Chapter 3

Association Mapping

Jodie N. Painter, Dale R. Nyholt, and Grant W. Montgomery

Abstract

Association mapping seeks to identify marker alleles present at significantly different frequencies in cases carrying a particular disease or trait compared with controls. Genome-wide association studies are increasingly replacing candidate gene-based association studies for complex diseases, where a number of loci are likely to contribute to disease risk and the effect size of each particular risk allele is typically modest or low. Good study design is essential to the success of an association study, and factors such as the heritability of the disease under investigation, the choice of controls, statistical power, multiple testing and whether the association can be replicated need to be considered before beginning. Likewise, thorough quality control of the genotype data needs to be undertaken prior to running any association analyses. Finally, it should be kept in mind that a significant genetic association is not proof positive that a particular genetic locus causes a disease, but rather an important first step in discovering the genetic variants underlying a complex disease.

Key words: Genetic association, allele, single nucleotide polymorphism (SNP), genotype, genome-wide association (GWA), genetic power.

1. Introduction

Association mapping seeks to identify marker alleles present at significantly different frequencies in cases carrying a particular phenotype (a disease or trait) compared with controls. This contrasts with linkage mapping (covered in **Chapter 2**) which searches for chromosomal regions shared by family members who are affected by the disease under study. Association mapping relies on linkage disequilibrium (LD), where some combinations of alleles at loci close together in the genome occur more often than by chance because of previous population history. A genetic marker close to and in LD with a causal variant will show significantly different allele frequencies in cases compared with appropriate control

B. Yu, M. Hinchcliffe (eds.), *In Silico Tools for Gene Discovery*, Methods in Molecular Biology 760,
DOI 10.1007/978-1-61779-176-5_3,

individuals. One consequence is that in a successful association mapping study the marker associated with a disease is unlikely to be the casual variant. The marker locus will likely be located close to the variant or variants contributing to disease risk, but further studies will be required to locate and characterise these casual variants.

Association studies can include family-based designs (1, 2). In general, case–control studies have greater power than family-based designs when genotyping equivalent numbers of individuals and for simplicity in this chapter we have restricted the discussion and examples to case–control studies. Association mapping designs range from genotyping a single marker in a 'candidate' gene through to genome-wide association (GWA) studies. Current GWA studies genotype 300,000–2,500,000 single nucleotide polymorphisms (SNPs) per individual, and this number will soon reach 5,000,000 SNPs per individual. Prior to the development of genotyping methods using high-density SNP 'chips', studies concentrated on genotyping markers in candidate genes chosen from an understanding of the biological mechanisms thought to contribute to the disease under study. Most studies genotyped small numbers of selected variants within the target gene and sample sizes were often low.

Genome-wide association strategies developed from advances in genotyping technology, greater understanding of the structure of common variation in the human genome and continued advances in computing power and software tools. The discoveries from many association studies in complex diseases clearly demonstrate that the effect size for common risk variants for most diseases is low with odds ratio's for the risk allele in the range of 1.1–1.5 (3–5). Large studies with several thousand cases and equivalent numbers of controls are required to have sufficient power to detect these small effects. For some traits, large international collaborations have developed to conduct association studies with samples sizes around 100,000 cases.

There are a number of different options for the analysis of association studies. However, as noted above, recent studies demonstrate that large sample sizes are necessary to have sufficient power to detect 'true' associations for most diseases. We have therefore discussed association mapping and provide examples using software that can be applied to both small and large studies.

2. Materials

The basic requirements for an association study are DNA genotypes and a computer with an Internet connection (in fact, many genetic data analysts work exclusively with electronic data!). First,

DNA samples must be extracted and genotyped to a high quality, and the methodology and techniques used to do both will depend on the scale of the project. While most laboratories are equipped to extract DNA and perform basic genotyping, a growing number of laboratories offer extraction and genotyping as a commercial service. This can be efficient for large-scale projects (*see* **Note 1**).

For association analyses there are a growing number of software programs being developed by commercial companies; however, all quality control and analysis stages can be either conducted using programs with a worldwide web interface or downloadable onto a personal computer directly from a Web site. Specialist software and the Web sites from which these are available will be presented in the relevant sections below (*see* **Note 2**).

3. Methods

3.1. Planning Your Study

Aspects that should be considered during the project planning phase are outlined briefly below, and the overall order of steps to perform an association study is shown in **Fig. 3.1**. There are now increasing numbers of reviews in the genetic literature dealing with specific aspects of association study design that should be referred to for more detail (*see* **Note 3**). These include reviews on issues to consider while planning association studies (6–9), data quality control (10), data analysis (11) and interpretation (4, 12) and replication analyses (13).

3.1.1. Selection of Cases and Controls

The first aspects to consider when designing an association study are that the trait or disease under investigation is heritable, and that all of your case subjects have the same phenotype. These seem obvious; however, complex diseases may have low levels of heritability and many are likely to be genetically heterogeneous and influenced by environmental risk factors. As a result, studies of complex diseases typically require extremely large sample sizes (e.g. 1000s of cases and controls) (10) in order to detect an association, particularly if the effect size is expected to be low. Careful phenotyping during case recruitment should reduce the risk of collecting a heterogeneous case sample, although the sample size attainable may need to be balanced with the cost of obtaining phenotypic information (e.g. clinical assessment versus detailed questionnaires).

