities are the data we use to infer the genotypes. Since all preprocessing and calling procedures are mostly automatic, we do not usually work with these intensities directly. Nonetheless, we can still use them if we are specially interested in checking a few specific SNPs. To do so, signal intensity plots are mainly used. For a given SNP, these plots are useful to visually inspect the intensity values for both alleles across all samples. Under an ideal situation we will observe three clouds, one for each genotype (Figure 1, left panel). However, some issues may influence this intensity values, thus distorting the plot and making it more difficult to visually define the three clusters (Figure 1, right panel). This will happen mainly for bad quality SNPs, SNPs with poor intensities, SNPs with homologous sequences in different parts of the genome or SNPs involved in a CNV (Copy Number Variant), among other reasons.

Since checking this plots is mainly a visual inspection that needs to be done SNP–by–SNP, it will be virtually infeasible to have a look at the hundreds of thousands of SNPs contained in an array. Thus, more than a pre–analysis quality control step, this should be considered a post–analysis quality control procedure. That is, when we have a few candidate SNPs and we want to ensure the reliability of the obtained results, we can plot their intensities and see how the calling algorithm has created the different groups. Furthermore, if we find a strange negative result (i.e. lack of association with the disease when we already expected it) we can do the same to check if it has been caused by any of the technical reasons stated above.

3.3 Sample–level quality control

As well as performing SNP–level quality control, it is important to check whether there are any poor–quality samples in our dataset or not. This fact could happen

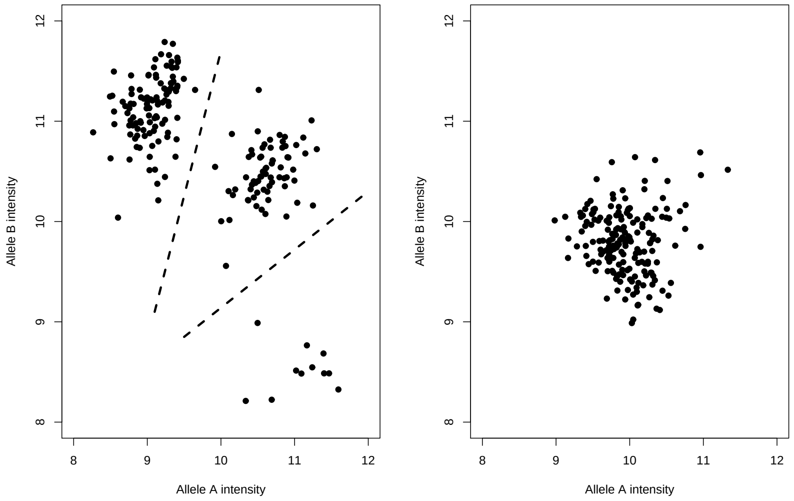

Fig. 1 Intensity plots of allele A vs. allele B for two SNPs. In the left panel we see a good quality SNP, where the boundaries between the three genotype regions (dashed lines) can be clearly defined. The right panel belongs to a defective SNP, where no clear boundaries between genotypes regions can be defined, yielding to potentially incorrect calling results.

not only because of bad quality of the biological material to be hybridized (e.g. DNA), but also because there may be some problem with the hybridization. That is, since samples and arrays are confounded, sometimes it will be hard to tell where the problem comes from unless we perform replicates of our experiments, which is rather expensive. Independently of the underlying reasons though, all samples with a poor quality level should be removed from our dataset for further analysis, as they could be a potential source of error in our study. In the following sections we will review which parameters can be used to detect these defective hybridizations.

3.3.1 Percentage of present calls

Much in the same way as we explained in section 3.2.1, percentage of present calls in a sample can be a helpful quality index. In the case of sample percentage of present calls, this fact could be mainly related to the quality of the hybridized DNA (i.e. it may be degraded or the amount of DNA hybridized may be too small), as well as with some technical problem with the hybridization or with the microarray itself. Usually, those samples with less than an 95 or 97% of present SNPs should be discarded, since missing genotypes tend to be non–randomly distributed. This threshold may be increased or decreased depending on how strict we want to be with our data. If for any reason we decide to lower it significantly (e.g. less than

90%), we must always bear in mind that our final results may be influenced by this potential artifact.

3.3.2 Sample heterozygosity

Total heterozygosity, understood as the number (or proportion) of heterozygous SNPs in one sample, can be a good quality indicator at the sample level. As an example, individuals having a large proportion of heterozygous SNPs may be more likely to have their DNA contaminated. On the contrary, a too low level of heterozygosity could indicate that there may be some problem with the hybridization or even a sign of inbreeding for that individual. A rather simple but typically used methodology to filter out those samples with an odd level of heterozygosity is to compute the mean and the standard deviation of this index across all samples and then filter out those individuals falling outside the mean ± 3 SD.

Regarding heterozygosity analysis, it is interesting to remark that we should pay special attention to SNPs located in the X chromosome, since presence or absence of heterozygous SNPs in a specific sample will help us decipher the gender of that individual (i.e. only females can have heterozygous SNPs in the X chromosome). This will enable us to check for possible mistakes during the process of sample annotation.

3.3.3 Using Principal Components Analysis as a method to detect outliers or related samples

Even after removing those defective SNPs and samples by methods such as the ones described in previous sections, to reach the maximum level of quality in our data we must still ensure that none of our individuals displays an irregular genotype pattern. As an example, this could happen if, by mistake, a Chinese or African individual falls into a study of Caucasians. By applying all the filters mentioned above we may not detect this fact, so we need to use techniques capable of discovering this kind of outliers. Additionally, another issue we must take into account is to search for any underlying relationships between individuals in our cohort. That is, if we have samples with a higher level of concordance between their genotypes than we would expect by chance. This relationship could be due to technical (e.g. date of hybridization, batch effect, etc.) or biological reasons (e.g. inbreeding).

A useful technique to perform all these quality controls is Principal Components Analysis (PCA). Basically, it is a dimensionality reduction technique which transforms an undefined number of correlated variables (which in this case would be the samples) into a smaller number of uncorrelated and ordered variables, called principal components. The order of the principal components is arranged according to the amount of variability explained by each one of them, being the first principal component the one that accounts for most of the variability.

Before doing the analysis, we will need to perform a simple transformation of the genotype matrix into a numerical one. That is, we will recode genotypes, such as AA–AB–BB, into numbers (e.g. 0–1–2). Although this may be enough to detect outliers in our dataset, a transformation of the matrix as the one suggested by Price et al [63], which takes into account the differences in the MAF of all the SNPs, may be appropriate. Performing a PCA is a rather straightforward task. Nonetheless, it may be memory consuming for very large datasets, so in this case we may need a computer with a fairly big amount of RAM memory for this purpose.

Once the analysis is done, a rather simple but useful way to check for outliers or unlikely relationships among samples is to plot the values of the first principal components for all the samples contained in our dataset. As an example, in Figure 2 we can see a plot of the two first principal components for a dataset containing 94 HapMap samples and 6359 SNPs located in the genomic region 8q24. Two of the samples have Asian origin (Japanese in Tokyo, Japan), and two more come from Africa (Yoruba in Ibadan, Nigeria), being the remaining ninety CEPH (Utah residents with ancestry from northern and western Europe). In the plot we can clearly see how JPT and YRI individuals can be clearly distinguished from CEU, which form a relatively homogeneous group. Therefore, this procedure has enabled us to uncover those samples that may be defective or have a different origin than we could have expected. Although the example is show to reveal different ethnic origins, it can be also useful to detect technical biases or batch effects in our dataset. A most extensive application of PCA to genome-wide association studies is reviewed in section 6.1.

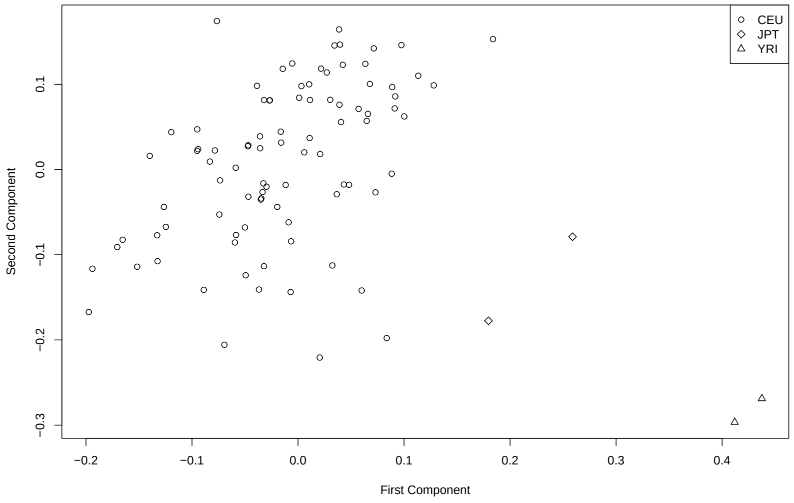

Fig. 2 Principal components analysis plot of 94 HapMap samples using 6359 SNPs located in the 8q24 region. We can clearly see how the two African and the two Asian individuals separate from the rest.

4 Single–SNP analysis: association between SNPs and a trait

Single–SNP analysis is usually the first step of the analytical process after performing quality control in our dataset. Essentially, it consists on assessing the association between the different genotypes of a SNP and a response variable. This has usually been the most straightforward and computationally feasible type of analysis in association studies. Nonetheless, with the recent advent of more dense platforms such as Affymetrix's SNP array 6.0 or Illumina's Human1M–Duo BeadChip, single–SNP association analysis has also become a challenge both for statisticians and bioinformaticians, specially for those association studies with a large sample size. The vast amount of data generated by these arrays demands good computational and statistical skills, as well as a computing infrastructure powerful enough to handle it properly. In the following sections we are going to review which are the different available tests to evaluate association between a SNP and a specific trait. We will divide the different types of analysis in terms of the type of outcome we have: binary, quantitative or prognosis.

4.1 Binary outcome

The scenario with a categorical outcome is one of the most, if not the most, usually found in population association studies. More specifically, what we typically have is a binary *case–control* phenotype, with unrelated affected and unaffected samples. Cases and controls are usually matched by third variables such as gender or age. In this case, the most straightforward manner to test the SNP–outcome association can be done based on the 3 × 2 contingency table (Table 2). To test the null hypothesis of no association between genotypes and the response variable we can perform a 2 df χ^2 test. Nonetheless, when there are low genotype frequencies in one ore more cells a Fisher's exact test would be desirable.

Genotype	**Controls**	**Cases**
AA	$n_{AA_{co}}$	$n_{AA_{ca}}$
AB	$n_{AB_{co}}$	$n_{AB_{ca}}$
BB	$n_{BB_{co}}$	$n_{BB_{ca}}$

Table 2 Contingency table with genotype counts for cases and controls.

The model which takes into account the full table with the three genotypes is usually called codominant. Basically, it assumes a phenotypic intermediate effect (but not necessarily half–way) for the heterozygotes compared to the two homozygotes. However, sometimes we will expect our SNPs to follow other inheritance patterns, such as dominant or recessive. To do so, we will rearrange the full contingency table as shown in Tables 3 and 4.

