Citation: BoneKEy Reports **4**, Article number: 635 (2015) | doi:10.1038/bonekey.2015.2

www.nature.com/bonekey

LABORATORY METHODS

Big data challenges in bone research: genome-wide association studies and next-generation sequencing

Nerea Alonso[1], Gavin Lucas[2] and Pirro Hysi[3]

[1]Rheumatic Diseases Unit, MRC Institute of Genetics and Molecular Medicine, Centre for Genomic and Experimental Medicine, Western General Hospital, University of Edinburgh, Edinburgh, UK. [2]Clear Genetics SL, Barcelona, Spain. [3]Department of Twin Research and Genetic Epidemiology, King's College London, Saint Thomas Hospital, London, UK.

Genome-wide association studies (GWAS) have been developed as a practical method to identify genetic loci associated with disease by scanning multiple markers across the genome. Significant advances in the genetics of complex diseases have been made owing to advances in genotyping technologies, the progress of projects such as HapMap and 1000G and the emergence of genetics as a collaborative discipline. Because of its great potential to be used in parallel by multiple collaborators, it is important to adhere to strict protocols assuring data quality and analyses. Quality control analyses must be applied to each sample and each single-nucleotide polymorphism (SNP). The software package PLINK is capable of performing the whole range of necessary quality control tests. Genotype imputation has also been developed to substantially increase the power of GWAS methodology. Imputation permits the investigation of associations at genetic markers that are not directly genotyped. Results of individual GWAS reports can be combined through meta-analysis. Finally, next-generation sequencing (NGS) has gained popularity in recent years through its capacity to analyse a much greater number of markers across the genome. Although NGS platforms are capable of examining a higher number of SNPs compared with GWA studies, the results obtained by NGS require careful interpretation, as their biological correlation is incompletely understood. In this article, we will discuss the basic features of such protocols.

BoneKEy Reports **4**, Article number: 635 (2015) | doi:10.1038/bonekey.2015.2

Introduction

Background

Genome-wide association studies (GWAS) have been very successful in identifying genetic factors associated with disease or other human traits. GWAS allow inference over the whole length of the genome by acquiring direct information on a relatively small number of loci, taking advantage of the presence of blocks of high linkage disequilibrium (LD), separated by recombination hot spots in the genome.[1–4] A comprehensive list of GWAS can be found in the NHGRI Catalogue of published GWAS at http://www.genome.gov/gwastudies.

Next-generation sequencing (NGS) relies on direct acquisition of information from all amenable loci. The applicability of one over the other depends on the fine balance between the lower costs of the former, allowing for analyses of larger cohorts, versus the superior power of the latter to analyse variants that are not easily characterised through other variants in LD with them.

Genome-wide studies

Previous generations of genetic studies made important contributions in understanding the genetic basis of rare

Correspondence: Dr N Alonso, Rheumatic Diseases Unit, MRC Institute of Genetics and Molecular Medicine, Centre for Genomic and Experimental Medicine, Western General Hospital, University of Edinburgh, Crewe Road, Edinburgh EH4 2XU, UK.
E-mail: n.alonso@ed.ac.uk
or Dr G Lucas, Clear Genetics SL, C/ Muntaner 53, Barcelona 08011, Spain.
E-mail: gavin.lucas@cleargenetics.com
or Dr P Hysi, Department of Twin Research and Genetic Epidemiology, King's College London, Saint Thomas Hospital, Westminster, London SE1 7EH, UK.
E-mail: pirro.hysi@kcl.ac.uk

Received 30 April 2014; accepted 12 December 2014; published online 11 February 2015

diseases that showed clear patterns of familial inheritance. They required mathematically very complex procedures to identify genes by demonstrating their cosegregation with the phenotype. In contrast, linkage was less successful for common and more genetically complex diseases. Owing to advances in genotyping technologies and availability of new information on the structure of the human genome (Human Genome (http://www.genome.gov/10001772) and HapMap (http://hapmap.ncbi.nlm.nih.gov/) projects), the road was paved for LD association that would allow a better understanding of these common complex diseases, using genome-wide genotyping or sequencing.

The methodologies used in modern genetic research are different from those of previous generations (e.g., linkage analysis). Genome-wide genetic analyses have to find informatics solutions to deal with extremely large quantities of data and select the results that are most likely to be true positives from a distribution of probabilities.

It is important to remember that quantitative genetics does not provide definite proofs, but rather a probabilistic evaluation of the likelihood of a hypothesis. Several factors influence the ability to discriminate between results that are likely to represent true positives versus random noise, such as statistical power (which is a direct function of sample size), amplitude of the effect from that genetic locus, frequency of the risk factor (risk allele) and degree of the multiple testing of independent hypotheses. Although power is a direct result of a predetermined study design, multiple testing is calculated on the basis of the degrees of independence (LD) between common polymorphisms in the genome. This will depend on the allele spectrum studied and the ethnicity of the sample. On the basis of permutational work, it is established that for most ethnic groups the multiple testing threshold is $\sim 5e-8$[5] for the average HapMap2-based GWAS. This practical criterion has been demonstrated in practice to be a useful guide, and it has withstood the test of time. Nevertheless, this threshold is equivalent to an empirical significance of $P = 0.05$ after multiple testing, and to minimise the risk of false positives there is an imperative need to validate results by replicating them in independent samples.[6] If a robust association between a phenotype and a gene is thus confirmed, we can be more confident that the gene is implicated in the phenotype under investigation. However, research should ideally not be limited to the most strongly associated variant, but it rather needs to consider the whole gene as a strong candidate, requiring more extensive exploration at the genetic level, as well as through other approaches (e.g., omics).

Scope of this review

This review summarises and outlines some practical details of the key steps in the quality control (QC) and analysis of genetic data for GWAS and NGS studies. We offer a step-by-step protocol, providing the future user with a list of tools and information enough to help them decide how to perform the analysis.

Materials

In this protocol, we refer to a series of informatics tools, which are mostly freely available. The recommended software is listed in **Table 1**.

PLINK software

Storage, management and analysis of high-throughput genetic data are made significantly easier by the PLINK software.[7]

This tool has been designed to perform a range of analysis in a computationally efficient manner. It is both freely available and user-friendly, and it works through command lines for data management, QC statistics, population stratification detection, single SNP (single-nucleotide polymorphism) association analyses, haplotypic tests, copy number variant analysis and meta-analysis.

Genotyping results must be presented into any of the accepted PLINK formats. For example, the PED file sets a linkage-style pedigree file that contains genotype and phenotype data, whereas a map file contains SNP location data. The tutorial shows examples of the input and output files in each step.

Table 1 Software suggested for the management and analysis of data obtained from GWAS and NGS techniques

Software	*Web*	*OS*	*Task*
PLINK	http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml	Windows, Mac, Linux	Data management, quality control, statistical analysis
R	http://www.r-project.org/	Windows, Mac, Linux	Statistical computing software, graphics
IMPUTE2	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Windows, Mac, Linux	Genotype imputation and haplotype phasing
SHAPEIT2	http://www.shapeit.fr/	Mac, Linux	Estimation of haplotypes
MACH 1.0	http://www.sph.umich.edu/csg/abecasis/MACH/download/	Windows, Mac, Linux	Inferring missing genotypes, resolve long haplotypes
Minimac	http://genome.sph.umich.edu/wiki/Minimac	Linux	Implementation of MACH genotype imputation
BEAGLE	http://faculty.washington.edu/browning/beagle/beagle.html	Windows, Mac, Linux, Unix	Phasing, inferring missing genotypes, imputation, association analysis
METAL	http://www.sph.umich.edu/csg/abecasis/metal/	Linux, Unix, DOS	Meta-analysis
Haploview	http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview	Windows, Mac, Linux, Unix (using JAR)	Manhattan plot, linkage disequilibrium
GRAIL	http://www.broadinstitute.org/mpg/grail/	Windows	Gene relationships
GATK	http://www.broadinstitute.org/gatk/download	Linux	Quality control, analysis of NGS data
Vcftool	http://vcftools.sourceforge.net/options.html	Linux	Working with VCF files (1000G)

Abbreviations: GWA, genome-wide association; NGS, next-generation sequencing.

If there are more SNPs than individuals, as it almost always is the case for GWAS analysis, it is convenient using transposed (TPED and TFAM) input file format, resulting in shorter lines, which are faster to read using Perl or even PLINK.

One of PLINK's most important features is the binary file format, which stores voluminous genotype data more efficiently, using less computation resources. Genotype data is stored in the BED file, pedigree information is stored in the FAM file and SNP location data is stored in the BIM file (the latter two files are flat files). These files can be created using the following script (where in this particular example PLINK will expect the files *mydata.ped* and *mydata.map* as input):

```
PLINK --file mydata --make-bed
```

This command will create three files: *mydata.bed*, *mydata.fam* and *mydata.bim*.

To read the binary file format, use `--bfile mydata` instead of `--file mydata`, which is used to read linkage-style pedigree files.

Methods

Genome-wide association studies

Study design. Study design is a key consideration before undertaking a genome-wide association analysis. GWAS can be performed using any of the types of analytical designs that could be applied for any other clinical variable of interest, including case–control, family-based, cross-sectional or cohort studies. Regardless of the design, informed consent must be obtained from all participants beforehand, following established ethical standards and institutional ethical approvals, as for all studies involving human subjects.