Control subjects should ideally come from the same population as the cases to avoid issues of sample stratification (*see* **Section 3.1.2**). Controls can be 'selected', where each individual has been screened for the disease under investigation, or 'unselected', where individuals are taken from a general population

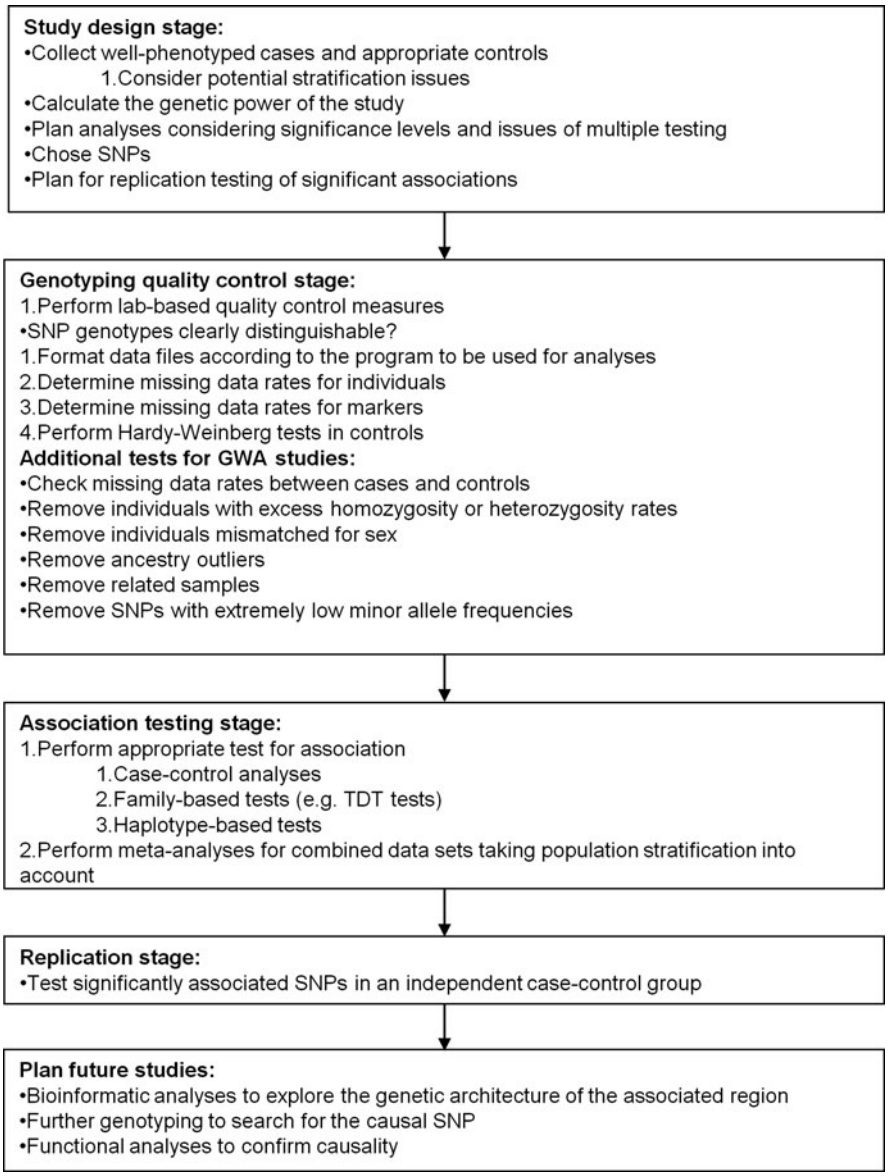

Fig. 3.1. Association study flow from initial planning to investigate a significant, replicated, association signal.

and for whom there is typically no information on disease status. Depending on the population prevalence a non-negligible proportion of unselected controls may carry the disease under investigation, hence a higher number of unselected controls is typically required than if controls are selected. The number of controls should equal or exceed the number of cases.

3.1.2. Stratification

Stratification refers to differences in allele frequencies of genetic variants between cases and controls due to the underlying

sampling scheme, which may lead to false-positive signals of association. To avoid technical stratification resulting from systematic differences in the way samples are handled, the collection of biological samples, DNA extraction and subsequent storage and genotyping should be performed in the same manner (and where possible in the same laboratories) for both cases and controls. Where case and control datasets have been collected and processed separately, for example GWA studies using large publically available control datasets (14), the quality of the genotyping should be compared prior to conducting association analyses.

Cases and controls should also be matched for ancestry to avoid population stratification due to genetic admixture (15). The inclusion of subgroups of genetically distinct individuals may lead to false-positive signals of association particularly if one ancestral population is over-represented amongst either the case or the control groups. Apparent association signals may then be due to differences in frequencies at ancestry-informative alleles, which have systematic differences in frequencies amongst populations. Ancestry should be determined at the subject recruitment stage by including questions on the birthplace of the subject and/or the subject's parents and grand parents in questionnaires, although ancestry outliers can now easily be detected with GWA data (**Section 3.3.4**, QC measure 7).

3.1.3. Statistical Power

The power of a study is the probability of rejecting your null hypothesis (*Ho*: that there are *no* differences in allele frequencies between cases and controls) when it is false (i.e. when association between a gene or locus and a disease actually exists). Power calculations allow estimation of the power of the study given the sample size, frequency of the disease-associated allele and the effect size associated with the risk allele.