Genotype	**Controls**	**Cases**
AA	$n_{AA_{co}}$	$n_{AA_{ca}}$
AB+BB	$n_{AB_{co}} + n_{BB_{co}}$	$n_{AB_{ca}} + n_{BB_{ca}}$

Table 3 Contingency table for the dominant model, with allele B being the risk allele. In this case heterozygous individuals are expected to have the same phenotype as BB homozygotes, so both categories are collapsed into a single one.

Genotype	**Controls**	**Cases**
AA+AB	$n_{AA_{co}} + n_{AB_{co}}$	$n_{AA_{ca}} + n_{AB_{ca}}$
BB	$n_{BB_{co}}$	$n_{BB_{ca}}$

Table 4 Contingency table for the recessive model, with allele B being the risk allele. Heterozygous individuals are only *carriers* of the disease, but they display a non–disease phenotype. Thus, they are merged with the AA subjects, which have a wild–type genotype.

Analysis of Tables 3 and 4 can also be done both with a χ^2 test (with 1 df) or a Fisher's exact test, depending on the number of counts found in each cell.

One important fact to take into account is that, when dealing with complex traits, genetic effects of single SNPs will likely be additive instead of dominant, recessive or codominant. Additive models assume that heterozygotes risk will be half–way between the two homozygote risks. Since the statistical tests explained above (2 df χ^2 and Fisher's exact test) have a reduced power to detect this kind of effects, we need to address this issue and find other ways to detect this additive association. The Cochran–Armitage test [5] is a good model to detect this trend in the proportion of cases for each one of the genotypes. It tests against the null hypothesis of a zero slope for the line that fits the three genotype risks best. Actually, this test corresponds to the *multiplicative* model for effects of alleles on odds scale. An important characteristic of this model is that it does not rely on the assumption of Hardy–Weinberg equilibrium, so it may be useful in case HWE does not hold for our complete population of individuals (cases and controls altogether).

Compared to the contingency table approach, logistic regression offers a more flexible environment to assess the association between a SNP and a binary outcome. For large sample sizes, the likelihood ratio test of the logistic model against the null hypothesis $\beta_{AA} = \beta_{AB} = \beta_{BB}$ is equivalent to the 2 df χ^2 test. However, logistic regression can be extended to further SNPs (epistasis), environmental or clinical variables which usually need to be taken into account.

To specify inheritance models in a logistic regression, we just need to restrict the values of the β coefficients. Thus, forcing that $\beta_{AB} = \beta_{BB}$ or $\beta_{AA} = \beta_{AB}$ would test for a dominant and recessive effects, respectively. If we restrict β_{AB} to be half–way between β_{AA} and β_{BB}, then the logistic model will be equivalent to the Cochran–Armitage trend test.

4.2 Quantitative outcome

A typical example of association study with a quantitative outcome is the one where we want to test if the expression value of a gene is affected by the genotype of a specific SNP (which may or may not be located in the same gene). This kind of association may be relevant for diseases with an important genetic basis, such as cancer.

A first and simple approach to assess the degree of association between a SNP and a trait would be to categorize the quantitative response into two classes (e.g. ”low value”, ”high value”), and then apply one of the approaches described in section 4.1. Nevertheless, this approach is suboptimal, since it would carry a loss of statistical power to detect significant changes between groups. Therefore, instead of that approach, a more natural and optimal model to test association with a quantitative response is to use statistical tests such as ANOVA and linear regression.

While ANOVA model is equivalent to the 2 df χ^2 test, linear regression assumes linearity between genotypes and the response means, so the degrees of freedom are reduced to one. Furthermore, both tests require the trait to be normally distributed and with equal variance across all genotypes.

In a similar manner as what is explained in section 4.1, inheritance models can also be specified in this case by merging the proper genotypes to generate a dominant, recessive or additive genetic pattern.

4.3 Prognosis outcome

In the last few years, a vast number of studies have investigated the association between polymorphisms and cancer survival. Some of the more recent findings include studies for breast cancer [22, 44, 55, 97], colorectal [12, 40, 47, 90] or prostate [40, 39, 88, 101, 102].

From a statistical point of view, the Kaplan–Meier estimator is the most widely used to estimate the survival function. As an example, we could model the time to develop metastasis after the resection of a primary tumor in terms of the genotype of a specific SNP. To assess the significance of survival differences in different groups a log–rank test can be used. However, Cox proportional hazards model will allow us to quantify the increase or decrease of risk for each one of the genotypes. Analogously to what is explained in section 4.1 and 4.2, in this case we can also force our SNPs to follow a specific inheritance pattern.

5 Multiple–SNP analysis

Association studies may not restrict only to single genetic markers, specially when most recent large–scale techniques have broaden the experiments up to more than a million SNPs. Although useful as a first approach to detect potential association with a trait, single–SNP analyses have shown to be somehow inefficient, because they do not integrate information of nearby markers. Since it may be rather unlikely that we have the *real* causative marker genotyped, multiple–SNP associations can provide a great advantage over pointwise estimations.

There are two main approaches to assess multiple marker association: regression and haplotype–based methods. Regression methods are mainly based on logistic or linear models (depending on the type of response we have). Nonetheless, as genotyping densities have dramatically increased over the last few years, correlations among neighboring SNPs can cause model instability. Backward or forward stepwise procedures may overcome this limitation, but they tend to overfit the observed genotype and phenotype data, making permutation testing procedures necessary to control the type I error rate. Another feasible approach is to select only tag–SNPs (i.e. loci that can serve as proxies for many other SNPs, see section 5.2), but at the expense of losing potentially valuable information. This motivates us to focus on haplotype–based approaches, which constitute an attractive alternative. In the following sections we are going to introduce the basic concepts of haplotype theory, and then we will review haplotype–based association methods.

5.1 Introduction to haplotypes

Haplotypes are combinations of alleles at multiple polymorphic loci along a chromosome. Although an entire chromosome could be seen as a haplotype, usually only regions no longer than 100 Kbp with highly linked polymorphisms are considered. Thus, for a given set of markers, each person has two haplotypes, each one inherited from one of the progenitors. As one can easily calculate, a set of n biallelic SNPs generate 2^n potential haplotypes in the population. However, recombination rates commonly make the actual occurring number of haplotypes be much smaller than this theoretical upper bound.

The usefulness of haplotypes in association studies is justified by several reasons. Firstly, since they are combinations of multiple SNPs, haplotypes have been demonstrated to be more informative than individual markers. Furthermore, haplotype association studies show greater statistical power than single–SNP association analyses [3]. From a biological perspective, there are evidences that a set of pointwise cis–mutations (i.e. located in the same copy of the chromosome) within the same gene can interact to have a greater effect on a subject's phenotype. Despite this, the association of a haplotype with a phenotype does not necessarily mean that the haplotype itself is biologically related to the trait, since it may be possible that an unexplored locus located in the haplotype region was the marker biologically functionally related with the phenotype.

One serious drawback of any analysis involving haplotypic information is that genotyping studies usually generate unphased data. That is, for a given subject we do not really know which alleles come from each one of the progenitors. Laboratory techniques which allow to obtain phase information, such as allele–specific PCR or cloning, are rather expensive and time consuming. Thus, to overcome this lack of information we need a statistical approach that enables us to *infer* haplotypes for a given set of unrelated samples and genotypic markers.

5.2 Linkage disequilibrium, linkage blocks and tag–SNPs

Linkage disequilibrium (LD) statistics describe the deviation of observed haplotype frequencies from what is expected. Let A and B be two SNPs with alleles A_1, A_2, B_1 and B_2. Thus, the combination of these SNPs can generate four possible haplotypes: A_1B_1, A_1B_2, A_2B_1 and A_2B_2, with relative frequencies $f_{A_1B_1}$, $f_{A_1B_2}$, $f_{A_2B_1}$ and $f_{A_2B_2}$, respectively. The basic statistic to assess the LD between both markers, named D, is defined as follows:

$$D = f_{A_1B_1} - f_{A_1} \cdot f_{B_1} \tag{1}$$

D equals to 0 in the case of *complete equilibrium*. Positive D values indicate that A_1 and B_1 tend to appear together more than expected by chance, while negative values would indicate the opposite. A major inconvenient with the D statistic is that its range depends on the MAF of the two SNPs, making it desirable to find a measure with a standardized range. Thus, a normalized version of D, called D', is defined as:

$$D' = \frac{D}{D_{max}} \tag{2}$$

where

$$D_{max} = \begin{cases} \frac{D}{min(f_{A_1}f_{B_1}, f_{A_2}f_{B_2})} & if\ D > 0 \\ \frac{D}{min(f_{A_1}f_{B_2}, f_{A_2}f_{B_1})} & if\ D < 0 \end{cases} \tag{3}$$

D' ranges from -1 to 1, and usually takes extreme values when allele frequencies are small. If $D' = 1$ or $D' = -1$, it means there is no evidence for recombination between the two markers. Moreover, if allele frequencies are similar, high D means the SNPs are good surrogates for each other. Nonetheless, this statistic has an important drawback, which is that it is inflated for small sample sizes or when one allele is rare. Therefore, another measure based on the correlation between alleles, called r^2, can be defined as follows:

$$r^2 = \frac{D^2}{f_{A_1} \cdot f_{A_2} \cdot f_{B_1} \cdot f_{B_2}} \tag{4}$$

r^2 ranges from 0 (i.e. perfect equilibrium) to 1 (i.e. both markers provide identical information), and its expected value is $1/2n$. It has become one of the most used statistics to assess LD between pairs of markers.

Comparison of haplotypes and the scope of LD across individuals allows us to identify segments or haplotype blocks that correspond to minimal units of recombination. Usually, one or few alleles within these haplotype blocks will be predictive of the other alleles. This predictive SNPs are called *tag–SNPs*. Therefore, genome–wide association studies (GWAS) can be accomplished by genotyping a collection of tag–SNPs which define the haplotype blocks along the complete genome. As an example, this is the approach followed by Illumina's BeadChip technology.

5.3 Haplotype inference

In the last two decades several methods of haplotypic reconstruction have been developed in order to solve the problem of haplotype inference. Since Clark, in 1990 [18], developed a parsimony algorithm to estimate haplotype frequencies from a sample of genotypes, quite a large number of methods have been developed. Most of them rely on the use of different techniques to calculate the Maximum Likelihood Estimator (MLE).

In 1995, Excoffier and Slatkin [31] adapted the Expectation–Maximization algorithm, an iterative algorithm of maximization developed by Dempster in 1977 [24] to maximize the likelihood function of the haplotypes given the genotypes at specific loci. This method has some limitations and convergence to a local maximum may occur in some situations [15].

Some authors have attempted to minimize these limitations in their works, like Qin *et al.* [67] using *Divide and conquer* strategies, or David Clayton, implementing an EM–algorithm which adds SNPs one by one and estimates haplotype frequencies, discarding haplotypes with low frequency as it progresses. In the context of Bayesian statistics, Stephens *et al.* in 2001 proposed an algorithm based on coalescent theory [87] with a especial prior based on the general mutational model. Niu *et al.* [58] implemented another Bayesian approach using a Markov Chain Monte Carlo method. In general, algorithms dealing with Bayesian models are suitable to infer haplotypes from genotypes having a large number of polymorphisms.