One of the key determinants of success is power, which will depend on a set of factors, including the genetic architecture of the phenotype of interest. The phenotype itself cannot be modified, and it can have a naturally stronger or weaker causal relationship with the underlying genetic architecture (e.g., heritability, penetrance, presence of pleiotropy, phenocopies, misclassification and so on), or simply be more or less difficult to measure precisely (e.g., Marfan syndrome or osteogenesis imperfecta). Other factors may have strong influence on power, including magnitude of the effect of an associated variant on the phenotype (effect size), the frequency of the variant and how well the genotyped variants (otherwise known as genetic markers) capture the causative variants through LD. Thus, a GWAS will have greater power to detect associated alleles that have stronger effects on the phenotype, are more common or are more strongly correlated with one or more genotyped variants. Clearly, therefore, not all truly associated variants are equally likely to be detected by a GWAS.

However, a key and modifiable determinant of power is sample size. A good study design should take this into account by including a sufficient number of individuals displaying the phenotype of interest (i.e., number of cases). It may be possible to increase the sample size of the control group using control samples (or even cases of unrelated disease) from previous studies or publicly available data sets. At any rate, it is not recommended that the number of controls be more than four times that of cases, as only modest power gains at the price of added genotyping work can be achieved beyond this ratio.[8] Other modifiable power determinants include genomic coverage, imputation accuracy (where applicable) and laboratory procedures that may affect genotyping accuracy, sample matching, genotype QC and so on.

Sample matching is the preferable means to minimise Type I error rate as per other epidemiologic studies. Yet, the study design should also consider the clinical characteristics of the samples, the presence of potential confounders and sources of bias, as these variables can be included in the association analysis as covariates.

Biological sample preparation. Good practice when storing and tracking the samples is essential to minimise spurious results and loss of overall study power. DNA samples are stored long term at temperatures of $-4\,^{\circ}C$ and backed up for future validation and troubleshooting.

The manufacturer's instructions should then be followed to guide the DNA quantity required for a particular assay or chip. DNA quantification should be performed using picogreen quantification method.

Genotyping. Depending on the genotyping array selected, between 300 000 and 2.5M SNPs can be genotyped through current commercially available SNP chips. Data from The HapMap Project suggest that the majority of SNPs in the human genome with a minor allele frequency (MAF) of $\geqslant 5\%$ may be detected using 550 000 tagging SNP markers in European and Asian populations, or 1 100 000 tagSNPs in African populations.[4] The major providers of high-throughput genotyping array platforms on the market are Illumina (San Diego, CA, USA) and Affymetrix (Santa Clara, CA, USA). Each SNP array has its unique specifications, costs, coverage and extra features (e.g., copy number variations or mitochondrial SNPs).[9] Results from chips can be visualised using the proprietary GenomeStudio or BeadSudio Suite (Illumina), or the freely available Genotyping Console suite (Affymetrix).

Genotype QC. A proposed global strategy for QC testing is shown in **Figure 1**. Apart from PLINK, extensively described below, various QC steps can be carried out using other software packages such as qtool, GenABLE, GS2 and so on. In this protocol, we recommend performing the individual-level QC, because individual samples are more likely to be unreliable than commercialised SNP assays. All command lines listed in the following section have been inspired from the PLINK software tutorial. In this protocol, we aim to summarise this extensive tutorial, to perform a basic QC test. Further information can be found at the following website: http://pngu.mgh.harvard.edu/~purcell/plink/tutorial.shtml.

Individual-level QC. For some of the QC steps described below, individuals can be filtered from the data set directly using some argument (e.g., a 'missingness' threshold), whereas others require the user to create a list of subject IDs to be removed using the `--filter` argument in PLINK, or to be retained using the `--keep` argument.

The PLINK tutorial provides a complete guide for the QC process. It is recommended that a clean data set is produced at the conclusion of the QC steps described here, so that the integrity of future GWAS analyses is not compromised by shifting criteria.

Genotyping efficiency: When a large proportion of SNPs fail in the same sample, this indicates poor-quality DNA for this

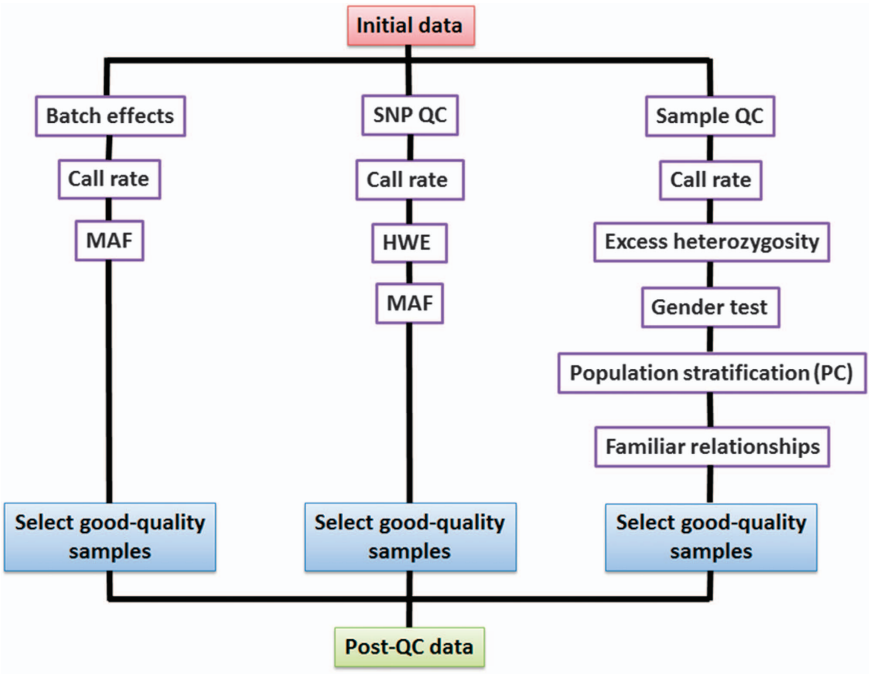

Figure 1 Flowchart showing the different steps to take into account when performing the quality control testing on a GWAS.

sample, and thus the results provided for this sample are generally less trustworthy. When genotyping a large number of SNPs, the recommended threshold is 98–99%,[10] such that in a GWAS individuals would be eliminated if they had missing genotypes for more than ~5000 SNPs. The following command from the PLINK tutorial can be used to obtain a summary of missingness in your sample:

```
plink --bfile mydata --missing --out mydata_missing
```

Individuals with an excess of missing genotype data can be removed as follows:

```
plink --bfile mydata --mind 0.01 --make-bed --out mydata_missing
```

Excess of heterozygosity is a clear indicator of DNA contamination. The test essentially compares the observed (O(HOM)) and expected (E(HOM)) number of homozygotes, using the following command:

```
plink --bfile mydata --het --out mydata_het
```

The PLINK tutorial also advises that if an individual has fewer homozygotes than expected by chance, this could reflect sample contamination, showing a strongly negative F-value. Extreme values (suggested values: <3 standard deviation units from the mean[11]) are considered outliers and should be removed from the data set.

Excess of homozygosity: Alternatively, samples can also show an excess of homozygosity, carrying values more than three standard deviation units from the mean.[11] These values can be an indicator of population substructure. Samples with excess of homozygosity are maintained in the study. It should be noted that these parameters are population specific; population isolates and level of inbreeding or migration into a given population will affect the acceptable values for this parameter.

Gender information: Phenotype data collected before commencing the GWAS analysis should contain gender information for every sample. Nevertheless, PLINK can check that gender data are consistent with the genotypes by using data from chromosome X to determine sex:

```
plink --bfile mydata --check-sex --out mydata_sex
```

When there is an inconsistency between sex determined in input pedigree files and sex obtained by X-chromosome analyses, PLINK highlights this as a PROBLEM. It uses a threshold of $F > 0.8$ for males and < 0.2 for females.

Such discrepancy can also arise owing to chromosomal abnormalities, such as Turner or Klinefelter syndrome, mosaicism or females with long stretches of loss of heterozygosity,[10] and thus results from this procedure need to be critically evaluated. GenomeStudio (Illumina) shows X-chromosome results per sample, and it is a visual way of identifying gender mismatches, as well as other alterations in females.

Population stratification: One of the key tests to perform in a GWAS is the population stratification test. Although some GWAS studies tend not to have strong population stratification, fine-scale genetic substructure is quite likely, and it can lead to spurious results. Population stratification can be prevented in various ways, the most common of which is a principal components analysis (PCA), which essentially uses all of the genetic data to compute the relatedness of every sample to every other sample. PLINK contains some tools to perform this test, using pairwise IBS (identity-by-state) distances to create a relatedness matrix.

Calculating the IBS matrices may be time consuming, especially for studies with large sample sizes performed in low-powered computer systems.

The `--genome` script will perform IBS metrics:

```
plink --bfile mydata --genome --out mydata_IBS
```

The above is a computationally intense step and may benefit from parallelisation. PLINK's tutorial offers an example of how to perform this, by creating subsets of samples and running all

unique pairwise combinations in parallel with the resulting files combined in the end.

Once IBS distances have been calculated, group differences are obtained using the following script:

```
plink --bfile mydata --read-genome plink.genome
--ibs-test
```

This permutes cases to controls and recalculates some metrics giving 12 separate tests. PLINK recommends T1 as most appropriate for case–control studies.