Power calculations should be performed over a range of allele frequencies and effect sizes, as these are typically unknown for complex diseases (*see* **Note 4**). The web-based Genetic Power Calculator (16) (http://pngu.mgh.harvard.edu/~purcell/gpc/) allows power to be estimated for a number of study types (e.g. case–control studies for discrete or quantitative traits and family-based studies), while the program Power for Association with Error (17) (http://linkage.rockefeller.edu/pawe/) calculates power in the presence of a small proportion of genotyping errors. Power for studies using a two-stage design, where a large number of SNPs are first genotyped in a case–control group (stage 1) and only the most promising SNPs subsequently genotyped in a larger, independent case–control group (stage 2), and under different genetic models can be calculated using the CaTS program (18) (http://www.sph.umich.edu/csg/abecasis/cats/). The program Quanto (http://hydra.usc.edu/GxE/)

allows power to be calculated in the presence of gene × environment or gene × gene interactions.

3.1.4. Significance and Multiple Testing

Multiple testing refers to the increasing number of hypotheses that can be tested in a genetic association study, such as testing for association with over 500,000 markers in a GWA study or testing multiple SNPs per gene or multiple subgroups for a particular disease. As multiple testing increases the chance of obtaining a false-positive result the significance threshold should take the number of tests performed into account (19). The simplest method is the 'Bonferroni' correction, where the pre-determined threshold for significance is divided by the number of tests performed. For example, the significance threshold for a study that genotyped 30 'tagging' SNPs (*see* **Section 3.1.5**) would require a p value lower than $0.05/30 = 0.00167$ to claim association with a type I error (false-positive) rate of 5%. However, this method can be overly conservative and result in stringent thresholds for evidence of association (19). More specifically, performing such a correction assumes that each test is completely *independent* of all others, which is typically not the case due to LD between SNPs located within the same genomic region. The web-based program SNPSpD (20) (http://gump.qimr.edu.au/general/daleN/SNPSpD/) allows p value correction in the presence of LD by estimating an 'effective' number of independent SNPs. However, this method may also result in overly conservative p values, hence permutation and/or simulation procedures are considered the 'gold standard', although these are generally computationally intensive. Significance thresholds for GWA studies can also be determined on a per-project basis using permutation, although values $<5–10 \times 10^{-8}$ (i.e. 0.05/500,000–1,000,000 independent tests) are often taken as evidence of significant association (21, 22).

The goal of such Bonferroni-type corrections and adjusted significance thresholds is to guard against any single false positive occurring. However, in the high-dimensional setting of GWA studies one may also aim to identify as many true positive findings as possible while incurring a relatively low number of false positives. The false discovery rate (FDR) (23, 24) is designed to quantify this type of trade-off, making this particularly useful to identify loci worth further investigation.

3.1.5. Choosing SNPs

Nowadays, association studies are typically performed using SNPs as the genetic marker (*see* **Chapter 1** for more discussion on genetic markers). Candidate gene/region studies may include SNPs chosen as biologically plausible candidates for which positive associations have previously been reported (usually as an attempt to replicate such a finding) and, increasingly, 'tagging' SNPs to comprehensively account for common genetic variation

across a gene. Tagging relies on LD, the correlation between alleles at SNPs located in the same chromosomal region, and effectively reduces the numbers of SNPs that need to be typed as the genotype at one locus is highly correlated with loci in high LD, thus testing these loci too. Tagging can be performed through the HapMap database (http://www.hapmap.org) and using programs such as Haploview (25) (http://www.broadinstitute.org/haploview/haploview) or Goldsurfer2 (26) (http://www.well.ox.ac.uk/gs2/) (9). SNPs for GWA studies are provided on commercially produced genotyping chips designed to tag >90% of the common variation present in the human genome.

3.1.6. Replication

The gold standard for accepting that an association between a marker and a disease potentially exists and is worthy of further investigation is replication, where significant association to the same allele is detected in a completely independent case–control group. Many candidate gene associations fall down at this stage, as subsequent studies fail to replicate the initial findings. This may be due in part to the 'winner's curse', where the first report of an association is typically the most significant (27), or underlying risk alleles differing between populations. However, a careful review of associations that subsequently fail to replicate typically reveals that initial associations were either weak (with p values close to the significance threshold or where multiple testing was not taken into account) or underpowered due to low sample size. Replication studies should therefore be interpreted with respect to their level of power and whether there is strong statistical or biological evidence underpinning the original association.

3.2. Programs for Association Mapping Analyses

Association analyses test for differences in genotype or allele frequencies between cases and controls using for instance Chi-squared (χ^2) tests. For small numbers of SNPs, association tests can be performed by hand or using a pocket calculator by constructing a 2×2 table of allele counts. For the larger numbers of SNPs typically included in a candidate gene SNP tagging or GWA study, specialist analysis software is more practical. Any program that performs χ^2 tests can be used to test association, including SAS or SPSS. There are also a growing number of programs written specifically for the analysis of genetic association data, including plugins for the R package of statistical programs (e.g. GenABEL (28), http://mga.bionet.nsc.ru/~yurii/ABEL/GenABEL/) and SNPTEST (10, 29) (http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html) for the analysis of GWA data or the web-based SNPStats (30) (http://bioinfo.iconcologia.net/snpstats/start.htm) for smaller association studies. In the sections below we provide the options for performing quality control and association analyses using the program PLINK

(31) (http://pngu.mgh.harvard.edu/~purcell/plink/). While PLINK requires the user to be familiar with MS-DOS or Unix/Linux environments, it is a user-friendly program for the analysis of GWA data that allows rapid and flexible analyses to be performed on 1 to more than 1 million SNPs for 1000s of individuals, with an upper limit to the size of datasets determined only by computing power (*see* **Note 5**).