More recent methods work with clusters of haplotypes in order to avoid the major limitations of many current haplotype–based approaches [96].

Once the frequencies have been estimated by any of the methods mentioned above, the next goal is to test the association between haplotypes and the disease. The most accurate strategy in order to take into account the uncertainty of the sample is to estimate simultaneously haplotype frequencies and haplotype effects. Some works are focusing on this approach [45, 89, 92].

5.4 Haplotype association with disease

Since haplotypes capture variation in small genomic regions, the analysis of association between haplotypes and disease is a potentially more powerful way to identify cancer genes when the causal variant is unknown [3]. Haplotypes inferred from a series of SNPs should also capture variation in VNTRs and CNVs.

From an analytical point of view, the possible haplotypes at a given region conform a categorical variable that can be analyzed in regression models when appropriately coded with indicator variables. There are a few technical difficulties with this analysis. First, haplotypes are inferred from genotypes, as we have seen previously, and for subjects heterozygous at more than two SNPs there is uncertainty about the pair of haplotypes. A similar problem arises when one or more genotypes are missing. For these cases, the inference algorithm used provides a posterior probability for each compatible combination. These probabilities can be used as weights in a regression model to transfer the uncertainty in the haplotype estimation to the estimates of association. This is the method used in haplo.stats software (see section 8.2.3). The second problem is the inheritance model. Since each subject carries two haplotypes (though in the dataset is further expanded to account for uncertainty), the most frequent inheritance model used is the log–additive, where the risk for each haplotype is compared to one selected as reference in logistic regression. Usually the most frequent haplotype is this reference. The odds ratio estimates obtained should be interpreted as per–haplotype relative risks, similar to the per–allele relative risk in a log–additive model for genotypes. There is a possibility to encode the haplotypes to model dominant or recessive effects [45], but the interpretation is not as simple and for the recessive effects the power is generally very limited. A final consideration in these analyses is the treatment of rare haplotypes (i.e. those with an observed frequency lower than a previously defined threshold–usually 0.01%–). The inference process usually results in a series of haplotypes with very low inferred frequency in the studied population. If these rare haplotypes are considered in the logistic regression, the model becomes unstable because most probably these haplotypes have only been observed or inferred for cases or for controls, and the standard errors of the regression coefficients become very large. The typical solution is to pool these rare haplotypes into a common group that is not interpreted. If the cumulative frequency of these rare haplotypes is not high and the reference category is large enough, a better option might be to pool them into the reference category. With this method one degree of freedom is gained for the test of global association between the haplotypes and the disease.

The analysis of haplotypes can also be useful to identify genes associated to prognosis of cancer. For this analysis, if a Cox proportional hazards model is desired, the combined estimation of haplotypes and regression parameters is more difficult computationally due to the semi–parametric nature of the Cox model. This analysis, could be approached within a Bayesian framework.

6 Genome–Wide Association Studies (GWAS)

As already mentioned in 2, candidate gene association studies are designed to assess association between a moderate number of SNPs and disease. These kind of studies can be viewed as hypothesis–based studies in which the a priori knowledge of the disease plays an important role. Another reason for adopting candidate gene approach is the low costs of genotyping since only a moderate number of makers have to be genotyped. The price we have to pay, however, is that only those genes with known functional impact are included in the analysis. The continuous improvements in genotyping technologies and, above all, their decreasing cost has made it possible to perform genome–wide association studies (GWAS) where the entire human genome is interrogated. GWAS use large–scale genotyping technologies to assay hundreds of thousands of SNPs and relate them to clinical conditions or measurable traits. One of the main strength of a GWAS is that they are unconstrained by prior hypotheses with regard to genetic associations with disease [41]. Recently, there has been a vast number of GWAS to determine new susceptibility locus contributing to complex diseases such as Alzheimer [8], cancer [30], Schizophrenia [59] or Parkinson [61] as well as quantitative traits such as metabolic traits [70] or serum IgE [98], among others. A remaining obstacle of GWAS is that a massive number of statistical tests is performed. This may lead to a huge number of false–positive results making necessary to adopt further statistical corrections in the assessment of association or replication.

In this section, we will give an overview about GWAS, including study designs and statistical tests. We will also illustrate how to determine the statistical power and suitable sample size of our study, as well as addressing the multiple comparisons problem.

6.1 Study designs

GWAS can be defined as the study of common genetic variations across the entire human genome. They are designed to identify associations with observable or quantitative traits (such as blood pressure or weight), or the presence or absence of a disease or condition (i.e. discrete traits). Depending on the nature of the trait (e.g. quantitative or discrete) the study design may differ.

The mostly used design for GWAS has been, so far, the case–control design, which has been frequently used in traditional epidemiological studies. They are less expensive to conduct than cohort studies, and genetic studies can be easily incorporated. This is possible because many epidemiological studies collect blood samples at the beginning of the study that afterwards can be used to perform genetic tests. The aim of genetic case–control studies is to compare allele (or genotype) frequencies in diseased individuals with frequencies in healthy controls. The only difference between GWAS and other genetic association studies is simply the number of SNPs analyzed. Their characteristic limitations are those of case–control studies such as recall, selection and confounding bias. Recall bias arises when cases report their exposure history in a different manner than controls. This is not a problem when assessing genotype–phenotype associations, because genotypes (i.e. *exposure*) are measured from DNA samples. Nonetheless, this may be relevant when studying gene–environment (GxE) interactions. Furthermore, in some occasions DNA collections may differ with regard to storage, technicians or genotyping methods that could induce to some systematic bias [20]. On the other hand, selection bias occurs when controls do not come from the same population as cases. In this case, genetic or environmental background may differ as a result of the study design and not due to genetic differences. This can be a concern in studies that use controls who were selected and genotyped for a previous study [21]. Finally, confounding bias occurs when a risk factor for disease is also associated with the marker. Some authors stated that genetic association studies are protected against this bias since genotypes at a given locus are independent of most environmental factors [19]. In genetic association studies, there is an special case of confounding bias known as *population stratification*. This situation appears when both disease and allele frequencies are correlated across ethnicity. This difficulty may be overcome either in the study design or in the analysis process. When designing the study, one may select controls matched by ethnicity with cases or select controls from the same family as cases (paired design). However, if this matching cannot be performed, population stratification needs to be addressed at the analytical stage. There are several procedures for addressing population stratification from a statistical point of view. Some of them are based on correcting the test statistics used to assess association by computing an inflation parameter, while others try to find the underlying structure of the data and its variability and incorporate it into the analysis.

Population stratification inflates chi–square values when assessing association among makers and disease. After estimating the inflation parameter, λ, one can correct the chi–square test of association dividing it by λ. Different methods exist to estimate λ. One of them, known as *genomic control* [25], uses the variance inflation factor to correct for the variance distortion estimates. In this case, the inflation parameter can be estimated as:

$$\hat{\lambda} = \frac{\texttt{median}(\chi_1^2, \chi_2^2, \ldots, \chi_M^2)}{0.466},$$

where $\chi_1^2, \chi_2^2, \ldots, \chi_M^2$ are the chi–square test statistics for the M markers analyzed. Another approach, known as *delta centralization* [37], centralizes the non–central chi–square distribution of the test statistic. In the presence of population stratification, the test statistic used for assessing association follows a non–central chi–square distribution with non–centrality parameter δ^2. In this case, δ^2 is used to correct the test statistics as λ [37]. Finally, [20] uses the relationship between observed and expected test statistics for disease association. The tests are ranked from smallest to largest and plotted against their expected order statistics under the global hypothesis of no association. Under no population stratification the points should be in the diagonal line. The authors estimate λ, by calculating the ratio between the mean across the smallest 90% of observed test statistics and the mean of the corresponding expected values. 90% of tests are considered because it is expected to have a little proportion of SNPs that are truly associated with the disease (i.e. they do not hold null hypothesis) for which the observed test statistic should be inflated. Nevertheless, methods based on adjusting association statistics at each marker by uniform overall inflation factor, may be insufficient for markers having unusually strong differentiation across different populations, leading to a loss in power [63].

Another approach is based on finding structured associations with the program STRUCTURE [64]. This method assigns individuals to discrete subpopulation clusters and then aggregates evidence of association within each cluster. It produces highly accurate assignments using few loci. One limitation of STRUCTURE, though, is that is computationally intensive and cannot be applied to the whole set of SNPs comprised in a GWAS. However, this method has been demonstrated to work reasonably well only with a subset of the SNPs. Furthermore, the assignment of individuals to clusters is very sensitive to the number of clusters, which is not well defined [63]. A better alternative is proposed in [63]. The authors proposed a method called EIGENSTRAT consisting of three steps. Firstly, principal components analysis of genotype data is applied to infer continuous axes of genetic variation. They show that the top two axes of variation describe most of the observed variability and can be used to correct for population stratification. Secondly, they adjust genotypes and phenotypes by amounts attributable to unobserved population (e.g. ancestry) along each axis. Finally, in the third step, association statistic is computed using population–adjusted genotypes and phenotypes. This last step can be performed by fitting a logistic regression model adjusted by the first (generally the two first) principal components. One of the main advantages of using continuous measurements for adjusting population stratification is that it provides a better description of genetic variation among individuals. Another important point is that EIGENSTRAT is computationally tractable on a genome–wide scale.

To illustrate how EIGENSTRAT works, we use HapMap samples (see section 8.3.2 for more details). We randomly selected a set of 9,307 SNPs from the entire genome for 270 individuals from different CEPH (Utah residents with ancestry from northern and western Europe–abbreviated as CEU–), subjects with African origin (Yoruba in Ibadan–YRI–) and Asian individuals with Japanese (JPT) or Chinese (CHB) origin. Figure 3 shows the two first components (axes) of variation and the position for each individual. We observe that the first component reflects genetic variation between YRI and CEU plus CHP+JPT populations, while the second component separates CEU and CHB+JPT populations. This example illustrates how well EIGENSTRAT is able to capture the genetic difference among individuals of different populations. If we are then interested in assessing association between two groups of individuals we will use a logistic regression model (see section 4.1) adjusted by subject score components (loading values). This incorporates genetic differences among individuals due to ancestry correcting population stratification.

The often abbreviated description of participants and lack of comparison of key characteristics can make evaluation of potential biases and replication of findings quite difficult, as described in [16]. To overcome the difficulties typical of case–control studies, other different designs based on trios or in cohort studies can be adopted. The trio design includes affected case individuals and both parents [83].