Pairwise IBS metrics only need to be calculated once for a set of individuals and can be reused. Using `--read-genome` option, cluster analysis can be run multiple times including different constraints, such as pairwise population concordance (`--ppc 0.0001`) or maximum cluster size (`--mc 2`). PLINK's tutorial provides a list of arguments to be used.

Group relatedness in the sample may be visualised using R or PLINK, using the proposed command:

```
plink --file mydata --read-genome plink.genome
--cluster --mds-plot 4
```

In the output file, parameters C1 to C4 correspond to principal components 1 to 4. PCAs are measures of particular sample collections and are not invariable properties of the samples. They are meant to show the relative relationship of samples. Plotting each principal component against the other will offer a scatter plot with a point per individual (**Figure 2**).

PLINK's tutorial advises to detect which individuals are outliers by calculating a sample mean and variance (transforming this measure into a *Z*-score). Extreme *Z*-scores (usually $<$ or >4 standard deviation units) are considered outliers.

Cryptic familial relationships: The IBS matrix can also be used to detect cryptic familial relatedness. In a homogeneous sample (individuals are similar, after removing IBS outliers) with a large number of SNPs available, IBS can be used to calculate genome-wide IBD (identity-by-descent) information; for this, we use the `--genome` command as indicated above.

Barring contamination during DNA sample handling stages before the genotyping, PI_HAT values above 0.1 indicate a definite and close familial relationship between that pair of samples, and therefore one sample per pair should be excluded from the analysis (unless software that can accommodate for family structure is used for the GWAS). Deciding which sample will be removed from the pair showing familial relationship could be based on user criteria—that is, removing the one that minimises the sample number loss or phenotype availability and interest.

PI_HAT values above 0.1 could also indicate contamination, and the samples involved need to be further investigated.

Occasionally, especially in inbreed populations, SNPs will be strongly correlated with each other, which may bias the IBS analysis. In these circumstances, the need to select a subset of SNPs that are not correlated (or rather are only weakly correlated) with each other arises. The software tutorial recommends applying this test to a subset of SNPs in linkage equilibrium, using an r^2 threshold of, for example, 0.2:

```
plink --bfile mydata --indep-pairwise 50 5 0.2 --
out mydata_IBS
```

This creates one list of the SNPs to be retained in the analysis (mydata_IBS.prune.in) and another for those to be excluded (mydata_IBS.prune.out).

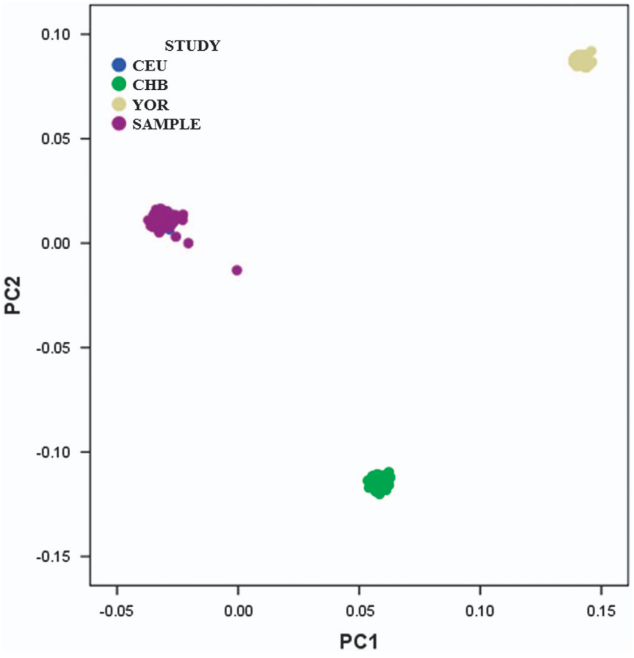

Figure 2 Example of scattered plot in SPSS (IBM Corp., IBM SPSS Statistics for Windows, Armonk, NY, USA) for the PC1 versus PC2, showing some outliers from the Caucasian population (blue dots = CEU, green dots = CHB + JPT, cream dots = YOR, purple dots = study samples).

SNP-level QC. It is recommended that at the end of the QC process a clean dataset is created to be used for future analyses, to avoid differences due to shifting QC measures that future users may adopt. As for the individual-level QC steps described above, SNPs can be filtered from the data set directly using various arguments (e.g., MAF threshold), or using a list of SNP IDs to be removed with the `--exclude` argument in PLINK, or to be retained using the `--extract` argument.

Genotyping efficiency: SNPs are tested for their genotyping efficiency (call rate), and those that fail in a large number of samples should be removed because they may produce unreliable results. The recommended threshold for call rate is 98–99%.[10] This stringency could be reduced for small sample size studies. SNPs with low call rates can be removed using the following command:

```
plink --bfile mydata --geno 0.01 --make-bed --out
mydata_geno
```

Hardy–Weinberg equilibrium (HWE): Deviations from HWE may indicate systematic genotype miscalling or population stratification,[12] but they may also be owing to intense natural selection pressures of the variant associated with the trait under study.

HWE can be assessed using the following command:

```
plink --bfile mydata --hardy --out mydata_hardy
```

The tutorial shows that for a case–control study each SNP will have a test for ALL, AFF (cases only) or UNAFF (controls only). For quantitative traits, only ALL(QT) will appear for SNP. This procedure will only consider founders in the sample—that is, individuals from whom the pedigree files specify that the father and mother are unknown and not included in the sample. It will not take into account the relatedness between two individuals, as calculated by PI_HAT or any other method.

In case–control studies, it is important to examine HWE in controls separately (considering that some true associations are expected to be out of HWE). If there are multiple populations included in the study, HWE needs to be analysed in each population separately.

Once the HWE information is obtained per SNP, a specific threshold can be set, and those SNPs failing it can be removed, using `--hwe` command. Popular exclusion threshold is $P < 10e-5$:[13]

```
plink --bfile mydata --hwe 0.00001 --make-bed --out mydata_hwe
```

Information for failing SNPs will be provided for cases and controls separately:

```
Writing Hardy-Weinberg tests (founders-only) to [ mydata.hwe ]
30 markers failed HWE test ( p <= 0.00001 ) and have been excluded
34 markers failed HWE test in cases
34 markers failed HWE test in controls
```

In a case–control study, this test is based on controls only.

MAF: Statistical power to detect associations is extremely low for rare SNPs, which, especially for older chips, are also more susceptible to biases in genotype calling;[10,14] thus, these SNPs can be removed using the following command:

```
plink --bfile mydata --maf 0.05 --make-bed --out mydata_maf
```

In the above case, alleles with an MAF lower than 0.05 will be excluded. PLINK sets a default value of MAF = 0.01. The threshold selected depends on the size of the study and the effect size expected. There are some tools such as CaTS Power (http://csg.sph.umich.edu/abecasis/CaTS/index.html) or Quanto (http://biostats.usc.edu/Quanto.html) that provide a frequency below which the study is underpowered,[15] but these calculations should also take into account the fact that power will differ if these results are intended to contribute to larger meta-analyses consortia or stand-alone publications.

Batch effects. Batch effects must be taken into account when combining samples from different platforms or those that have been processed in different laboratories. Turner *et al.*[10] suggest a simple approach to detect batch effects, by testing differences in average MAF and genotype call rate for the same SNP in each plate. They also propose that testing one plate against the others in a GWAS analysis can also be performed. Although it is more time consuming, it gives a clear impression of any batch-related significance. To perform this, one of the batches or plate is coded as case and compared with the rest of the batches, coded as controls. A simple association test should be performed to detect any deviation from the expected uniform distribution of the *P*-values. The same procedure should be applied to all the batches. Should moderate batch effects be detected, they can be treated in the same way as population stratification.[10]

Association testing. A standard post-QC analysis protocol is summarised in **Figure 3**.

In a case–control setting, a basic χ^2 analysis comparing allele frequencies between cases and controls can be performed using the following command:

```
plink --bfile mydata --assoc --out mydata_assoc
```

Confidence intervals can also be computed using the argument `--ci X`, where X may usually be 0.95 or 0.99.

PLINK's tutorial also provides information on association tests under other genetic models, including the Cochran–Armitage trend test, a 2 d.f. genotypic test and tests under a dominant or recessive (1 d.f.) model. The results for all of these tests can be obtained using the argument `--model` instead of `--assoc`.

Exact test statistics, using Fisher's exact test, can be obtained using the argument `--fisher` instead of `--assoc`, or using `--fisher` in addition to `--model`.