3.3. Data Quality Control

3.3.1. Laboratory-Based Quality Control

Genotyping studies should include appropriate controls to test for correct orientation of DNA sample tubes or plates and repeatability of genotype results across the study. Initial quality control (QC) is typically performed in the laboratory in which the genotypes were generated. Laboratory-based QC measures should focus on ensuring that all individuals and SNPs have genotyped accurately. For each SNP, homozygotes for the alternative alleles and heterozygotes for both alleles should be clearly distinguishable. For large-scale projects conducted using genotyping chips, this can be visualised as genotype 'clusters' (**Fig. 3.2**) (4).

3.3.2. Analysis File Preparation

The minimum requirement for an association study is individual genotype data assigned to case and control individuals and knowledge of the order and position of SNPs along a chromosome. File formats will depend on the analysis program that is utilised. Depending on the amount of data, files can be created and edited in a spreadsheet program such as Microsoft Excel or a text editor, and association analysis programs can then produce correctly formatted files. PLINK requires two files: a pedigree file containing individual information including disease status and genotype data and a map file containing the chromosomal positions of SNPs (*see* **Note 6**).

3.3.3. Running PLINK

PLINK can be driven by a graphical user interface (GUI) in Windows, command-line instructions typed directly into a MS-DOS or Unix command window or via a –script option, where the options are read from a text file. To save disk space and reduce analysis time, particularly for large GWA data files, PLINK can covert pedigree and map files to 'binary' files, which will have the endings *.bed* (containing genotype information), *.fam* (containing pedigree information) and *.bim* (containing map information). In the following sections we assume that data files have been converted to binary format. Note also that by default all output files will be named *plink.xxx* unless an outfile name (using the '--out' option) is specified.

The command line for a basic PLINK run has the format:

plink --bfile *mydatafile* --assoc --out *mydatafile*

This will call PLINK to perform an association analysis using information included in the *mydatafile.bed*, *mydatafile.fam*

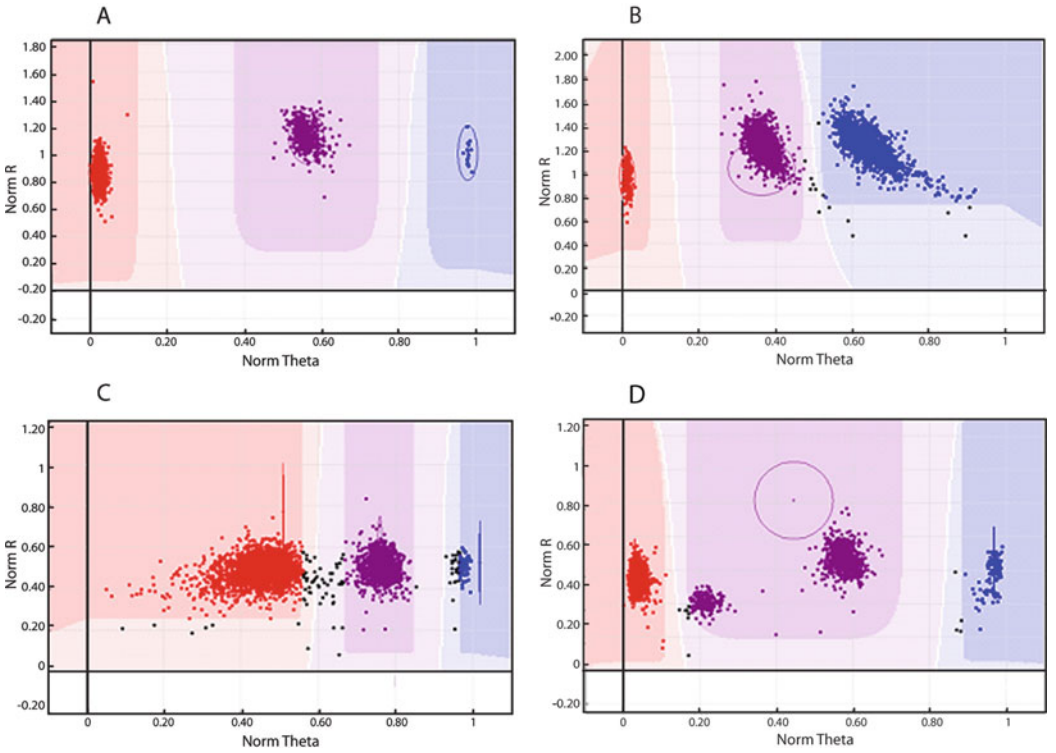

Fig. 3.2. SNP genotyping cluster plot examples. (**a**) SNP genotypes cluster well and homozygotes for either allele as well as heterozygotes carrying both alleles can be clearly distinguished from each other even though few samples are homozygous for one of the alleles. (**b**) Homozygotes for one allele cluster well but the range of values seen for heterozygotes and the homozygotes for the alternative allele cluster reasonably loosely. *Black dots* indicate samples with values outside of which genotypes can be confidently called. Such a SNP may fail quality control. (**c**) Although the clusters appear reasonably tight a number of samples carrying all three genotypes fall outside of the values for which genotypes can be confidently called. Data for SNPs clustering as in b and c could be rescued following visual inspection of the cluster plots. (**d**) Heterozygous genotypes for this SNP appear in two clusters, indicating the presence of a so-called null or non-amplifying allele (typically due to the presence of a SNP in the PCR primer site, preventing amplification in any DNA sample carrying such a variant). Such a SNP would likely *pass* quality control, but should in fact be removed from the dataset. Cluster plots of at least all significantly associated SNPs should be inspected before data analyses are taken further. This is particularly important for GWA studies that include 100,000 s of SNPs.