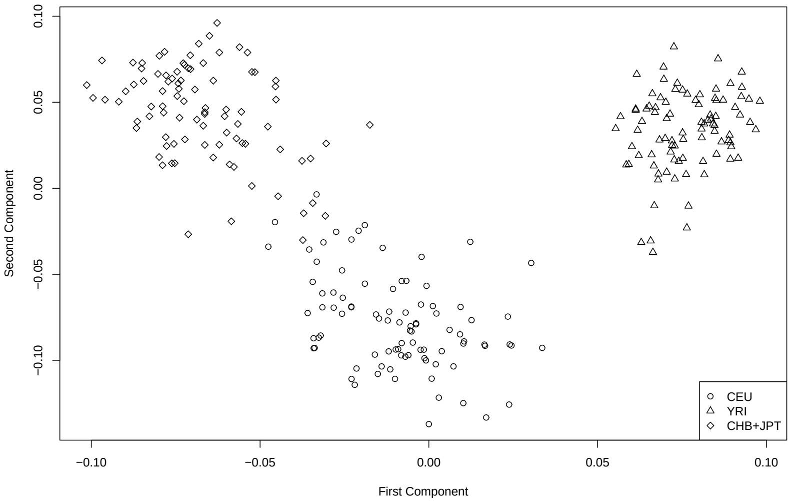

Fig. 3 The top two components (axes) of variation of HapMap samples obtained by using EIGENSTRAT approach. Data corresponds to 270 individuals from European (CEU), Yoruba (YRI) and Chinese plus Japanese (CHB+JPT) populations. This example analyze 9,307 randomly selected SNPs from the entire genome.

Classification of affected status is given only in the offspring and only affected offspring are included, but genotyping is performed in all 3 trio members. The frequency with which an allele is transmitted to an affected offspring from heterozygous parents is then estimated [83]. Under the null hypothesis of no association with disease, the transmission will be 50%. Therefore alleles associated with the disease will be transmitted in excess to the affected case individuals. Cohort studies collect baseline information in a large number of individuals, who are then assessed for the appearance of the disease over time depending on different genetic variants. These studies are normally more expensive, but the results are often more representative than case–control studies. Consequently, GWAS have been recently performed using cohort studies such as the European Prospective Investigation on Cancer (EPIC) [55], Study French Prostate Cancer Study (CeRePP) and MultiEthnic Cohort Study, and Physicians' Health Study [102]. One of the main disadvantages of cohort studies is the large investment required both in time and money. A large cohort needs decades of follow–up time to achieve the number of cases required to detect moderate genetic effects. However, if the cohort is established, genetic association studies can be performed using a nested case–control strategy, where observed cases and a subset of appropriately matched controls are genotyped, rather than the entire cohort [50].

Most of GWAS studies also adopt a multistage design, mainly aimed to reduce the number of false–positive results while minimizing the number of genotyped indi-

viduals and keeping statistical power [41]. In practice, analysis of GWAS is often performed in 2 (or sometimes even more) stages. First, at step 1, the full marker set is genotyped for an initial group having a moderate number of cases and controls. Then, at stage 2, only the most promising markers (e.g. SNPs that have been statistically significant associated with the trait in the first step) are re–genotyped for another group using smaller SNP arrays. The number of SNPs and individuals included in these consecutive steps may vary depending on budget. The question of what significance threshold is appropriate for GWAS studies is somewhat unresolved. In Section 6.4 we outline some approaches that are currently adopted.
[60] pointed out that two–stage designs can be seen as a special case of general group sequential designs used in clinical trials, consisting of 1 single interim analysis (stage 1) and a final analysis (stage 2). They extend the special case of two–stage designs to the general framework of multistage designs, deriving optimal multistage designs with respect to both minimal overall study costs and maximal study power. The authors concluded that economical benefits can be obtained by using more than 2 stages. In particular, they found that using no more than 4 stages is sufficient for practical purposes [60].

6.2 Assessing association in GWAS

After designing the study and obtaining genotype information, we have to assess association between the variants and the disease. The significance of association between a marker and disease is done by analyzing each SNP at time, determined by calculating any of the test statistics described in section 4. For autosomal SNPs, the test statistic can be based on assuming dominant, recessive and additive models. This means that we have to perform 3 times the number of SNPs tests. Considering that currently Affymetrix or Illumina platforms are able to genotype 1 million SNPs, 3 million of tests are needed to be computed. One may be tempted to avoid performing such a high number of tests by calculating the most powerful test (additive model) or Armitage's trend test [33, 80]. However, assuming a model different from the real one leads to loss of power, [33, 72, 80]. Therefore, when the underlying genetic model is unknown, association may be assessed using the max–statistic, which selects the largest test statistic from the dominant, recessive and additive models [36]. This statistic is written as:

$$\chi^2_{MAX} = \max\{\chi^2_{DOM}, \chi^2_{REC}, \chi^2_{ADD}\} \tag{5}$$

A naive approach to determine whether a given marker is associated with the disease using max–statistic is to consider the smallest p–value between dominant, recessive and additive tests [36]. This approach does not maintain the overall type I error rate since it does not account either for multiple testing nor correlation among the three tests as showed by several authors [33, 36, 72, 80]. Hence, the statistical significance of association using max–statistics has to be addressed with other

methods.
[79] consider χ^2_{MAX} test to identify novel risk variants for type 2 diabetes. The authors stated that as the distribution of max–statistic is unknown, a permutation approach can be used to estimate statistical significance. This procedure is extremely expensive computationally for GWAS. For instance, [79] needed to calculate around 11,800 million tests (only in the first stage) for 392,935 markers and 3 inheritance models performing 10,000 permutations to compute p–values for the max–statistic.
[36] derived the asymptotic form for max–statistic that can be used to compute the correct p–value when it is used. The authors also found through simulations studies that the effective number of tests when dominant, recessive and additive models are fitted is 2.2. This number can be used to correct the significance level or p–values as a rule of thumb. Since this value is based on simulations large sample sizes might be required.

6.3 Statistical power calculations

As in many other situations, when a GWAS is performed, one can be interested in estimating the probability that the statistical test used to assess association yields significant results when the null hypothesis is false (e.g. power). In other words, we would like to know the chance that the study will be successful in detecting a true effect. Power calculations are normally performed during the planning stages of a study. However, in some occasions, they can also be used to interpret negative results. Herein, we are describing the fundamentals of statistical power for genetic case–control studies. We will also illustrate how to perform power calculations in the context of GWAS, where other issues such as multiple testing, coverage and staged designs need to be considered.
Power in association studies depends on a number of factors, such as the total number of available samples, the size of the genetic effect, its mode of inheritance, the prevalence of the disease, the ratio of case to control individuals, allelic frequencies, and the extent of linkage disequilibrium between marker and trait loci [13, 65]. It is possible to obtain closed–form expressions to compute statistical power for genetic associations [75]. These formulas can be used to calculate expected power under a range of scenarios. For instance, Table 5 displays the effect of varying sample size, linkage disequilibrium or disease allele frequency on the power to detect association under a case–control setting. Power of case–control studies changes as a function of linkage disequilibrium (section 5.2), as can be seen in Table 5. Values of $D' = 1$ indicate an absence of ancestral recombination between marker and disease loci and thus complete disequilibrium. In contrast, $D' = 0.8$ indicates independence between marker and trait loci. In this case, we can observe that power to detect association is greater when linkage disequilibrium is high, as well as when trait and marker loci have similar allele frequency. As a final conclusion of these examples, we can also observe how power to detect association is strongly related to allele frequency.
Power of two–stage GWAS depends on the same factors as case–control studies.

	Sample size					
	100	200	500	1000	1500	2000
$D' = 1$						
0.1	0.24	0.42	0.79	0.98	1.00	1.00
0.2	0.19	0.33	0.67	0.92	0.99	1.00
0.3	0.15	0.25	0.53	0.82	0.94	0.98
0.4	0.12	0.19	0.40	0.67	0.84	0.92
0.5	0.09	0.14	0.28	0.49	0.66	0.79
$D' = 0.9$						
0.1	0.20	0.36	0.71	0.94	0.99	1.00
0.2	0.16	0.28	0.58	0.86	0.96	0.99
0.3	0.13	0.21	0.45	0.74	0.89	0.96
0.4	0.11	0.16	0.33	0.58	0.76	0.87
0.5	0.09	0.12	0.24	0.42	0.58	0.70
$D' = 0.8$						
0.1	0.17	0.30	0.61	0.89	0.97	0.99
0.2	0.14	0.23	0.49	0.78	0.92	0.97
0.3	0.11	0.18	0.38	0.64	0.81	0.91
0.4	0.09	0.14	0.28	0.49	0.66	0.78
0.5	0.08	0.11	0.20	0.35	0.48	0.60

Table 5 Power calculation for a case–control study varying sample size, linkage disequilibrium (D'), and allele frequency.

Additionally, these studies also depend on how markers are selected for being analyzed in the second stage, how samples are divided between stages 1 and 2, and the proportion of markers tested in stage 2. Power also depends on the significance level we consider for the entire genome α_{genome}. Table 6 shows the obtained power for a hypothetical two–stage GWAS where a case–control study was used including 1,000 cases and 1,000 with a prevalence of the disease equal to 0.1, the allele frequency equal to 0.4, $\alpha_{genome} = 0.,5$, and 300,000 markers (M). The table presents different scenarios by varying the proportion of individuals genotyped at first step, and the proportion of makers tested in second stage. Following the recommendations given by [78], we also present the power for the joint analysis (e.g. joint analysis of data from both stages) that is expected to be more efficient. We observe that the power for two–stage design increases as the number of individuals genotyped in the first stage decreases. For example, two–stage design has 31% power to detect association in the case of analyzing 50% of cases in the first step and genotype 10% of SNPs in the second phase (30,000 markers), while the power is 61% when 20% of cases are analyzed in the first step. We also notice that joint analysis can achieve nearly the same power as the one–stage design in which all samples are genotyped on all markers.

			Power		
%n	%M	OR	step 1	step 2	joint
1.00	1.00	1.40	1.00	0.00	0.74
0.50	0.10	1.40	0.99	0.31	0.74
0.50	0.05	1.40	0.99	0.36	0.74
0.50	0.01	1.40	0.94	0.48	0.74
0.40	0.10	1.40	0.98	0.46	0.74
0.40	0.05	1.40	0.96	0.50	0.74
0.40	0.01	1.40	0.87	0.58	0.71
0.30	0.10	1.40	0.94	0.57	0.73
0.30	0.05	1.40	0.90	0.60	0.71
0.30	0.01	1.40	0.74	0.58	0.63
0.20	0.10	1.40	0.84	0.61	0.67
0.20	0.05	1.40	0.75	0.58	0.62
0.20	0.01	1.40	0.52	0.45	0.47

Table 6 Power calculation for a two–stage GWAS varying the percentage of individuals (%n) to be genotyped at stage 1, the percentage of markers (%M) to be re–genotyped in the second stage. The results are computed assuming a case–control design with 1,000 cases and 1,000 controls, where the prevalence of the disease is 0.1, the allele frequency is equal to 0.4, the α_{genome} is 0.05, and 300,000 markers (M) are analyzed. Table also shows the power of a joint analysis test.

6.4 Statistical level correction for multiple testing

It is well–known that multiple testing problem arises when many hypotheses are tested simultaneously using the same data because some test statistics can be extreme even if no associations exists. Multiple correction procedures are designed to control the set of hypotheses and to prevent false positive conclusions that could be attributed to chance. Correcting for multiple comparisons requires determining a threshold for which p–values are considered as statistically significant. There are several approaches to establish such threshold.