PLINK can also perform stratified analyses, when a cluster variable is specified (using the `--within` command). Tests for both, overall disease/gene association and heterogeneity of the disease/gene association, can be performed, taking into account this clustering. Selecting the right stratified analysis will depend on the cluster structure, either a small number of clusters with a large number of cases and controls or a very large number of clusters with small number of individuals per cluster, as indicated in the tutorial. Statistical tests include Cochran–Mantel–Haenszel, Breslow–Day of homogeneity of odds ratio or partitioning the total association χ^2 to perform

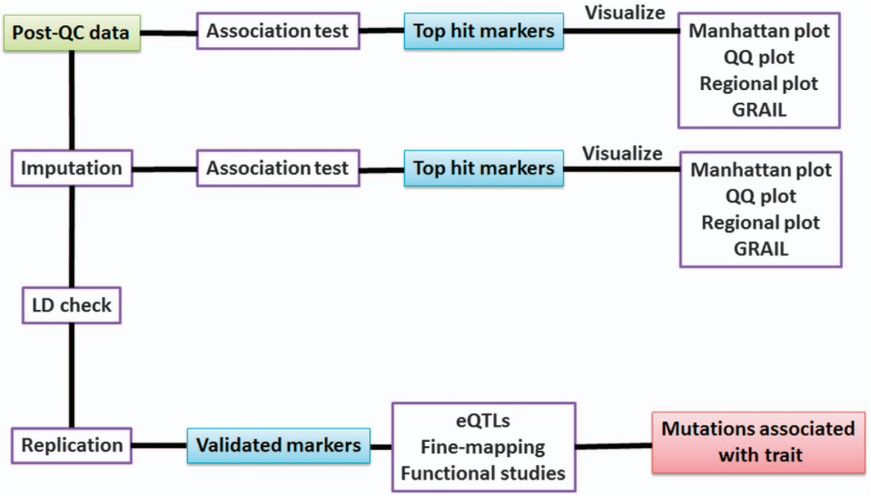

Figure 3 Data analysis flowchart.

between- and within-cluster association.

One of PLINK's most powerful analysis options is its ability to fit regression models, using the `--linear` or `--logistic` argument for quantitative and dichotomous/disease response phenotypes, respectively. Covariate data are provided in a separate text file (read with `--covar` command), and this option can also be combined with a separate phenotype data file, using the `--pheno` argument, which allows the user to fit a range of complex models against a range of response phenotypes, using the following commands:

```
plink --bfile mydata --linear --pheno phe.txt
--covar cov.txt --out mydata
plink --bfile mydata --logistic --pheno phe.txt
--covar cov.txt --out mydata
```

For the test, information will be provided for additive effects of allele dosage and results for each covariate.

Finally, it is possible to include an adjustment for multiple testing, such as the Bonferroni correction. The command below will generate a file with *P*-values corrected for the total number of tests performed:

```
plink --file mydata --assoc --adjust --out
mydata_adjusted
```

All command lines were obtained from PLINK's tutorial. It also offers detailed information on different association tests.

A conventional *P*-value threshold for GWAS association is 5×10^{-8} in samples of European ancestry,[5] but a more stringent threshold may be required for samples of African ancestry, owing to their greater genetic diversity.[6]

Family-based GWAS analysis. Some family-based GWAS analysis can be carried out using PLINK (trios), but for others software such as QTDT and MERLIN will be needed.

PLINK: This software allows the user to analyse family-based samples using the TDT (transmission disequilibrium) test for linkage-given association. For an SNP, TDT analyses parents who are heterozygous for a variant and checks whether this SNP has the same frequency among the inherited alleles compared with the noninherited ones. The TDT test is not affected by population stratification. To perform the analysis, the following PLINK command can be used:

```
plink --bfile mydata --tdt
```

PLINK also facilitates analysis of parent of origin, separating heterozygous fathers from heterozygous mothers, using the following command:

```
plink --bfile mydata --tdt --poo
```

QTDT software: This package contains a number of modules that facilitate TDT analysis and a variety of association analyses. The software and its modules can be used under various scenarios and are highly flexible. However, the output is not well suited to a large-scale analysis.

MERLIN software: This package does not correct for population stratification, and thus if the user is concerned about stratification in their samples then this parameter should be included as covariate, or genomic control methods should be applied. However, MERLIN is a widely used software that accommodates family structure into the analyses. Its main advantage over alternatives such as GenAbel is its immediate compatibility of its input files with PLINK. MERLIN requires three files: a PED file, which is identical to that used by PLINK; an MAP file, which is a three-column file that can be obtained through simple editing of PLINK's MAP file; and a DAT file, which is essentially the list of SNPs names (or phenotypes) in the exact order in which they appear in the PED file, preceded by the letter 'M' ('A' for binary diseases or 'T' for quantitative traits).

Imputation. Imputation is an essential tool for maximising the information obtained from GWAS data. GWAS studies are generally based on the genotyping of 300 000–2.5 million variants throughout the genome,[16] which is only a small fraction of the many millions of variants across the genome. To improve genome coverage, imputation allows the genotypes of SNPs that are not present in the SNP chips to be estimated on the basis of genetic linkage and founder haplotype mapping studies. In this case, genotyped markers are phased and compared with a reference panel such as from the HapMap or 1000 Genomes projects (among others). Identification of shared haplotypes allows imputing the missing genotypes according to the reference panel.[17] Multiple tools are available for imputing missing genotypes and nongenotyped SNPs, such as IMPUTE,[16,18–21] MACH and fastPHASE/BIMBAM, which consider all genotypes and are more accurate for rare polymorphisms. Other and less used options include PLINK, TUNA and BEAGLE.[17]

SHAPEIT2 software:[22,23] This tool is very useful for phasing haplotypes before using IMPUTE2 for imputation. The software accepts diverse types of input files, including PLINK files (PED/MAP or BIM/BED/FAM). Before phasing can be performed, genotype data must be split into separate chromosomes, which can be done using the `--chr` argument with `--recode` or `--make-bed` in PLINK.

The following script, obtained from the tutorial, is used to specify the input files (PLINK input files, as an example):

```
shapeit --input-ped chr20.unphased.ped
chr20.unphased.map -M chr20.gmap.gz --output
-max chr20.phased
```

`--input-thr` option allows the user to change the threshold for uncertainty in the genotypes. The default value is 0.9.

This tool performs some data checking, including tests for individuals or SNPs with >5% missing data, detection of singleton SNPs and detection of completely missing SNPs or individuals. The software tutorial indicates how to modify some algorithm parameters, including the number of threats (`--threat n` option) and the default number of MCMC iterations (`--burn X, --prune Y, --main Z`). Model parameters, including the 'number of conditioning states on which haplotype estimation is based' (`--states`), window size (`--window`), genetic map of recombination rates (`--input-map`) and effective population size (`--effective-size`) can also be modified.

`--chrX` command can be used to phase data from chromosome X.

A key point before imputation is alignment of the physical positions of SNPs in the GWAS data to those in the reference panel. Most recent reference panels use the Single Nucleotide Polymorphism Database (dbSNP) build 37 coordinates, and if the GWAS data are based on an older build, the LiftOver tool can be used to translate these positions to build 37. The following command, as shown in the tutorial, can then be used to check the alignment of SNPs:

```
shapeit -check -B chr20.unphased --input
-ref chr20.reference.hap.gz chr20.reference.
```

```
legend.gz chr20.reference.sample --output-log
chr20.alignments
```

To list all alignment problems, the tutorial recommends the use of the following command in Linux:

```
cat chr20.alignments.snp.strand|grep 'strand'
```

Once GWAS data are correctly aligned, the next step is to phase the chromosomes. The software recommends phasing whole chromosomes, instead of making chunks, using the following command:

```
shapeit -B chr20.unphased -M chr20.gmap.gz
-O chr20.phased
```

IMPUTE2 is the recommended tool for imputation after phasing with SHAPEIT2.

The website http://mathgen.stats.ox.ac.uk/impute/pre-phasing.with.SHAPEIT_IMPUTE2.html offers an example on how to use SHAPEIT and IMPUTE2 to perform a prephasing imputation, including some files that are publicly available.

Further information about the use of SHAPEIT and the complete tutorial can be found at https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#home.

IMPUTE:[19] Imputation with one-phased reference panel (pre-phasing) is the most common imputation scenario when imputing non-genotyped SNPs into a study from a reference panel. The following commands from the IMPUTE website show how to perform the analysis in IMPUTE v.2:

```
impute2
-m ./Example/example.chr22.map \
-h ./Example/example.chr22.1kG.haps \
-l ./Example/example.chr22.1kG.legend \
-g ./Example/example.chr22.study.gens \
-strand-g ./Example/example.chr22.study.strand\
-int 20.4e6 20.5e6 \
-Ne 20000 \
-o ./Example/example.chr22.one.phased.impute2
```

IMPUTE tutorial provides detailed information of each argument.

Once all parts of the chromosome have been imputed, the `cat` command (in LINUX computer systems, or 'cp' in DOS-based ones) allows the user to merge the chunks into a file for the whole chromosome.

The IMPUTE tutorial offers a list of scripts to perform the analysis in IMPUTE v.2 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#home). It also contains a series of examples to illustrate the use of this software.

Haplotype reference and haplotype legend files, as well as recombination rates and strand reference data, are provided on the IMPUTE website.

Imputation can be speeded up by prephasing, thus splitting the imputation process into two parts: (1) phasing the chromosomes to be imputed and (2) imputing these from the reference panel. IMPUTE tutorial provides the scripts used for both steps:

Step 1: Pre-phasing:

```
impute2
-prephase_g \
-m ./Example/example.chr22.map\
-g ./Example/example.chr22.study.gens \
-int 20.4e6 20.5e6 \
-Ne 20000 \
-o ./Example/example.chr22.prephasing.impute2
```

Step 2: Imputation into prephased haplotypes:

```
impute2
-use_prephased_g \
-m ./Example/example.chr22.map \
-h ./Example/example.chr22.1kG.haps \
-l ./Example/example.chr22.1kG.legend \
-known_haps_g    ./Example/example.chr22.pre-
 phasing.impute2_haps \
-strand_g ./Example/example.chr22.study.strand\
-int 20.4e6 20.5e6 \ #values similar to prephase
step
-Ne 20000 \
-o ./Example/example.chr22.one.phased.impute2
-phase
```

The IMPUTE tutorial provides further information about each argument. This method is also useful when imputing into different reference panels, because chromosomes can be prephased once and stored for multiple uses later on.