and *mydatafile.bim* files and produce a result file named *mydatafile.assoc* (or by default *plink.assoc* if no output file name was specified). Output files can be viewed and manipulated in a text editor or spreadsheet program (*see* **Note 7**).

3.3.4. In Silico Quality Control

In silico QC measures can be performed either in PLINK, other genetic analysis software or a spreadsheet program.

1. Remove individuals with low genotyping rates. The threshold for the removal of individuals due to missing data is dependent on the number of SNPs genotyped but is often set at 5%. Low individual genotyping rates are generally due to low concentration or poor quality DNA, and such

individuals should be removed as the genotypes obtained for other SNPs may be incorrect.
To run a 'missingness' analysis in PLINK the basic command line is

plink --bfile *mydatafile* --missing --out *mydatafile*

PLINK will produce two files: the *.imiss* file contains missing genotype rates per individual and the *.lmiss* file contains missing genotype rates per SNP.
Individuals with an excess of missing data (e.g. missing genotypes for >5% of the total number of SNPs genotyped) should be omitted from the next round of QC. This can be achieved in PLINK by including a 'remove' file containing the family and individual IDs of each individual to ignore during an analysis run. Alternatively, new data files excluding these individuals could be created, but such files take up a greater amount of disk space than a 'remove' file. It is also advisable to keep a copy of the original data files including all individuals and marker genotypes for future reference.

2. Remove SNPs with low genotyping rates. The next step is to remove SNPs with high rates of missing genotypes in the individuals remaining in the dataset following the first round of QC. Typically, SNPs missing more than 1–5% of data are excluded from further analyses.

   Re-run the PLINK '--missing' option. Note that the *.imiss* file will be overwritten if a new outfile name is not provided.

   plink --bfile *mydatafile* --missing --remove *removeindividualslist.txt* --out *mydatafileQCII*

   SNPs with an excess of missing data should be excluded from the next round of analysis by including the names of the SNPs in a separate file using the '--exclude' option or by creating new data files.

3. Investigate SNPs for evidence of Hardy–Weinberg disequilibrium (HWD). Under a neutral genetic model the frequencies of homozygote and heterozygote genotypes for a particular SNP are expected to equal the products of the allele frequencies: if the frequency of an allele 'A' = p and the frequency of the alternative allele 'G' = q, then the frequencies of AA, AG and GG genotypes should equal p^2, $2pq$ and q^2, respectively. Departures from Hardy–Weinberg equilibrium (HWE) may occur due to true association, where certain genotypes are over-represented in the case group. In GWA studies, HWE tests are generally performed only in controls where such departures are often due to poor genotyping. Departures from HWE in cases should be checked for any SNPs showing association to ensure the

departure is in the expected direction of the genotype over-representation/association. SNPs with extremely low HW p values ($<10^{-6}$) should be excluded.
To perform HWE tests in PLINK:

plink --bfile *mydatafile* --remove *removeindividualslist.txt* --exclude *excludeSNPlist.txt* --hardy --out *mydatafile*

PLINK will output a *.hardy* file. Any SNP markers that are not in HWE and have clear clustering issues should be added to the list of SNPs to exclude in further analyses (*see* **Note 8**).

The following additional QC measures are undertaken in GWA studies where the numbers of SNPs and individuals are large enough to provide accurate estimates. Typically, exclusion thresholds are determined on a per-project basis.

4. Check missing data rates between cases and controls and across all genotyping chips. To avoid stratification due to technical issues SNP missing data rates should be equivalent in the case and control groups, and across all genotyping chips included in an association analysis. In PLINK the test for differences in missing rates between cases and controls is performed using the '--test-missing' option. More complex group comparisons, such as chip effects (where genotyping success rates differ per chip), can be performed using the '--loop-assoc' option which automatically tests each group (defined by a categorical factor) versus all other individuals for a variety of statistics.
5. Remove individuals showing excessive homozygosity or heterozygosity. Excessive levels of either homozygosity or heterozygosity can indicate poor genotyping, typically due to low-quality DNA. 'Normal' individual heterozygosity levels using GWA data are typically of the order of ~0.3. Heterozygosity is measured as an inbreeding coefficient (denoted as F, the observed versus expected number of homozygous genotypes). The PLINK option to calculate heterozygosity is '--het'.
6. Remove individuals mismatched for sex. Large-scale genotyping chips include markers for both the X and the Y chromosomes. Males should be homozygous for X chromosome markers, while females should show a degree of X chromosome heterozygosity and have no genotypes for Y chromosome markers. The PLINK option to calculate F for the X chromosome is '--check-sex'.
7. Remove ancestry outliers. Many SNPs have systematic differences in allele frequencies between populations. On a genome-wide scale such SNPs can be used to determine

genetic ancestry. Programs such as Eigenstrat (32) (http://genepath.med.harvard.edu/~reich/Software.htm) are used to perform multidimensional scaling of pairwise 'identity by state' (IBS) values in comparison to reference populations from, e.g. the HapMap (*see* **Note 9**). The first two principle components of the IBS values for each individual are then plotted to reveal ancestry. In PLINK, outlying individuals can either be excluded from further analysis or the principal components be included as covariates ('--covar') in a logistic regression ('--logistic') test for association.