The simplest one is based on controlling the family–wise error rate (FWER), defined as the probability of committing at least one type–I error. FWER can be controlled in a weak sense by using procedures such as Bonferroni or Sidak corrections or in a strong sense by considering that any subset of hypothesis is true [42]. On one hand, Bonferroni correction for multiple testing simply requires a p–value of α/M, α denoting the desired nominal level which normally is set equal to 0.05 and M is the number of genotyped SNPs. On the other hand, Sidak's correction needs a p–value of $1-(1-\alpha)^{(1/M)}$, which is similar to Bonferroni's.

By using these corrections, a GWAS including 1,000,000 SNPs will lead to a Bonferroni significance threshold of 5.0×10^{-8} and 5.12×10^{-8} using Sidak's formula. These assumptions are too conservative, which may lead to a high false–negative rate [28]. The main problem of using a FWER correction in GWAS is that, by applying it, we are assuming that all markers are independent. This hypothesis makes sense in the context of targeted studies, where SNPs are selected to cover a given region. However, genome–wide scans includes SNPs that are in a strong LD, making the independent assumption too restrictive. In these situations, permutation test can be used to estimate the ”effective number of tests” [17]. FWER can then be applied to correct for multiple comparisons using the effective number of tests.

Due to linkage disequilibrium between SNPs, FWER control may be too conservative for GWAS, where the goal is to screen the genome for a further study of very few promising candidates (e.g. replication, fine mapping, functional studies, ...). As a consequence, several authors proposed other methods for false positive rate control, such as false discovery rate (FDR) [9], posterior error rates [54, 95], or permutation testing. [27] pointed out that FDR is not appropriate for association studies and that other methods such as permutation approach based on minimum p–value should be employed.

Permutation tests are based on computing significance levels by estimating the underlying null distribution when theoretical results do not hold. The unknown null distribution is estimated by simulating data in a given manner. Before addressing the problem of how to compute the corrected p–value in a GWAS using permutation procedure, let us start by illustrating how to use it in a single analysis where association between trait and a given SNP is assessed. For the benefit of simplicity, let us assume that our trait is dichotomous (case–control), and that we are interested in comparing the proportion of carriers of a rare allele (e.g. assume a dominant model) between the two groups of individuals. As mentioned in section 4.1, the null hypothesis of no association can be addressed by using the χ^2 test. In general, the test is based on comparing the proportion of cases who carry the susceptibility allele with the proportion of controls who do not. In the case of having association, we will observe a large value for χ^2 statistic. That is, far away from the null hypothesis of no association where χ^2 is equal to 0. Following theoretical results, the significance of this observed statistic is computed by using a χ^2 distribution with 1 degree of freedom. If this distribution cannot be assumed by any reason, the significance has to be computed by estimating the distribution of the test statistic under the null hypothesis. Under these circumstances, permutation can be used as following. Case–control labels are exchanged among individuals by keeping genotypes fixed and test statistic is then computed. It is expected that after random assignment of case and control status the test statistic would be close to the null hypothesis (e.g. near 0). This procedure is repeated B times and the significance level is the proportion of replicate statistics exceeding the observed statistic [99]. If there is no association, we expect to have the χ^2 value as one of those obtained by permutating data and a large p–value, meaning that the null hypothesis of no association cannot be rejected. On the other hand, if the variant is related to the disease, the number of χ^2 values larger that the obtained by analyzing the observed data will be low. In this case the permuted p–value will be lower that the nominal level, rejecting the null hypothesis and concluding that there is association between the marker and the disease.

This permutation procedure can be extended in GWAS to compute a corrected p–value. By permuting data, we are able to capture the correlation among markers. Table 7 shows the main steps we have to perform for obtaining the corrected level of significance by permutation testing. As in the single case, we first randomly assign case and control labels to individuals by keeping genotypes fixed. Then we compute p–values of association for each SNP using a statistical test (i.e. dominant, log–additive, max–statistic…). For each permutation, we retain the minimum p–value obtained among all SNPs analyzed. The corrected significance level is computed as the 5% quantile point of the empirical distribution for the minimum p–values. Alternatively, one can assume that the minimum p–value follows a Beta distribution with parameters $(1, n_E)$, being n_E the number of effective tests [29]. One can fit the Beta distribution to the minimum p–value of the permutation replicates and then estimate the 5% quantile point form this theoretical distribution. Using the approximation to a Beta distribution, a moderate number of permutations (e.g. 10,000) can be enough to correct estimate the corrected nominal level. This approach has also another important advantage. Let us assume that we are in the context of performing a GWAS where several diseases are analyzed by using a shared group of controls like in the Wellcome Trust Case Control Consortium study [21]. In this case, for each analysis, we should have to perform a permutation analysis for each disease. However, [29] pointed out that this approach needs to be done once only because the distribution of p–values under the null hypothesis is the same in all studies. Using Beta approximation, we can also test whether the minimum p–value is consistent with an effective number of independent tests, by testing whether the first parameter is 1 [29].

```
Repeat B times:
  1. Randomly assign traits among samples, while keeping the genotype fixed
  2. Compute the p–value for each SNP by using any selected test (e.g. log–additive,
max–statistic, ...)
  3. Keep the minimum p–value
The corrected significance level is estimated by selecting the 5% quantile of the replicate
minimum p–values
  or
Estimate $Beta(1, n_E)$ by using the replicated minimum p–values and computing its 5% quantile
```

Table 7 Steps to correct nominal level by using permutation approach in GWAS

To illustrate how the permutation approach works, we used 9,307 SNPs from the HapMap [1] project randomly selected from the entire genome. We compared genotype frequencies between European (CEU) and Yoruba (YRI) populations. The corrected nominal level by using Bonferroni correction is equal to 5.37×10^{-6}. However using the distribution of the minimum p–value we obtained a corrected p–value of 2.32×10^{-5} (Figure 4). The p–value from the empirical distribution is 2.22×10^{-5}. This shows an excellent agreement between the empirical and theoretical distributions of minimum p–values, as expected. Notice that, using this permutation approach, we obtained a not so stringent significance level leading to an increase in power. This procedure is implemented in SNPassoc software (see section 8.1.2).

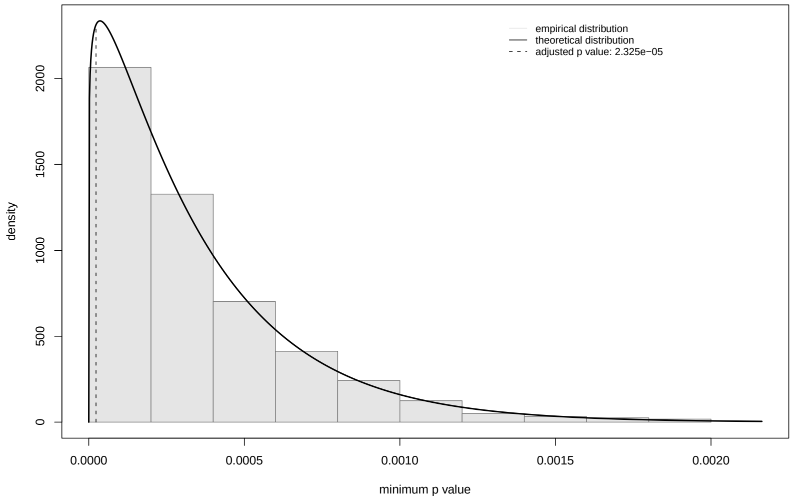

Fig. 4 Empirical (histogram) and Beta distributions (straight line) of minimum p–values obtained from the permutation procedure applied to 9307 randomly selected SNPs from HapMap project. The p–values are obtained by assessing differences between CEU and YRI populations assuming an additive model. The results are based on 10,000 permutations

7 Gene–gene and gene–environment interactions

The analysis of interactions involves the assessment of an effect modification: the risk associated with a genotype varies according to a third variable, which can be an environmental exposure or other genetic factors. The assessment is a comparison of the effects among subgroups of the data. Usually regression models are used to test for interactions. In these models, next to main effects (gene and environment), an additional term corresponding to the product of the main effects is added.

The coefficient for this product term, if zero, is interpreted as no interaction: gene and environment are independent and act additively in the model scale. If the coefficient is different from zero, it is interpreted as a difference in the effect for the combination of gene and environment respect to the expected under independence. A positive coefficient means synergism: the combined effect is larger than expected. A negative coefficient means antagonism or that the combined effect is not larger than the sum of the main effects. Usually, when this analysis is performed for a case–control study, logistic regression is used and the interactions must be interpreted in a multiplicative scale. A negative interaction coefficient may mean that the combined effect is additive and not multiplicative, as the logistic regression imposes.

Power to detect interactions is smaller than power to detect main effects because it is effectively a comparison between two (or more) risk estimates. Power also de-

pends on the degrees of freedom used to test the interaction. For this reason, it is usually desirable to define a binary environmental factor and collapse the genotypes into an inheritance model with only one degree of freedom (dominant, recessive or log–additive).
One strategy to increase the power to detect gene–environment or gene–gene interactions is to use a case–only analysis [62]. The association between gene and environment when the sample is restricted to cases only estimates the interaction under the condition that there is no association between the gene and the environment in controls and it is important to recognize this requirement [4]. Recent developments in this methodology allow using the advantages of the case–only design when the independence assumption is met and use the complete dataset otherwise in a weighted empirical–bayes approach [57].

8 Bioinformatics tools and databases for genetic association studies

In previous sections we have described the whole process of a population–based genetic association analysis, from initial quality control to correction for multiple testing in GWAS. As it can be easily seen, the illustrated statistical analysis is complex and composed of many different parts. Thus, it is difficult and time–consuming to perform a complete association analysis using only general–purpose statistical suites (e.g. R, SAS). Fortunately, there are plenty of bioinformatics tools that will allow researchers to successfully complete the whole analysis without the need of a deep knowledge of computing or bioinformatics. Before using these tools, however, we need to be careful about choosing the right type of analysis for our data, so that we avoid any potential errors and take the most advantage of our data.
In the next sections we will briefly list some of the most used software tools. For a better clarity, they have been classified according to their main functionalities.

8.1 Genetic association suites

This pieces of software deal with most of the analytical steps explained. They are suitable tools for quality control, single–SNP association or haplotype analysis, and in the case of PLINK it can handle GWAS datasets.