Similarly to imputation, IMPUTE2 tutorial shows the steps to follow for pre-phasing and provides examples.

MACH 1.0 software:[17,24] This tool requires Merlin-data-format input files, as well as a pedigree file. Reference haplotype panels are based on HapMap and 1000G data. The first step is to build a model to relate the GWAS data to the reference haplotypes. The most important considerations are the number of iterations used to estimate the model parameters (`--round`) and the number of individuals in the sample. Then, the following script, containing a marker list file, a linkage-format genotype file and a haplotype reference and legend file, can be used to build the model:

```
mach1 -d gwas.dat -p gwas_subset.ped -s hapmap.
legend -h hapmap.phased -hapmapFormat -greedy -r
100 -prefix step1
```

The second step consists of carrying out the genotype imputation, using the following script:

```
mach1 -d gwas.dat -p gwas.ped -s hapmap.legend -h
hapmap.phased -hapmapFormat --crossover step1.rec
--errormap step1.erate --greedy --mle --mldetails
--prefix step2
```

`--mle` indicates that maximum-likelihood genotype imputation should be carried out.

Both scripts have been extracted from the MACH 1.0 tutorial. These guidelines also provide information about the output files: *.mlgeno, *.mldose, *.mlqc and *.mlprob files.

MACH2DAT software can use MACH data to run an association test for quantitative and qualitative traits, requiring *.ped and *.dat files in Merlin format, and using the following script, provided in the Mach2dat wiki (http://genome.sph.umich.edu/wiki/Mach2dat:_Association_with_MACH_output):

```
mach2dat -p myfile.ped -d myfile.dat --infofile
myfile.mlinfo --dosefile myfile.mldose
```

Minimac software:[16] This tool is an implementation of MACH that runs with lower memory and is able to handle thousands of haplotypes.[16] Minimac tutorial describes the software workout as 'involving an initial step to estimate haplotypes for the entire sample and then imputing missing genotypes using the reference panel'.

This tutorial recommends the first step to be run in MACH, using the following command:

```
mach -d sample.dat -p sample.ped --rounds 20
--states 200 --phase --interim 5 --sample 5
--compact
```

To perform the imputation step, the tutorial specifies that 'Minimac requires a file listing the markers, extracted from the second column of the *.dat file'. Then, the software will work, based on MACH-selected haplotypes from the previous test, using the following script, as described in the tutorial:

```
minimac --refHaps ref.hap.gz --refSnps ref.snps.gz --haps target.hap.gz --snps target.snps.gz --rounds 5 --states 200 --prefix results
```

It is also possible to impute chromosome X using Minimac, although the X-chromosome pedigree file first needs to be split by sex. Minimac tutorial provides a simple description of the protocol to follow.

BC|SNPmax software: This software can be purchased from the BC Platforms Ltd (Esbo, Finland), which also offers imputation as a service. BC|SNPmax is a database platform (https://www.bcplatforms.com) that provides tools for integrated genetic and clinical data management and analyses with a queue system that enables segmentation of large analyses. The BC|SNPmax system has user-friendly interface tools for data preprocessing, alignment, variant calling, data cleaning and epidemiological data analyses. In the genotype imputation workflow, BC|SNPmax supports programs MACH, SHAPEIT and PHASE in prephasing. For imputation, users can use programs IMPUTE, BEAGLE and MINIMAC. The system has integrated dbSNP marker maps and consequent reference panels. The results of imputation can be analysed using a variety of most commonly used GWAS and linkage analysis programs such as PLINK, SNPTEST, MACH2QTL, Eigenstrat and ProbABLE.

In BC|SNPmax, PLINK files can be easily used as input files, and they must be identified as PLINK files when uploaded to the system. SNPs for imputation can be selected using PLINK parameters, such as `--maf` (MAF), `--max-maf` (maximum MAF), `--geno` (maximum per-SNP missing), `--mind` (maximum per-person missing) and `--hwe` (Hardy–Weinberg disequilibrium *P*-value). When both MACH and MINIMAC are used for imputation, the first step consists of phasing the data using MACH. It is necessary to identify the dbSNP build used for phasing (build 36 or build 37, as above), and the tool offers the option to fragment the chromosomes into the desired number of SNPs. All of the parameters supported by MACH can also be used here (`--states, --rounds, --interim, --sample, --compact`). Phased haplotypes and pedigree information should be uploaded into the BC|SNPmax data set for imputation, and imputation will then be performed using Minimac and R postprocessing. The interface shown in **Figure 4** will be displayed, allowing selection of the dbSNP build for imputation, the number of chromosomes to be included in the analysis, the reference panel (HapMap or 1000G) and the population. As for the previous phasing step with BC|SNPmax, any of the parameters used with MACH can be applied here. A final QC test is run after imputation, and markers with an r^2 value of <0.3 should be removed.[24]

BEAGLE software.[25–27] The working guidelines provided with the software describe all steps to perform the analysis. We summarise this process below:

Input files for this software contain markers in rows and individuals in columns, with every allele in a different column. An initial column will indicate 'I' for indicator, 'A' for affection status and 'M' for marker. Genotypes can be provided unphased or phased. For GWAS data, each chromosome should be phased separately using the following command:

```
java -Xmx1000m -jar beagle.jar unphased=fileA.bgl phased=fileB.bgl markers=markers.txt missing=? out=example
```

BEAGLE will impute nongenotyped markers in the study file (file A) based on a reference panel (file B). The tutorial explains each argument of the script and offers a number of parameters that can be included in the script, such as `niterations=`<number of iterations>, `nsamples=`<number of samples>, `excludecolumns=`<excluded columns file>, `excludemarkers=`<excluded markers file> and so on. Once phasing is completed, a *.log file summarising the process and a *.phased.gz file are created. BEAGLE can also perform an association test for haplotypes, using the following command:

```
java -Xmx800m -jar beagle.jar data=data.bgl trait=T2D out=example
```

As previously mentioned, the BEAGLE tutorial includes detailed information about each of the arguments in the script. All commands have been extracted from the BEAGLE tutorial.

Meta-analysis. This approach allows the researcher to increase the number of samples analysed, as well as the number of markers, by combining results from different GWAS studies.[1] Data for SNPs that have not been genotyped in one or more studies are inferred through imputation.[28] Thus, meta-analysis is a powerful tool for discovering variants with small effects or those that are very rare, which require large sample sizes.

All included data sets should use the same parameters and definitions for each variable, and samples must be unrelated. To avoid any relatedness between individuals, it is possible to correct the number of times that the χ^2 statistic for the association is inflated for the inflation factor λ,[1] which is considered normal with values ~ 1.[29] Heterogeneity in meta-analyses may occur when phenotypes are difficult to assess and standardise, or due to different ancestry, and can be reported as a value of Cochran's *Q* (significant when $P<0.10$). When meta-analyses include a large number of studies (typically 10 or more), heterogenetity is quantified using I^2, a parameter that shows overlap (or rather lack of it) between results of individual studies. It ranges from 0 to 100%, where 0–25% is considered to reflect low heterogeneity and >75% is considered to reflect very strong heterogeneity.[30] If the number of studies included in the meta-analysis is rather small, a *P*-value for the differences in estimated effect sizes of each population may be calculated instead.

Meta-analysis assigns a weight to each study result, whereby those studies with greater precision (typically proportional to sample size) are given higher weights. The most commonly used meta-analysis method is the fixed-effect approach, which assumes that there is no between-study heterogeneity, and meta-analysis is performed to increase power. Inverse variance weighting is the most common model.[1,31] The Cochran–Mantel–Haenszel approach is an alternative method that returns very similar results. A random-effect meta-analysis assumes that the effect varies across studies following a normal distribution (reviewed in Evangelou and Ioannidis[13]).