8. Remove related samples. Cryptic (unknown) relatedness between samples can produce erroneous association results due to increased allele sharing amongst relatives. Relatedness is generally estimated by calculating measures of IBS and/or 'identity by descent' (IBD). In PLINK the option '--genome' will generate a *.genome* file containing estimated IBD values for each pair of individuals in the dataset.
9. Exclude SNPs with very low minor allele frequencies (MAFs). While rare SNPs (MAFs <1%) of large effect are likely to be important in at least some complex diseases, SNPs with very rare minor allele frequencies may cause spurious association signals if present in one group (e.g. cases) but not the other (e.g. controls). Very rare alleles are more likely to differ in frequency due to chance, and due to their rarity any truly disease-associated variant must be of particularly large effect for a real association signal to be detected. Unless the effect is large (odds ratio >2) most case–control samples (those with <10,000 individuals) have minimal power to detect association to less common alleles (e.g. MAFs <5%).
   Allele frequencies can be calculated in PLINK using the '--freq' option, and SNPs with MAFs less than the threshold for inclusion can then be added to the 'exclude' file. Alternatively, PLINK can filter SNPs based on allele frequencies using the option '--maf 0.01' (for example), which will exclude SNPs with MAFs <0.01 from the analysis.

3.4. Association Analyses

1. Running a basic association analysis. Once all QC steps have been completed the next step is to perform the association analysis on the clean dataset.
   In PLINK the '--assoc' option will perform χ^2 tests of association on each SNP in the data file, producing an *.assoc* output file containing *p* values and odds ratios for each marker. PLINK allows for the use of separate phenotype files that over-ride the phenotypes in the main .fam or .ped files. This

option is particularly useful if the trait under investigation has distinct, well-characterised sub-phenotypes that could be run in alternative analyses without the need for multiple pedigree files to be produced. Individuals failing QC could also be removed from the analysis by setting their phenotype to '0' in the phenotype file to avoid the need to include a '--remove' file.

--bfile *mydatafile* --pheno *phenotypefilename.txt* --remove *removeindividualslist.txt* --exclude *excludeSNPlist.txt* --assoc

2. Running an association analysis in the presence of population stratification. Due to the extremely large sample sizes required to find variants of even moderate effect, it is becoming increasingly common for researchers from different centres to combine their data sets prior to running GWA analyses to maximise the power to detect association. Analyses should then be run taking potential population stratification into account. This can be done in two ways depending on the data files at hand.
   If genotype data are available the Cochran–Mantel–Haenszel (CMH) test can be used to test for association in the presence of different population groups, while possible heterogeneity in disease-marker associations between the different groups can be estimated using Breslow-Day (BD) tests. An additional 'within' file is required to indicate the group (or 'cluster') to which each individual belongs:

   plink --bfile *mydatafile* --pheno *phenotypefilename.txt* --remove *removeindividualsfile.txt* --exclude *excludeSNPsfile.txt* --within *individualclusterinfo.txt* --mh --bd --out *mydatafile*

   PLINK will produce two files, *.cmh* and *.bd*. Confidence intervals for odds ratios can be calculated using the option '--ci 0.95'.

   PLINK can also run 'meta-analyses' using files containing only the results (*p* values, etc.,) for different projects. The options here are

   plink --meta-analysis *project1.assoc project2.assoc* --out *metaanalysisresults*

3.5. Visualising and Interpreting Your Results

The results generated in small-scale studies can be easily accessed from the output of the program used to perform the association testing. The interpretation of GWA studies, where 100,000 s of tests have been performed, is simpler if the results are visualised as plotted figures. The first plot that should be made is a diagnostic quartile–quartile (q–q) plot (**Fig. 3.3**), produced by plotting

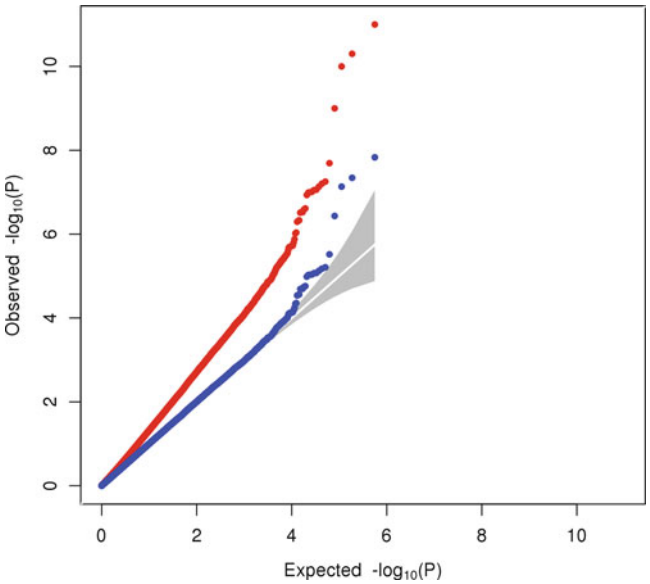

Fig. 3.3. Quartile–quartile (q–q) plot of hypothetical GWA results. The solid *white line* represents the expected (reference) distribution under the null hypothesis of no association and the *grey shaded region* indicates the point-wise 95% confidence interval envelope based on the standard errors of order statistics. The *red points* indicate population stratification and/or cryptic relatedness (substructure), while the blue points show no evidence for substructure, but convincing evidence for an excess of disease associations.