8.1.1 SNPStats

SNPStats [81] is an easy and ready–to–use web tool for association analysis, specially designed for small to mid–size studies (i.e. up to a thousand of SNPs approximately). It can perform quality control procedures (genotyping rate, HWE, allele

and genotype frequencies), analysis of association with a response variable based on linear or logistic regression, accepts multiple inheritance models (e.g. co–dominant, dominant, recessive, over–dominant and log–additive) and analysis of gene–gene or gene–environment interactions. If multiple SNPs are selected, SNPStats offers the possibility to compute LD statistics between SNPs, haplotype frequency estimation and association with the response and analysis of haplotype–environment interactions.
Website: http://bioinfo.iconcologia.net/snpstats

8.1.2 SNPassoc

The R package SNPassoc [35] is useful to carry out most common parts of a GWAS analysis in an efficient manner. These analyses include descriptive statistics and exploratory analysis of missing values, calculation of Hardy–Weinberg equilibrium, analysis of association based on generalized linear models (either for quantitative or binary traits), and analysis of multiple SNPs (haplotype and epistasis analysis, p–value correction for multiple testing). Permutation tests and other related tests (sum statistic and truncated product) are also implemented. Compared to other R packages with similar purposes, SNPassoc offers a greater usability.
Website: http://www.creal.cat/jrgonzalez/software.htm

8.1.3 PLINK

PLINK [66] is a free, open–source GWAS toolset, designed to perform a complete range of basic, large–scale analyses in a computationally efficient manner. The focus of PLINK is purely on analysis of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data). PLINK is designed as a command–line tool, but through its recent integration with a JAVA graphical interface (called gPLINK) and Haploview, there is some support for the subsequent visualization, annotation and storage of results.
Website: http://pngu.mgh.harvard.edu/∼purcell/plink/

8.1.4 GAP

GAP (Genetic Analysis Package) [104] is implemented as a package for R. It has functions for Hardy–Weinberg equilibrium tests, measures of linkage disequilibrium between SNPs or multiallelic markers and haplotype analysis. It is also useful for two–stage case–control power calculations.
Website: http://www.mrc-epid.cam.ac.uk/Personal/jinghua.zhao/r-progs.htm

8.2 Haplotype–only software

Haplotype–related analysis is an important part of association studies. Thus, there are some tools focused specifically on this area. Regarding their functionality, these tools can be mainly divided into those which only infer haplotype frequencies and those which perform both the inference and the association with a trait.

8.2.1 Haploview

Haploview [7] is graphical tool nicely designed to simplify and expedite the process of haplotype analysis by providing a common interface to several tasks relating to such analyses. Haploview currently supports a wide range of haplotype functionalities such as: LD and haplotype block analysis, haplotype population frequency estimation, single SNP and haplotype association tests, permutation testing for association significance, implementation of Paul de Bakker's Tagger tag SNP selection algorithm, automatic download of phased genotype data from HapMap and visualization and plotting of PLINK genome–wide association results including advanced filtering options. Haploview is fully compatible with data dumps from the HapMap project and the Perlegen Genotype Browser. It can analyze thousands of SNPs (tens of thousands in command line mode) in thousands of individuals.
Website: http://www.broad.mit.edu/mpg/haploview/

8.2.2 PHASE/fastPHASE

PHASE [87, 85, 86] and fastPHASE [74] are command–line pieces of software used for haplotype reconstruction, as well as estimation of missing genotypes from population data. They do not compute association with a phenotype. Although fastPHASE can handle larger datasets than its previous version PHASE (eg hundreds of thousands of markers in thousands of individuals), it does not provide estimates of recombination rates (while PHASE does). Experiments suggest that fastPHASE haplotype estimates are slightly less accurate than from PHASE, but missing genotype estimates appear to be similar or even slightly better than PHASE.
Website: http://stephenslab.uchicago.edu/software.html

8.2.3 Haplo.stats

Haplo.stats [71, 72, 73] is a suite of R routines for the analysis of indirectly measured haplotypes. The statistical methods implemented assume that all subjects are unrelated and that haplotypes are ambiguous (due to unknown linkage phase of the genetic markers). The genetic markers are assumed to be codominant (i.e. one–to–one correspondence between their genotypes and their phenotypes). Some tools, such as SNPStats (see section **??**) and SNPassoc (see section 8.2) use Haplo.stats as

the underlying software to compute all haplotype related computations.
Website: http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm

8.2.4 Thesias

The aim of THESIAS [92, 94, 93] is to perform haplotype–based association analysis in unrelated individuals. The program is based on the maximum likelihood model and is linked to the Stochastic EM algorithm. THESIAS allows the simultaneous estimation of haplotype frequencies and of their associated effects on the phenotype of interest. Quantitative, qualitative, categorical and, more interestingly, survival outcomes can be studied. Covariate–adjusted haplotype effects as well as haplotype–environment interactions can be investigated. THESIAS began as a command–line tool which was not too user–friendly, but in the latest version its creators have added a JAVA interface which has improved the usability of the program, although it is still a bit rigid at some points.
Website: http://ecgene.net/genecanvas/uploads/THESIAS3.1/Documentation3.1.htm

8.3 Web databases

The dramatically increasing amount of genomic information generated in the last few years has made essential the development of information systems which provide easy procedures of storage and retrieval of data. Therefore, public repositories of genetic data have been created and are maintained on a daily basis due to the effort of large consortiums.

8.3.1 dbSNP

In collaboration with the National Human Genome Research Institute, The National Center for Biotechnology Information has established the dbSNP [76] database to serve as a central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms. Once discovered, these polymorphisms can be used by additional laboratories, using the sequence information around the polymorphism and the specific experimental conditions. Note that dbSNP takes the looser 'variation' definition for SNPs, so there is no requirement or assumption about minimum allele frequency. Data in dbSNP can be integrated with other NCBI genomic data. As with all NCBI projects, data in dbSNP is freely available to the scientific community and made available in a variety of forms.
Website: http://www.ncbi.nlm.nih.gov/projects/SNP/

8.3.2 Hapmap

The International HapMap Project [1, 91] is a multi–country effort to identify and catalog genetic similarities and differences in human beings. It describes what these variants are, where they occur in our DNA, and how they are distributed among people within populations and among populations in different parts of the world. Using the information in the HapMap, researchers will be able to find genes that affect health, disease, and individual responses to medications and environmental factors. In the initial phase of the project, genetic data are being gathered from four populations with African, Asian, and European ancestry. Ongoing interactions with members of these populations are addressing potential ethical issues and providing valuable experience in conducting research with identified populations.
Website: http://www.hapmap.org/

8.3.3 Genome Variation Server (GVS)

The Genome Variation Server is a local database hosted by the SeattleSNPs Program for Genomic Applications. The objective of this database is to provide a simple tool for rapid access to human genotype data found in dbSNP. The database includes a suite of analysis tools such as linkage disequilibrium plots, tag SNPs and more. In addition you can upload our own data and use the GVS analysis and visualization tools.
Website: http://gvs.gs.washington.edu/GVS/

8.4 Statistical power calculation

Statistical power calculation is an important issue in association studies, specially for GWAS. Before we proceed with the experiments, we need to ensure that with our sample size we will be able to detect changes of a specific size in terms of the minimum allele frequency we expect to have in our SNPs of interest. Ignoring this information may cause our failure in detecting existing genetic associations due to the inadequate sample size of our study.

8.4.1 QUANTO

QUANTO computes sample size or power for association studies of genes, gene–environment interaction, or gene–gene interaction. Available study designs include the matched case–control, case–sibling, case–parent, and case–only designs. It is a stand–alone 32–bit Windows application. Its graphical user interface allows the user to easily change the model and view the results without having to edit an input file

and rerun the program for every model.
Website: http://hydra.usc.edu/gxe/

8.4.2 Genetic Power Calculator

Designed by the same authors of PLINK, GPC [65] is a website tool that performs power calculations for the design of linkage and association genetic mapping studies of complex traits.
Website: http://pngu.mgh.harvard.edu/∼purcell/gpc/

8.4.3 CaTS

CaTS [78] is a simple and useful multi–platform interface for carrying out power calculations for large genetic association studies, including two stage genome–wide association studies.
Website: http://www.sph.umich.edu/csg/abecasis/cats/index.html.

References

1. The international hapmap project. *Nature*, 426(6968):789–96, 2003.
2. Brlmm: an improved genotype calling method for the genechip human mapping 500k array set. Technical report, Affymetrix, 2006.
3. J. Akey, L. Jin, and M. Xiong. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet*, 9(4):291–300, 2001.
4. P. S. Albert, D. Ratnasinghe, J. Tangrea, and S. Wacholder. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol*, 154(8):687–93, 2001.
5. P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–86, 1955.
6. J. E. Bailey-Wilson, B. Sorant, A. J. Sorant, C. M. Paul, and R. C. Elston. Model-free association analysis of a rare disease. *Genet Epidemiol*, 12(6):571–5, 1995.
7. J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–5, 2005.
8. G. W. Beecham, E. R. Martin, Y. J. Li, M. A. Slifer, J. R. Gilbert, J. L. Haines, and M. A. Pericak-Vance. Genome-wide association study implicates a chromosome 12 risk locus for late-onset alzheimer disease. *Am J Hum Genet*, 84(1):35–43, 2009.
9. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J R Stat Soc B*, 57:289–00, 1995.
10. S. T. Bennett, A. M. Lucassen, S. C. Gough, E. E. Powell, D. E. Undlien, L. E. Pritchard, M. E. Merriman, Y. Kawaguchi, M. J. Dronsfield, F. Pociot, and et al. Susceptibility to human type 1 diabetes at iddm2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nat Genet*, 9(3):284–92, 1995.
11. E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock,

R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, and Nobl. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

12. P. Broderick, L. Carvajal-Carmona, A. M. Pittman, E. Webb, K. Howarth, A. Rowan, S. Lubbe, S. Spain, K. Sullivan, S. Fielding, E. Jaeger, J. Vijayakrishnan, Z. Kemp, M. Gorman, I. Chandler, E. Papaemmanuil, S. Penegar, W. Wood, G. Sellick, M. Qureshi, A. Teixeira, E. Domingo, E. Barclay, L. Martin, O. Sieber, D. Kerr, R. Gray, J. Peto, J. B. Cazier, I. Tomlinson, and R. S. Houlston. A genome-wide association study shows that common alleles of smad7 influence colorectal cancer risk. *Nat Genet*, 39(11):1315–7, 2007.
13. J. Cai and D. Zeng. Sample size/power calculation for case-cohort studies. *Biometrics*, 60(4):1015–24, 2004.
14. B. Carvalho, H. Bengtsson, T. P. Speed, and R. A. Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide snp array data. *Biostatistics*, 8(2):485–99, 2007.
15. G. Celeux and J. Diebolt. The sem algorithm: a probabilistic teacher derived from the em algorithm for the mixture problem. *Computer Statistics Quarterly*, 2:73–82, 1985.
16. S. J. Chanock, T. Manolio, M. Boehnke, E. Boerwinkle, D. J. Hunter, G. Thomas, J. N. Hirschhorn, G. Abecasis, D. Altshuler, J. E. Bailey-Wilson, L. D. Brooks, L. R. Cardon, M. Daly, P. Donnelly, Jr. Fraumeni, J. F., N. B. Freimer, D. S. Gerhard, C. Gunter, A. E. Guttmacher, M. S. Guyer, E. L. Harris, J. Hoh, R. Hoover, C. A. Kong, K. R. Merikangas, C. C. Morton, L. J. Palmer, E. G. Phimister, J. P. Rice, J. Roberts, C. Rotimi, M. A. Tucker, K. J. Vogan, S. Wacholder, E. M. Wijsman, D. M. Winn, and F. S. Collins. Replicating genotype-phenotype associations. *Nature*, 447(7145):655–60, 2007.
17. G. A. Churchill and R. W. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–71, 1994.
18. A. G. Clark. Inference of haplotypes from pcr-amplified samples of diploid populations. *Mol Biol Evol*, 7(2):111–22, 1990.
19. D. Clayton and P. M. McKeigue. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*, 358:1356–1360, 2001.
20. D. G. Clayton, N. M. Walker, D. J. Smyth, R. Pask, J. D. Cooper, L. M. Maier, L. J. Smink, A. C. Lam, N. R. Ovington, H. E. Stevens, S. Nutland, J. M. Howson, M. Faham, M. Moorhead, H. B. Jones, M. Falkowski, P. Hardenbol, T. D. Willis, and J. A. Todd. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet*, 37(11):1243–6, 2005.
21. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 controls. *Nature*, 447:661–678, 2007.
22. A. Cox, A. M. Dunning, M. Garcia-Closas, S. Balasubramanian, M. W. Reed, K. A. Pooley, S. Scollen, C. Baynes, B. A. Ponder, S. Chanock, J. Lissowska, L. Brinton, B. Peplonska, M. C. Southey, J. L. Hopper, M. R. McCredie, G. G. Giles, O. Fletcher, N. Johnson, I. dos Santos Silva, L. Gibson, S. E. Bojesen, B. G. Nordestgaard, C. K. Axelsson, D. Torres, U. Hamann, C. Justenhoven, H. Brauch, J. Chang-Claude, S. Kropp, A. Risch, S. Wang-Gohrke, P. Schurmann, N. Bogdanova, T. Dork, R. Fagerholm, K. Aaltonen, C. Blomqvist, H. Nevanlinna, S. Seal, A. Renwick, M. R. Stratton, N. Rahman, S. Sangrajrang, D. Hughes, F. Odefrey, P. Brennan, A. B. Spurdle, G. Chenevix-Trench, J. Beesley, A. Mannermaa, J. Hartikainen, V. Kataja, V. M. Kosma, F. J. Couch, J. E. Olson, and E. L. Goode. A common coding variant in casp8 is associated with breast cancer risk. *Nat Genet*, 39(3):352–8, 2007.
23. D. J. Cutler, M. E. Zwick, M. M. Carrasquillo, C. T. Yohn, K. P. Tobin, C. Kashuk, D. J. Mathews, N. A. Shah, E. E. Eichler, J. A. Warrington, and A. Chakravarti. High-throughput variation detection and genotyping using microarrays. *Genome Res*, 11(11):1913–25, 2001.
24. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em-algorithm. *JRSS*, 39:1–38, 1977.