The *METAL* software described below is a very useful tool for performing meta-analysis based on GWAS data.[32]

Input files: Each input file should include the marker name, the coded allele and the other allele. Sample size-weighted

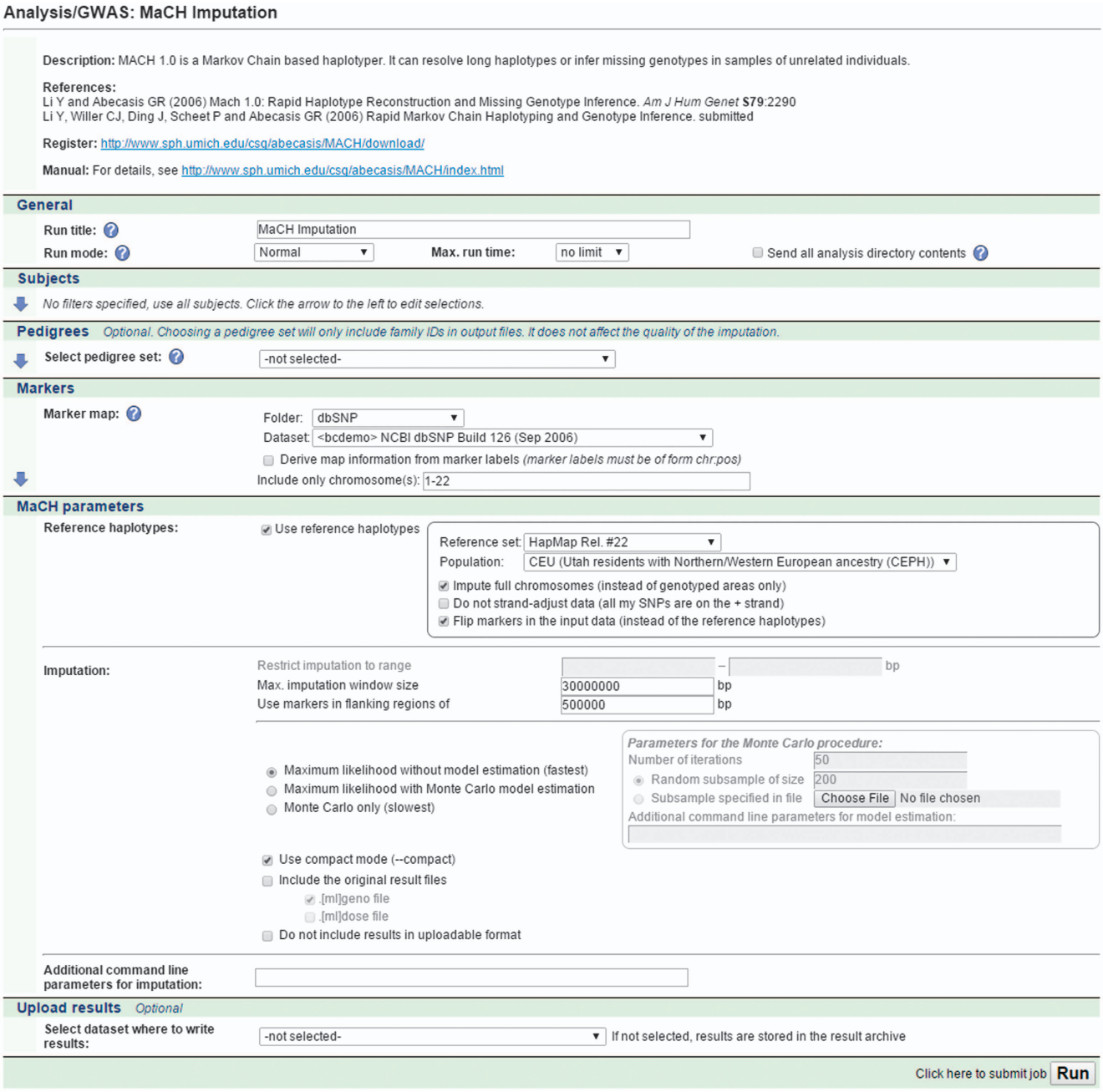

Figure 4 Interface of BCSNPmax software for imputation step, showing the different parameters available for the analysis.

analyses also require the direction of effect of the tested allele, corresponding *P*-value and sample size. If meta-analysis is based on standard errors, then the estimated effect size for each marker and standard error are required.

Analysis: The meta-analysis can be executed using the ANALYZE command.

Additional options: Other commands can be used to customise the meta-analysis according to the characteristic of the studies involved, including alternatives to the *P*-value analysis, genomic control for population stratification, strand used, filtering of SNPs for analysis, inclusion of noncomplete lines, tracking allele frequency and other options.

METAL tutorial provides more information about the above-mentioned steps, as well as an example of a METAL script.

Linkage disequilibrium. GWAS results need to be checked to detect whether the trait-associated SNPs are in linkage disequlibrium (LD), and therefore, containing the same genetic information. Regional LD can be examined using the Haploview[33] tool (http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview).

The first step includes selecting the HapMap download and the corresponding genomic positions for the SNP to test, within the specific version of HapMap.

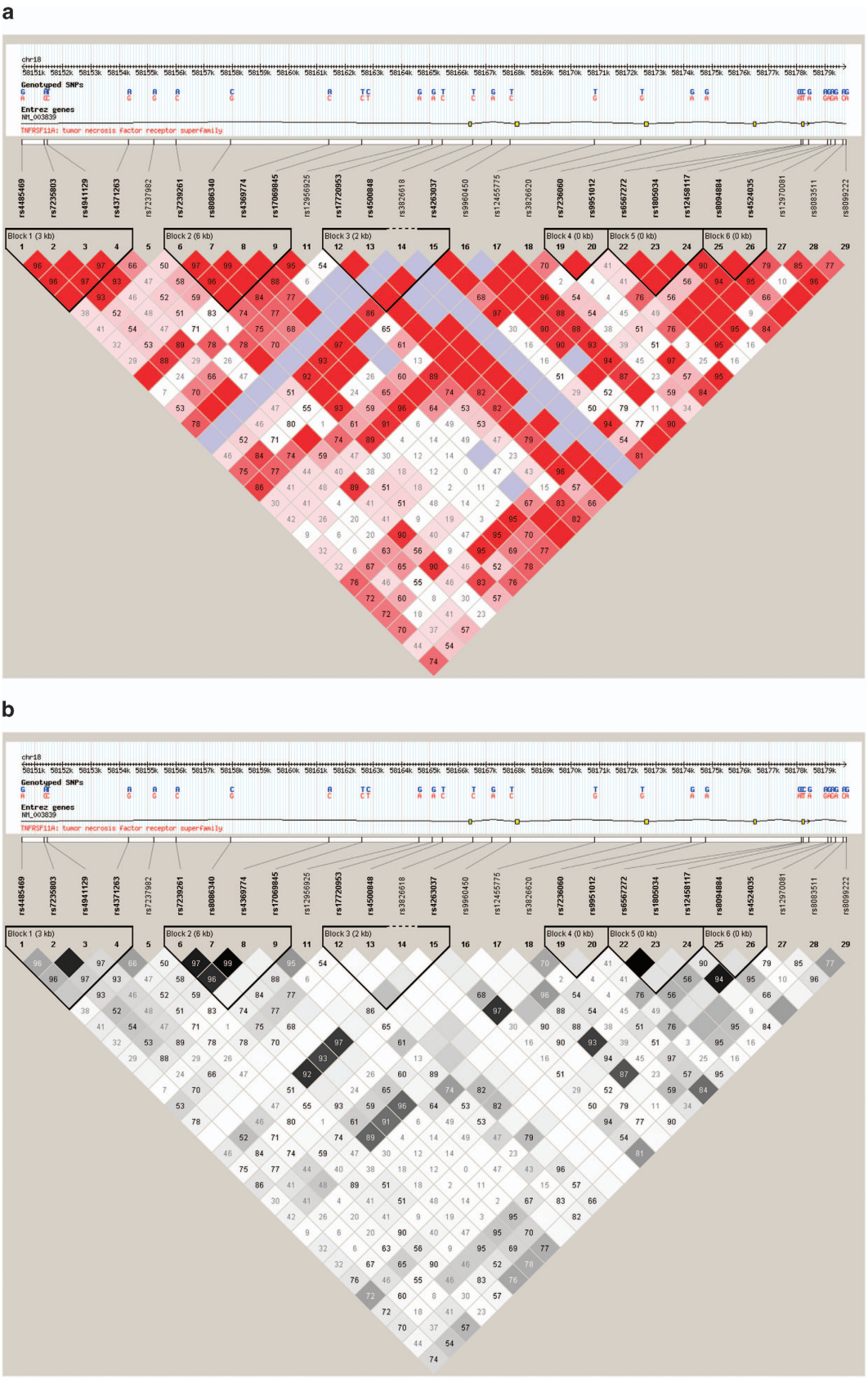

Figure 5 LD blocks for the selected area (TNFRSF11A as example) displaying results for (**a**) D' and (**b**) r^2 figures.

The software will show the LD blocks within the highlighted area, as shown in **Figure 5**. The two commonly used measures for LD are D' and r^2. D' ranges between 0 and 1, where 1 reflects complete LD. R^2 is the square of the Pearson's correlation between the two genotypes; an $r^2 > 0.8$ indicates that the two SNPs in question convey very similar information.[9]

The same results as shown in the LD plot are also provided in a table, where it is possible to select one SNP and obtain a list of SNPs in LD with it. In this case, the r^2 threshold can also be set, as well as the SNPs in the data set to be tagged and the aggressiveness of the tagging algorithm.

As mentioned in the section on Cryptic Familial Relationships (section Individual-level QC), LD can be used for SNP pruning, selecting only non-redundant SNPs from a region with a dense SNP data. Further information can be obtained in the Haploview tutorial.

Graphical representation of GWAS results. Manhattan plot (**Figure 3**): This provides a graphical representation of the results of a GWAS analysis; the $-\log_{10} P$-values for association are plotted on the y axis and the physical position of the corresponding SNPs are plotted on the x axis.

The simplest way to create a Manhattan plot is using Haploview, on the basis of the results obtained from the association test.

Haploview offers the option to upload PLINK-formatted result files, allowing the user to browse the association results file. The Integrated MapInfo option should be selected. Further plot options allow the user to enter the scale of the y axis ($-\log_{10}$) and the thresholds for significance and suggestiveness.