the observed values of the association statistics (e.g. the χ^2 or p values) ranked in order from smallest to largest against those expected under a null distribution. Deviations from the diagonal line give an indication of the quality of the data, in terms of controlling for population stratification and the strength of the associations detected (4, 12).

The second plot is a display of the association results themselves, termed a Manhattan plot (**Fig. 3.4**). Here the $-\log_{10}$ of the p values generated by the association analysis are plotted against chromosomal location, allowing interesting association signals to be clearly seen against background signals. Two user-friendly programs that can generate both q–q plots and Manhattan plots directly from PLINK output files are Haploview and WGAViewer (33) (http://people.genome.duke.edu/~dg48/WGAViewer/). These and other programs (e.g. SNAP, http://www.broadinstitute.org/mpg/snap/ldplot.phpcan – change the 'plot type' drop-down window to 'regional association plot' – and LocusZoom, http://csg.sph.umich.edu/locuszoom/) can then be used to focus on the areas harbouring interesting association signals, for example displaying the genes and features such as microRNAs present in the region and providing links to genetic databases that can serve as the starting point for in silico investigation of the area surrounding significant association signals.

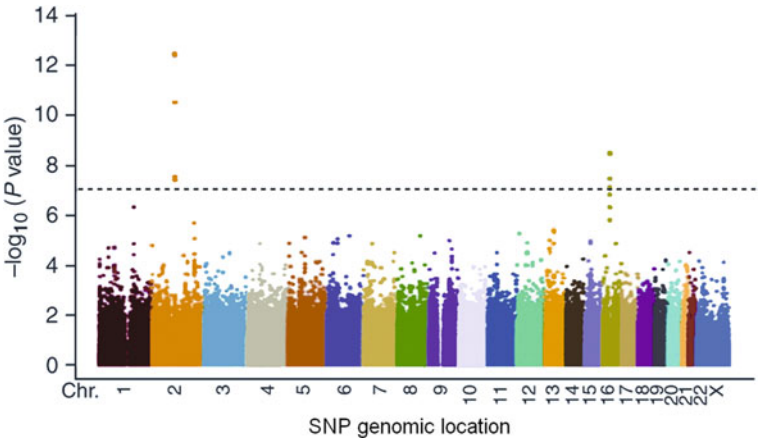

Fig. 3.4. Manhattan plot of hypothetical GWA results. *P* values for each SNP analysed in the GWA study are shown as their $-\log_{10}$ values. Each chromosome is represented by a different colour. The *dashed line* shows the threshold for genome-wide significance (accounting for the number of independent tests performed). Two regions in this example have reached genome-wide significance and should be targeted for replication.

3.6. Additional Considerations

It is increasingly clear that very large samples of well-phenotyped individuals are required to detect the typically modest effect sizes for risk alleles underlying susceptibility to complex genetic diseases or traits. The association analysis field is changing rapidly to adapt to the increasing complexities involved in mapping human disease genes. In addition to careful planning and QC, once at the analysis phase there are a growing number of methods that can be employed to increase the likelihood of detecting an association. For example, genotypes for untyped SNPs can be generated by imputation (34) in reference to individuals taken from HapMap or the 1000 Genomes project (http://www.1000genomes.org), increasing the total number of loci that can be analysed. Haplotype associations can be examined to determine the genetic background on which a causal variant may lie. Potential interactions between loci can also be investigated. While it should be remembered that even a highly significant association signal replicated in an independent case–control group is not absolute proof that a marker or gene is associated with the phenotype under investigation, such results are extremely encouraging and indicate that further in silico and genetic analyses and subsequent functional investigations should be carried out.

4. Notes

1. There is no substitute for high-quality DNA with carefully measured concentrations. The tissue type (e.g. blood), collection method, storage and transport of samples, DNA

extraction and subsequent storage of the DNA itself need to be carefully planned well in advance. Poor quality DNA will result in genotyping errors or failure, wasting considerable time and research funds.