25. B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.
26. X. Di, H. Matsuzaki, T. A. Webster, E. Hubbell, G. Liu, S. Dong, D. Bartell, J. Huang, R. Chiles, G. Yang, M. M. Shen, D. Kulp, G. C. Kennedy, R. Mei, K. W. Jones, and S. Cawley. Dynamic model based algorithms for screening and genotyping over 100 k snps on oligonucleotide microarrays. *Bioinformatics*, 21(9):1958–63, 2005.
27. F. Dudbridge, A. Gusnanto, and B. P. Koeleman. Detecting multiple associations in genome-wide studies. *Hum Genomics*, 2(5):310–7, 2006.
28. F. Dudbridge and A. Gustano. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*, 32:227–234, 2008.
29. F. Dudbridge and B. P. Koeleman. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet*, 75(3):424–35, 2004.
30. D. F. Easton and R. A. Eeles. Genome-wide association studies in cancer. *Hum Mol Genet*, 17(R2):R109–15, 2008.
31. L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–7, 1995.
32. W. D. Foulkes. Inherited susceptibility to common cancers. *N Engl J Med*, 359(20):2143–53, 2008.
33. B. Freidlin, G. Zheng, Z. Li, and J. L. Gastwirth. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered*, 53(3):146–52, 2002.
34. M. Garcia-Closas, N. Malats, D. Silverman, M. Dosemeci, M. Kogevinas, D. W. Hein, A. Tardon, C. Serra, A. Carrato, R. Garcia-Closas, J. Lloreta, G. Castano-Vinyals, M. Yeager, R. Welch, S. Chanock, N. Chatterjee, S. Wacholder, C. Samanic, M. Tora, F. Fernandez, F. X. Real, and N. Rothman. Nat2 slow acetylation, gstm1 null genotype, and risk of bladder cancer: results from the spanish bladder cancer study and meta-analyses. *Lancet*, 366(9486):649–59, 2005.
35. J. R. Gonzalez, L. Armengol, X. Sole, E. Guino, J. M. Mercader, X. Estivill, and V. Moreno. Snpassoc: an r package to perform whole genome association studies. *Bioinformatics*, 23(5):644–5, 2007.
36. J. R. Gonzalez, J. L. Carrasco, F. Dudbridge, L. Armengol, X. Estivill, and V. Moreno. Maximizing association statistics over genetic models. *Genet Epidemiol*, 32(3):246–54, 2008.
37. P. Gorroochurn, G. A. Heiman, S. E. Hodge, and D. A. Greenberg. Centralizing the non-central chi-square: A new method to correct for population stratification in genetic case-control association studies. *Genet Epidemiol*, 30(4):277–89, 2006.
38. S. W. Guo and E. A. Thompson. Performing the exact test of hardy-weinberg proportion for multiple alleles. *Biometrics*, 48(2):361–72, 1992.
39. C. A. Haiman, L. Le Marchand, J. Yamamato, D. O. Stram, X. Sheng, L. N. Kolonel, A. H. Wu, D. Reich, and B. E. Henderson. A common genetic risk factor for colorectal and prostate cancer. *Nat Genet*, 39(8):954–6, 2007.
40. C. A. Haiman, N. Patterson, M. L. Freedman, S. R. Myers, M. C. Pike, A. Waliszewska, J. Neubauer, A. Tandon, C. Schirmer, G. J. McDonald, S. C. Greenway, D. O. Stram, L. Le Marchand, L. N. Kolonel, M. Frasco, D. Wong, L. C. Pooler, K. Ardlie, I. Oakley-Girvan, A. S. Whittemore, K. A. Cooney, E. M. John, S. A. Ingles, D. Altshuler, B. E. Henderson, and D. Reich. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet*, 39(5):638–44, 2007.
41. J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108, 2005.
42. J. Hoh and J Ott. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet*, 4:701–9, 2003.
43. H. Hong, Z. Su, W. Ge, L. Shi, R. Perkins, H. Fang, J. Xu, J. J. Chen, T. Han, J. Kaput, J. C. Fuscoe, and W. Tong. Assessing batch effects of genotype calling algorithm brlmm for the affymetrix genechip human mapping 500 k array set using 270 hapmap samples. *BMC Bioinformatics*, 9 Suppl 9:S17, 2008.

44. D. J. Hunter, P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager, S. E. Hankinson, S. Wacholder, Z. Wang, R. Welch, A. Hutchinson, J. Wang, K. Yu, N. Chatterjee, N. Orr, W. C. Willett, G. A. Colditz, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, R. B. Hayes, M. Tucker, D. S. Gerhard, Jr. Fraumeni, J. F., R. N. Hoover, G. Thomas, and S. J. Chanock. A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*, 39(7):870–4, 2007.
45. R. Iniesta and V. Moreno. Assessment of genetic association using haplotypes inferred with uncertainty via markov chain monte carlo. In A. Keller, S. Heinrich, and H. Niederreiter, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2006*, pages 529–35. Springer-Verlag, 2008.
46. R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, 2003.
47. E. Jaeger, E. Webb, K. Howarth, L. Carvajal-Carmona, A. Rowan, P. Broderick, A. Walther, S. Spain, A. Pittman, Z. Kemp, K. Sullivan, K. Heinimann, S. Lubbe, E. Domingo, E. Barclay, L. Martin, M. Gorman, I. Chandler, J. Vijayakrishnan, W. Wood, E. Papaemmanuil, S. Penegar, M. Qureshi, S. Farrington, A. Tenesa, J. B. Cazier, D. Kerr, R. Gray, J. Peto, M. Dunlop, H. Campbell, H. Thomas, R. Houlston, and I. Tomlinson. Common genetic variants at the crac1 (hmps) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet*, 40(1):26–8, 2008.
48. J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, and D. Altshuler. Integrated genotype calling and association analysis of snps, common copy number polymorphisms and rare cnvs. *Nat Genet*, 40(10):1253–60, 2008.
49. S. Landi, F. Gemignani, V. Moreno, L. Gioia-Patricola, A. Chabrier, E. Guino, M. Navarro, J. de Oca, G. Capella, and F. Canzian. A comprehensive analysis of phase i and phase ii metabolism gene polymorphisms and risk of colorectal cancer. *Pharmacogenet Genomics*, 15(8):535–46, 2005.
50. B. Langholz, N Rothman, S Wacholder, and D. Thomas. Cohort studies for characterizing measured genes. *Monogr Natl Cancer Inst*, 26:39–42, 1999.
51. P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki. Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from sweden, denmark, and finland. *N Engl J Med*, 343(2):78–85, 2000.
52. S. Lin, B. Carvalho, D. J. Cutler, D. E. Arking, A. Chakravarti, and R. A. Irizarry. Validation and extension of an empirical bayes method for snp calling on affymetrix microarrays. *Genome Biol*, 9(4):R63, 2008.
53. W. M. Liu, X. Di, G. Yang, H. Matsuzaki, J. Huang, R. Mei, T. B. Ryder, T. A. Webster, S. Dong, G. Liu, K. W. Jones, G. C. Kennedy, and D. Kulp. Algorithms for large-scale genotyping microarrays. *Bioinformatics*, 19(18):2397–403, 2003.
54. K. F. Manly, D. Nettleton, and J. T. Hwang. Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res*, 14(6):997–1001, 2004.
55. J. D. McKay, R. J. Hung, V. Gaborieau, P. Boffetta, A. Chabrier, G. Byrnes, D. Zaridze, A. Mukeria, N. Szeszenia-Dabrowska, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, V. Bencko, L. Foretova, V. Janout, J. McLaughlin, F. Shepherd, A. Montpetit, S. Narod, H. E. Krokan, F. Skorpen, M. B. Elvestad, L. Vatten, I. Njolstad, T. Axelsson, C. Chen, G. Goodman, M. Barnett, M. M. Loomis, J. Lubinski, J. Matyjasik, M. Lener, D. Oszutowska, J. Field, T. Liloglou, G. Xinarianos, A. Cassidy, P. Vineis, F. Clavel-Chapelon, D. Palli, R. Tumino, V. Krogh, S. Panico, C. A. Gonzalez, J. Ramon Quiros, C. Martinez, C. Navarro, E. Ardanaz, N. Larranaga, K. T. Kham, T. Key, H. B. Bueno-de Mesquita, P. H. Peeters, A. Trichopoulou, J. Linseisen, H. Boeing, G. Hallmans, K. Overvad, A. Tjonneland, M. Kumle, E. Riboli, D. Zelenika, A. Boland, M. Delepine, M. Foglio, D. Lechner, F. Matsuda, H. Blanche, I. Gut, S. Heath, M. Lathrop, and P. Brennan. Lung cancer susceptibility locus at 5p15.33. *Nat Genet*, 40(12):1404–6, 2008.