A Manhattan plot can also be created using R. As Haploview cannot handle imputed data, this is the only option available. S Turner proposes a simple pipeline on his blogspot 'Getting Genetics Done' (http://gettinggeneticsdone.blogspot.co.uk):

```
>source("http://dl.dropbox.com/u/66281/0_Permanent/qqman.r")
>mydata=read.table("mydata.dat", header=TRUE)
>manhattan (mydata, colors=c("#FF6A6A", "#3A5FCD", "#E066FF", "#FFC1C1", "#878787", "#CD0000", "#00008B", "#32CD32", "#CDCD00", "#8B008B", "#00EEEE", "#4D4D4D", "#FF4040", "#3A5FCD", "#EE82EE", "#ADADAD", "#8B1A1A", "#00008B", "#006400", "#8B8B00", "#8B008B", "#00868B"))
```

Either using Haploview or R software, a Manhattan plot similar to that shown on Albagha *et al.*[34] can be displayed.

Quantile–quantile plot (Q–Q plot) (**Figure 3**): This is a graphical technique that displays the distribution of the results of a GWAS analysis. Expected *P*-values are plotted against observed *P*-values for each SNP under the null hypothesis of no association. True associations that deviate from the expected results will appear in the tail of the distribution. A Q-Q plot can be created using R,[35] for example, as shown in Albagha *et al.*:[34]

```
>pvals <- read.table("pvals.txt", header=T)
>observed <- sort(pvals$PVAL)
>lobs <- -(log10(observed))
>expected <- c(1:length(observed))
>lexp <- -(log10(expected / (length(expected)+1)))
>pdf("qqplot.pdf", width=6, height=6)
>plot(c(0,7), c(0,7), col="red", lwd=3, type="l", xlab="Expected (-logP)", ylab="Observed (-logP)", xlim=c(0,7), ylim=c(0,7), las=1, xaxs="i", yaxs="i", bty="l")
>points(lexp, lobs, pch=23, cex=.4, bg="black")
```

The script above has been extracted from http://www.inside-r.org/questions/generating-data-frame-based-qq-plot.

Regional plot (**Figure 3**): After finding SNPs associated with the trait of interest, it may be necessary to plot the GWAS results in the surrounding region, which can be done using LocusZoom (http://csg.sph.umich.edu/locuszoom/). LocusZoom can also plot information about LD for SNPs in the selected region (based on HapMap and 1000 Genomes Project data), and gene information from the USCS browser.[36] Albagha *et al.*[37] shows some examples of a typical result from LocusZoom, indicating SNPs within a recombination area (delimited by the peaks). Colours distinguish the strength of LD between each SNP and the selected marker.

Forest plot (**Figure 3**): This type of graph can be used to display the results of a meta-analysis. The effect size estimate (mean or odds ratio) for each study is represented by a box, whose size is proportional to the weight of the study in the overall analysis, and a line corresponding to the 95% confidence interval is also shown. The pooled effect size estimate across all studies is usually represented by a diamond at the bottom of the graph, whose width corresponds to the confidence interval of the pooled estimate. The position of no effect is shown as a vertical line crossing the x axis 1 (for odds ratios) or 0 (for means).

A forest plot can easily be created using R (the manual will help with the arguments included), as shown in Albagha *et al.*:[37]

```
>trials <-read.table ("mydata_forest_plot.txt", as.is=TRUE, header=TRUE)
> library(rmeta)
> attach(trials)
> metaplot(OR, se, labels=cohort, logeffect=FALSE, conf.level=0.95, xlab="Odds ratio", ylab="Study Reference", zero=1, summn=1.50, sumse=0.038, sumnn=692.52, colors=meta.colors(box="magenta", lines="blue", zero="red", summary="orange", text="forestgreen"))
>grid.text("Forest_plot_meta-analysis", .5, .9, gp=gpar(cex=2))
```

Note: Values for `sumn`, `sumse` and `sumnn` on the script above are random data.

GRAIL software[38] (**Figure 3**): This software establishes relationships between genes in different loci associated with a specific trait. The GRAIL tutorial offers some guidelines about the process: Input data consist of a selection of SNPs (listed as rs numbers, based on HapMap release 21 or release 22) from the GWAS, or a series of genomic regions (labelled as ID CHR START(bp) END(bp)). It is recommended that every region represent a unique gene and do not overlap. Query regions and seed regions need to be determined for the analysis. Seed regions are high-confidence associations, and query regions are those to be evaluated. GRAIL assigns a *P*-value to each region depending on its connectivity and picks the candidate gene. VIZ-GRAIL is required to visualise the results. Guidelines

for running grail, the files needed to create the input files and the graphic and Perl scripts can be found at http://www.broadinstitute.org/mpg/grail/vizgrail.html. An example of results obtained using GRAIL can be found in Estrada *et al.*[39]

Genetic risk scores. Individual SNPs identified by GWAS generally confer only a modest disease risk for a complex disease, and thus it may be useful to model the combined effects of multiple SNPs to explore the distribution of their effects in the general population. This can be done by constructing a genetic risk score (GRS), which is the number of risk-modulating alleles at a series of SNPs associated with the phenotype of interest.[40] These SNPs can be selected using a 'best-guess' approach, which includes SNPs in or near genes that are known to be relevant for the phenotype of interest, or using a 'top-hit' approach, which includes SNPs that are strongly associated with the phenotype of interest, regardless of what is known about the biological mechanisms underlying the association. It has been also proposed to extract all SNPs associated with the phenotype at a given threshold, cluster them according to LD patterns and select the SNPs depending on their statistical significance and replication.[41] The R package PredictABEL provides a function, riskScore, for constructing a crude (unweighted) GRS or one weighted by the magnitude of the risk effects observed for each SNP in a GWAS or meta-analysis:

```
riskScore (weights, data, cGenPreds, Type)
```

`weights` are the observed beta coefficients for each SNP, `data` is the data set analysed, `cGenPreds` indicates which genetic variables are to be included in the GRS and `type` is used to indicate whether the GRS should be weighted or unweighted.

Exploiting publicly available GWAS data. The database of genotypes and phenotypes (dbGap, http://www.ncbi.nlm.nih.gov/gap) is a platform for depositing and sharing the results and raw phenotype, genotype and sequencing data produced by genome-wide studies. Detailed documentation about the study (description, design, history and publications), phenotype summary data, genotyping and sample QC data and an association results browser are available via open access. Individual-level phenotype and genotype data, as well as precomputed univariate genotype–phenotype association results, are available to legitimise researchers via controlled access.

This data repository offers a powerful and versatile way of increasing the power of genetic association analyses, performing exploratory studies or informing and optimising research planning.

Methods for NGS

Current NGS assays measure genotype variation across loci, but they generally do not re-assemble *de novo* the sequencing reads. At the time of writing, NGS technologies obtain information from short sequences (between 35 and a few hundred base pairs, depending on the technology used), which are overlaid to a preexisting library of the genome sequence of that particular species.

Consequently, NGS is more suited to detecting single-base variations in the genome of a species and less efficient for larger variations, whereas larger insertions, deletions or balanced re-arrangement of chromosomes are less likely to be detected. There are, however, many ongoing attempts to implement algorithms that would improve calling of more complex polymorphisms, which may be useful in the future.

Given the large number of sequences that can be aligned with a library and the potentially large number of variants within them, a variant is observed ('called') after an automated system makes a decision that the evidence for a polymorphism is sufficiently strong at a particular site. This decision is probabilistic in nature. To increase calling accuracy and to protect against technical artefacts, a certain experimental redundancy is needed; for example, the same region should be sequenced numerous times in separate experiments in the same individual, and a variant is called when it is observed in a sufficiently large number of them. This notion is referred to as 'sequencing depth'.

Study design. There are a few considerations that must be made before running the NGS assays—for example, the practical balance between a perfect depth of the NGS coverage and the still significant economic costs associated with it. Study designs will vary depending on the samples available and the purpose of the study. For example, if the target is a rare, perhaps Mendelian, disorder there is an incentive to sequence the few available individuals at high depth (60–80 $\times$ or more). However, if samples are abundant and finding variants with low frequencies present in the general population is the purpose, then sequencing thousands of subjects at low depth (e.g., 4–10 $\times$) may represent a reasonable compromise.

Early steps: sequence information and QC. Obtaining raw sequence reads, aligning them to libraries and calling the variants is often done using proprietary software, which depends on the sequencing platform used. Non-profit academic centres often prepare tools that are either alternative or complementary to those suggested by the NGS platform manufacturers (such as GATK by the BROAD Institute, Cambridge, MA, USA; **Table 1**).

There are many alternatives for sequencing QC, which partly reflects different views and expectations of the variations expected from a given assay. The most popular tool is currently the variant quality score recalibration (VQSR), which can be used separately or as part of the GATK suite.

The use of VSQR is a highly specialised skill, and it is out of the scope of this review. Its purpose is to assign a confidence score to each observed variant based on raw read depth, mapping quality, haplotype scores and similar indices. Using these data, VSQR creates Gaussian-distributed information scores for each variant that can be used to determine QC thresholds.

NGS data exploration and analysis. Typically, results come in one of two formats: VCF (http://vcftools.sourceforge.net/specs.html) or BED (https://genome.ucsc.edu/FAQ/FAQformat.html). A commonly used tool that can handle both formats is *vcftool*. This program cannot generate meaningful association results, but it is particularly useful for exploring the data and for generating summary statistics.