2. Just as for laboratory work it is a good idea to keep written records of your work, including all quality control measures and analyses performed, as you will quickly amass a great deal of data that may be spread across a large number of computer files. This also ensures you can repeat analyses should you need to or pick up any errors in the analytical process.
3. The association mapping field is rapidly evolving, particularly as researchers gain more experience with GWA data. Literature searches (using a site such as the NCBI PubMed http://www.ncbi.nlm.nih.gov/pubmed/) should be undertaken regularly to keep up with new methodologies as they are published.
4. If risk allele frequencies and effect sizes are unknown, power calculations should be run over a realistic range, for example allele frequencies from 0.05 to 0.4 and odds ratios of 1.1–2.0 (although odds ratios for complex disease are typically in the range of 1.1–1.5). These can then be plotted on a graph to visualise the power expected over the range of values.
5. Another advantage of using PLINK is that the Web site has extensive, easy to read, documentation on all aspects of file preparation, quality control and analysis and is therefore an extremely useful resource even for those using other analysis programs.
6. The most common input file format used is the so-called linkage format, following that initially required by the original 'Linkage' program (*see* http://linkage.rockefeller.edu/soft/list2.html#l under 'L'). PLINK is reasonably flexible with regards to input file formats, consult the Web site to determine what is most appropriate for your study.
7. PLINK will also output a *.log* file containing all details of the analysis that has just run (input files and commands, etc.,) that should be kept for future reference. This *.log* file will have the default name *plink.log* and will be overwritten with each new analysis if an output file name is not specified via the '--out' option.
8. Various QC steps, including the 'missing' and 'hardy' steps, can also be performed during an association analysis by providing pre-determined thresholds for each measure in a single command line (*see* the 'filter' section of the PLINK

documentation). However, it is highly recommended that all QC steps are performed individually as this ensures a greater understanding of the quality of the dataset and allows decisions of what data should be excluded to be made based on actual missing data and HWD rates.

9. The HapMap (at http://www.hapmap.org) is an invaluable resource for association mapping projects providing genotyping data for millions of SNPs in a genomic context. Other easy to use databases that the reader should become familiar with as a starting point for genetic research are the UCSC Genome Browser (http://genome.ucsc.edu/), ENSEMBL (http://www.ensembl.org/index.html) and NCBI (http://www.ncbi.nlm.nih.gov/) databases.

References

1. Laird, N. M., and Lange, C. (2006) Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* 7, 385–394.
2. Benyamin, B., Visscher, P. M., and McRae A. F. (2009) Family-based genome-wide association studies. *Pharmacogenomics* 10, 181–190.
3. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106, 9362–9367.
4. McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9, 356–369.
5. Visscher, P. M., and Montgomery, G. W. (2009) Genome-wide association studies and human disease: from trickle to flood. *JAMA* 302, 2028–2029.
6. Zondervan, K. T., Cardon L. R., and Kennedy, S. H. (2002) What makes a good case-control study? Design issues for complex traits such as endometriosis. *Hum Reprod* 17, 1415–1423.
7. Hirschhorn, J. N., and Daly, M. J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6, 95–108.
8. Wang, W. Y., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 6, 109–118.
9. Pettersson, F. H., Anderson, C. A., Clarke, G. M., Barrett, J. C., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2009) Marker selection for genetic case-control association studies. *Nat Protoc* 4, 743–752.
10. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678.
11. Balding, D. J. (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7, 781–791.
12. Pearson, T. A., and Manolio, T. A. (2008) How to interpret a genome-wide association study. *JAMA* 299, 1335–1344.
13. Kraft, P., Zeggini, E., and Ioannidis, J. P. (2009) Replication in genome-wide association studies. *Stat Sci* 24, 561–573.
14. Zhuang, J. J., Zondervan, K., Nyberg, F., Harbron, C., Jawaid, A., Cardon, L. R., Barratt, B. J., and Morris, A. P. (2010) Optimizing the power of genome-wide association studies by using publicly available reference samples to expand the control group. *Genet Epidemiol* 34, 319–326.
15. Cardon, L. R., and Palmer, L. J. (2003) Population stratification and spurious allelic association. *Lancet* 361, 598–604.
16. Purcell, S., Cherny, S. S., and Sham, P. C. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19, 149–150.
17. Gordon, D., Haynes, C., Blumenfeld, J., and Finch, S. J. (2005) PAWE-3D: visualizing power for association with error in case-control genetic studies of complex traits. *Bioinformatics* 21, 3935–3937.

18. Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2007) Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol* 31, 776–788.
19. Cardon, L. R., and Bell, J. I. (2001) Association study designs for complex diseases. *Nat Rev Genet* 2, 91–99.
20. Nyholt, D. R. (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 74, 765–769.
21. Dudbridge, F., and Gusnanto, A. (2008) Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 32, 227–234.
22. Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 32, 381–385.
23. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001) Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* 125, 279–284.
24. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100, 9440–9445.
25. Barrett, J.C., Fry, B., Maller, J., and Daly, M. J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265.
26. Pettersson, F., Morris, A. P., Barnes, M. R., and Cardon, L. R. (2008) Goldsurfer2 (Gs2): a comprehensive tool for the analysis and visualization of genome wide association studies. *BMC Bioinformatics* 9, 138.
27. Kraft, P. (2008) Curses—winner's and otherwise – in genetic epidemiology. *Epidemiology* 19, 649–651.
28. Aulchenko, Y. S., Ripke, S., Isaacs, A., and van Duijn, C. M. (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23, 1294–1296.
29. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39, 906–913.
30. Sole, X., Guino, E., Valls, J., Iniesta, R., and Moreno, V. (2006) SNPStats: a web tool for the analysis of association studies. *Bioinformatics* 22, 1928–1929.
31. Purcell, S., Neale, B., Todd-Brown, K., et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81, 559–575.
32. Price, A. L., Patterson, N. J., Plenge, R. M., et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38, 904–909.
33. Ge, D., Zhang, K., Need, A. C., et al. (2008) WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res* 18, 640–643.
34. Li, Y., Willer, C. J., Sanna, S. and Abecasis, G. R. (2009) Genotype Imputation. *Ann Rev Genomics Hum Genet* 10, 387–406.