56. V. Moreno, F. Gemignani, S. Landi, L. Gioia-Patricola, A. Chabrier, I. Blanco, S. Gonzalez, E. Guino, G. Capella, and F. Canzian. Polymorphisms in genes of nucleotide and base excision repair: risk and prognosis of colorectal cancer. *Clin Cancer Res*, 12(7 Pt 1):2101–8, 2006.
57. B. Mukherjee and N. Chatterjee. Exploiting gene-environment independence for analysis of case-control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64(3):685–94, 2008.
58. T. Niu, Z. S. Qin, X. Xu, and J. S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet*, 70(1):157–69, 2002.
59. M. C. O'Donovan, N. Norton, H. Williams, T. Peirce, V. Moskvina, I. Nikolov, M. Hamshere, L. Carroll, L. Georgieva, S. Dwyer, P. Holmans, J. L. Marchini, C. C. Spencer, B. Howie, H. T. Leung, I. Giegling, A. M. Hartmann, H. J. Moller, D. W. Morris, Y. Shi, G. Feng, P. Hoffmann, P. Propping, C. Vasilescu, W. Maier, M. Rietschel, S. Zammit, J. Schumacher, E. M. Quinn, T. G. Schulze, N. Iwata, M. Ikeda, A. Darvasi, S. Shifman, L. He, J. Duan, A. R. Sanders, D. F. Levinson, R. Adolfsson, U. Osby, L. Terenius, E. G. Jonsson, S. Cichon, M. M. Nothen, M. Gill, A. P. Corvin, D. Rujescu, P. V. Gejman, G. Kirov, N. Craddock, N. M. Williams, and M. J. Owen. Analysis of 10 independent samples provides evidence for association between schizophrenia and a snp flanking fibroblast growth factor receptor 2. *Mol Psychiatry*, 14(1):30–6, 2009.
60. Roman Pahl, Helmut Schafer, and Hans-Helge Muller. Optimal multistage designs–a general framework for efficient genome-wide association studies. *Biostatistics*, 2008, (In press).
61. N. Pankratz, J. B. Wilk, J. C. Latourelle, A. L. DeStefano, C. Halter, E. W. Pugh, K. F. Doheny, J. F. Gusella, W. C. Nichols, T. Foroud, and R. H. Myers. Genomewide association study for susceptibility genes contributing to familial parkinson disease. *Hum Genet*, 124(6):593–605, 2009.
62. W. W. Piegorsch, C. R. Weinberg, and J. A. Taylor. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*, 13(2):153–62, 1994.
63. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, 2006.
64. J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59, 2000.
65. S. Purcell, S. S. Cherny, and P. C. Sham. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149–50, 2003.
66. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75, 2007.
67. Z. S. Qin, T. Niu, and J. S. Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet*, 71(5):1242–7, 2002.
68. N. Rabbee and T. P. Speed. A genotype calling algorithm for affymetrix snp arrays. *Bioinformatics*, 22(1):7–12, 2006.
69. R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–54, 2006.
70. C. Sabatti, S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen, S. Gabriel, P. Elliot, M. I. McCarthy, M. J.

Daly, M. R. Jarvelin, N. B. Freimer, and L. Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*, 41(1):35–46, 2009.

71. D. J. Schaid. Evaluating associations of haplotypes with traits. *Genet Epidemiol*, 27(4):348–64, 2004.
72. D. J. Schaid, S. K. McDonnell, S. J. Hebbring, J. M. Cunningham, and S. N. Thibodeau. Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet*, 76(5):780–93, 2005.
73. 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(2):425–34, 2002.
74. P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78(4):629–44, 2006.
75. N. J. Schork. Power calculations for genetic association studies using estimated probability distributions. *Am J Hum Genet*, 70(6):1480–9, 2002.
76. S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–11, 2001.
77. A. Shlien, U. Tabori, C. R. Marshall, M. Pienkowska, L. Feuk, A. Novokmet, S. Nanda, H. Druker, S. W. Scherer, and D. Malkin. Excessive genomic dna copy number variation in the li-fraumeni cancer predisposition syndrome. *Proc Natl Acad Sci U S A*, 105(32):11264–9, 2008.
78. A. D. Skol, L. J. Scott, G. R. Abecasis, and M. Boehnke. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*, 38(2):209–13, 2006.
79. R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D. Meyre, C. Polychronakos, and P. Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–5, 2007.
80. S. L. Slager and D. J. Schaid. Case-control studies of genetic markers: power and sample size approximations for armitage's test for trend. *Hum Hered*, 52(3):149–53, 2001.
81. X. Sole, E. Guino, J. Valls, R. Iniesta, and V. Moreno. Snpstats: a web tool for the analysis of association studies. *Bioinformatics*, 22(15):1928–9, 2006.
82. X. Sole, P. Hernandez, M. L. de Heredia, L. Armengol, B. Rodriguez-Santiago, L. Gomez, C. A. Maxwell, F. Aguilo, E. Condom, J. Abril, L. Perez-Jurado, X. Estivill, V. Nunes, G. Capella, S. B. Gruber, V. Moreno, and M. A. Pujana. Genetic and genomic analysis modeling of germline c-myc overexpression and cancer susceptibility. *BMC Genomics*, 9:12, 2008.
83. R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, 52(3):506–16, 1993.
84. S. E. Spruill, J. Lu, S. Hardy, and B. Weir. Assessing sources of variability in microarray gene expression data. *Biotechniques*, 33(4):916–20, 922–3, 2002.
85. M. Stephens and P. Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, 73(5):1162–9, 2003.
86. M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*, 76(3):449–62, 2005.
87. M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–89, 2001.
88. J. Sun, S. L. Zheng, F. Wiklund, S. D. Isaacs, L. D. Purcell, Z. Gao, F. C. Hsu, S. T. Kim, W. Liu, Y. Zhu, P. Stattin, H. O. Adami, K. E. Wiley, L. Dimitrov, T. Li, A. R. Turner, T. S. Adams, J. Adolfsson, J. E. Johansson, J. Lowey, B. J. Trock, A. W. Partin, P. C. Walsh, J. M. Trent, D. Duggan, J. Carpten, B. L. Chang, H. Gronberg, W. B. Isaacs, and J. Xu. Evidence for two independent prostate cancer risk-associated loci in the hnf1b gene at 17q12. *Nat Genet*, 40(10):1153–5, 2008.

89. M. W. Tanck, A. H. Klerkx, J. W. Jukema, P. De Knijff, J. J. Kastelein, and A. H. Zwinderman. Estimation of multilocus haplotype effects using weighted penalised log-likelihood: analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. *Ann Hum Genet*, 67(Pt 2):175–84, 2003.

90. A. Tenesa, S. M. Farrington, J. G. Prendergast, M. E. Porteous, M. Walker, N. Haq, R. A. Barnetson, E. Theodoratou, R. Cetnarskyj, N. Cartwright, C. Semple, A. J. Clark, F. J. Reid, L. A. Smith, K. Kavoussanakis, T. Koessler, P. D. Pharoah, S. Buch, C. Schafmayer, J. Tepel, S. Schreiber, H. Volzke, C. O. Schmidt, J. Hampe, J. Chang-Claude, M. Hoffmeister, H. Brenner, S. Wilkening, F. Canzian, G. Capella, V. Moreno, I. J. Deary, J. M. Starr, I. P. Tomlinson, Z. Kemp, K. Howarth, L. Carvajal-Carmona, E. Webb, P. Broderick, J. Vijayakrishnan, R. S. Houlston, G. Rennert, D. Ballinger, L. Rozek, S. B. Gruber, K. Matsuda, T. Kidokoro, Y. Nakamura, B. W. Zanke, C. M. Greenwood, J. Rangrej, R. Kustra, A. Montpetit, T. J. Hudson, S. Gallinger, H. Campbell, and M. G. Dunlop. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet*, 40(5):631–7, 2008.

91. G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein. The international hapmap project web site. *Genome Res*, 15(11):1592–3, 2005.

92. D. A. Tregouet, S. Escolano, L. Tiret, A. Mallet, and J. L. Golmard. A new algorithm for haplotype-based association analysis: the stochastic-em algorithm. *Ann Hum Genet*, 68(Pt 2):165–77, 2004.

93. D. A. Tregouet and V. Garelle. A new java interface implementation of thesias: testing haplotype effects in association studies. *Bioinformatics*, 23(8):1038–9, 2007.

94. D. A. Tregouet and L. Tiret. Cox proportional hazards survival regression in haplotype-based association analysis using the stochastic-em algorithm. *Eur J Hum Genet*, 12(11):971–4, 2004.

95. S. Wacholder, S. Chanock, M. Garcia-Closas, L. El Ghormli, and N. Rothman. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*, 96(6):434–42, 2004.

96. E. R. Waldron, J. C. Whittaker, and D. J. Balding. Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol*, 30(2):170–9, 2006.

97. Y. Wang, P. Broderick, E. Webb, X. Wu, J. Vijayakrishnan, A. Matakidou, M. Qureshi, Q. Dong, X. Gu, W. V. Chen, M. R. Spitz, T. Eisen, C. I. Amos, and R. S. Houlston. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet*, 40(12):1407–9, 2008.

98. S. Weidinger, C. Gieger, E. Rodriguez, H. Baurecht, M. Mempel, N. Klopp, H. Gohlke, S. Wagenpfeil, M. Ollert, J. Ring, H. Behrendt, J. Heinrich, N. Novak, T. Bieber, U. Kramer, D. Berdel, A. von Berg, C. P. Bauer, O. Herbarth, S. Koletzko, H. Prokisch, D. Mehta, T. Meitinger, M. Depner, E. von Mutius, L. Liang, M. Moffatt, W. Cookson, M. Kabesch, H. E. Wichmann, and T. Illig. Genome-wide scan on total serum ige levels identifies fcer1a as novel susceptibility locus. *PLoS Genet*, 4(8):e1000166, 2008.

99. W. J. Welch. Construction of permutation tests. *J Am Stat Assoc*, 85:693–8, 1990.

100. J. E. Wigginton, D. J. Cutler, and G. R. Abecasis. A note on exact tests of hardy-weinberg equilibrium. *Am J Hum Genet*, 76(5):887–93, 2005.

101. J. S. Witte. Multiple prostate cancer risk variants on 8q24. *Nat Genet*, 39(5):579–80, 2007.

102. M. Yeager, N. Orr, R. B. Hayes, K. B. Jacobs, P. Kraft, S. Wacholder, M. J. Minichiello, P. Fearnhead, K. Yu, N. Chatterjee, Z. Wang, R. Welch, B. J. Staats, E. E. Calle, H. S. Feigelson, M. J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F. R. Schumacher, E. Giovannucci, W. C. Willett, G. Cancel-Tassin, O. Cussenot, A. Valeri, G. L. Andriole, E. P. Gelmann, M. Tucker, D. S. Gerhard, Jr. Fraumeni, J. F., R. Hoover, D. J. Hunter, S. J. Chanock, and G. Thomas. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet*, 39(5):645–9, 2007.

103. C. C. Yeh, R. M. Santella, L. L. Hsieh, F. C. Sung, and R. Tang. An intron 4 vntr polymorphism of the endothelial nitric oxide synthase gene is associated with early-onset colorectal cancer. *Int J Cancer*, 124(7):1565–71, 2009.

104. J.H. Zhao. gap: Genetic analysis package. *J Stat Soft*, 23(8), 2007.