After obtaining and installing vcftools, it can be run in one of the two ways:

```
vcftools --vcf <input_filename.vcf>
```

or

```
vcftools --bed <input_filename.bed>
```

depending on whether the data are in VCF or BED format. Vcftools can be used to merge information contained in multiple input files, extract information from a large file and save it into a smaller file, or convert the data to other formats such as PLINK, IMPUTE and so on.

The results can be explored for a part of the data set (e.g., a chromosome by setting the filter `--chr`, or `--from-bp` and `--to-bp` for a region) or for the whole data set. It is also possible to remove all sites that do not have a high-enough quality score (`--minQ` filter) or depth (`--min-meanDP`) or missingness (`--geno`).

Vcftools includes other filters, such as:

`--maf` Minor Allele Frequency
`--mac` Minor Allele Counts
`--hwe` Hardy-Weinberg Equilibrium

Statistics that vcftools can generate include the following:

`--freq` generates minor allele related frequency information
`--counts` which reports the exact counts at each site
`--depth` reporting the mean depth per individual
`--site-depth` reporting depth by site
`--site-quality` which, as the name suggests, reports the quality of sequencing
`--hardy` reporting Hardy-Weinberg equilibrium
`--het` for the individual heterozygosity
`--TsTv` which calculates the Transition/Transversion ratio

It is noteworthy that several of these statistics have a relative value in terms of QC. For example, Hardy–Weinberg is expected to look normal for rarer alleles even in the presence of serious technical problems owing to lack of power to detect significant departure from equilibrium, whereas the Ts/Tv ratio at which a human genome-wide scale is about 2.1 will be a lot higher in cases in which the assays have targeted GC-rich regions of the chromosome, exons or can vary, higher or lower in other species.

NGS data analysis. Currently, a plethora of analytical tools are available for NGS data. There are two main approaches: single-variant and gene-based techniques. The single-variant methods are essentially applications of regression models, whereas the gene-based approaches look at accumulation of evidence, suggesting that a gene is involved in susceptibility to the phenotype of interest. Among many programs written to implement the analytical approaches, Plinkseq (http://atgu.mgh.harvard.edu/plinkseq/input.shtml) is the most flexible, and it implements most of the current approaches at the same time. This section will describe the use of Plinkseq and the ways it can implement these analytical algorithms.

The first step for running a Plinkseq analysis is to create a project. This is simply done by entering the following command:

```
Pseq <project_name> new-project --resources
<path_to_directory_containing_resources>
```

Note that before creating the new project it is advisable to create a directory containing all annotations and resources, which can be downloaded from the Plinkseq website.

At this stage, the project is just a skeleton, and it needs the actual NGS genotype data to become fully functional. New vcf files can be added to a Plinkseq project using the following command line:

```
pseq <project_name> load-vcf --vcf <input_
file_name.vcf>
```

and then the project will contain the basic information that is needed to run an analysis. Optionally, it is possible to load phenotypes into a project:

```
pseq <project_name> load-pheno --file <file_
containing_phenotypes>
```

Although the last step is not always necessary, it may be advantageous in certain circumstances (e.g., when covariate adjustment is required). The above commands will create large data sets, as large as the vcf files that contained the initial information. After the projects have been created, the original vcf files and phenotype data files can be deleted from the system to save space, whereas the resources will need to be accessible at all times.

Plinkseq is capable of performing association analyses for qualitative or quantitative phenotype variables. For the former case, the command is as follows:

```
pseq <project_name> v-assoc --phenotype
<phenol_name>
```

If at this point the genotype/phenotype files have not been incorporated into the project as described above (load-pheno), it is necessary to also specify the file:

```
pseq <vcf_file> v-assoc --phenotype <file_con-
taining_phenotypes> <phenol_name>
```

A quantitative analysis is run using the following command:

```
pseq <project_name> glm --phenotype <phenol_
name> --covar <covariate_1, Covariate_2 etc. >
```

Common gene-based analyses implemented in current studies include burden tests, case-unique analysis, variable threshold test, frequency-weighed test, C-alpha, summary of single site statistics test and, recently, Skat and Skat-o tests. In Plinkseq, it is possible to run one or more of these tests at the same time using the following command:

```
pseq <project_name> assoc --tests calpha vt fw
sumstat --phenotype <phenol_name> --mask
loc.group=refseq
```

This will run gene-based tests in which genes are defined as Refseq transcripts. The user can choose which tests to include after the `--test` parameter.

Interpretation of NGS association results. NGS data is different from GWAS data, and caution is required when interpreting the association results. The main difference is in the sheer number of extremely rare variants, which, unless filtered, may inflate the number of Type I errors in the results.

In many respects, this is a lesser problem in qualitative analyses, where it is possible to run exact tests and/or permutation procedures that efficiently control for false positives. However, there are no exact tests for quantitative analyses, and permutation procedures are very resource-intensive, and thus it is advisable to reduce the number of rare variants analysed as single variants by filtering those with allele frequencies that are unlikely to give any meaningful results. This will also contribute to relieving the burden of multiple testing.

It is currently unclear as to what the optimum whole-genome sequencing level of significance for association should be. The widely adopted 5×10^{-08} significance threshold of GWAS is based on a haplotype structure and LD distribution that is different from what we can expect for rare variants. As of the time of writing, no simulation studies addressing this issue have been published.

Another issue concerns the gene-based tests. Normally, the genes contain a large number of both common and rare variants. The former can dilute the effect of the rare variants and often would have been identified through GWAS. In practice, filtering the common variants is helpful, but the results become very dependent on the filtering thresholds and interpretation often becomes difficult. Another difficulty is the definition of the gene. Using the command line above, only transcription initiation through transcription termination sites will be accounted for. However, this definition of the gene is not necessarily suitable under all circumstances. Often *ad hoc* definitions, for example, coding variants only, or definitions including the promoter may be more successful for certain phenotypes. Finally, there is the added difficulty that transcription variants of the same gene may be counted separately. For example, it would be difficult to interpret a result when only some, but not other, alternatively spliced transcripts are associated with a given phenotype.

Three 'rules-of-thumb' for NGS results: replication, replication and replication. Given that NGS analyses are so recent and no comprehensive framework is universally accepted yet, it is important to have validation panels, either through internal resources or larger collaborations. This is far more critical for NGS studies than for GWAS, in which conventional significance thresholds are very conservative.[42] All results should be interpreted with extreme caution and taken as provisional when the validity is suspicious. Coding variants causing some rare syndromic disorder could be considered as possible exceptions, due to their reduced independent validation.

Yet, replication is not necessarily straightforward for NGS results, and there are some expected difficulties in replicating NGS results. Often, rare variants will display ethnicity specificity, that is, they will be relatively well-powered (present in sufficient numbers) in one population but not necessarily in another. It also remains to be seen whether ranking of variants according to their statistical probabilities of association is as efficient a tool as it was in the GWAS; standards error of estimated effect sizes of rare SNPs may depend less on total sample sizes than on common SNPs, making these estimates and, consequently, probability ranking less reliable. Finally, multiple rare but high-impact variants may be located within the same gene, and their effect may only be detected in the samples in which they happen to be enriched but not others.

Best practice guidelines for successful NGS analyses are likely to change in the near future, as more experience and data sets become available to the community.

Discussion

This protocol describes two methods for managing high-throughput data, such as GWAS and NGS. The crucial step in performing a GWAS study is to secure high-quality data. Various tests can be applied to curate the data, both at SNPs and individual level, to remove those results that do not achieve the desired level of quality. PLINK software is a useful tool for performing this selection. To achieve robust and significant results after a GWA study, a large number of samples should be included in the analysis. However, achieving a large sample size may require the analyst to merge several smaller cohorts in a meta-analysis, with the possibility that different cohorts are genotyped on different platforms and arrays. At this point, imputation is required to avoid reducing the number of SNPs that are available for analysis. Reference panels from the HapMap and 1000 Genomes projects provide a great number of *in silico* SNPs to fill in the gaps in the original genotyped results. Displaying the results in a Q–Q plot or Manhattan plot will help determine whether there is any marker that is significantly associated with the trait under study. However, the key point for a GWAS hit is being able to replicate it in an independent cohort.

NGS is a very recent approach for analysing a large quantity of data, obtained from whole genome, whole exome or targeted sequencing. Although the possibility of analysing the whole genetic content would be very useful for detecting markers that are not captured by other studies, such as GWAS, this approach is only used to detect single-nucleotide variations, not large deletions or insertions. Analysing the results from an NGS test and understanding their biological meaning also represents a significant challenge. As for GWAS results, replication of significantly associated variants in an independent cohort is crucial.

Both methods are valuable for identifying small areas of the genome that may be implicated in the development of diseases or other phenotypes of interest. Research will continue with the identification of the causal variant in the gene, which will subsequently lead to functional studies to explore the role of the gene in the phenotype.

Recommended Further Reading

We recommend reading the user manuals and tutorials provided in the website of each of the above-mentioned software: PLINK, R, IMPUTE, SHAPEIT2, MACH1.0, Minimac, BEAGLE, METAL, Haploview and GRAIL. They will provide a better understanding of the mechanism of action for each tool, as well as to perform any variation in the protocol proposed to fit specific data analysis.

Multimedia

Illumina Company: http://www.illumina.com/
Affymetrix Company: http://www.affymetrix.com/estore/

Acknowledgements

We acknowledge Professor Andre Uitterlinden and Dr Fernando Rivadeneira from Erasmus MC, Rotterdam (The Netherlands) whose research group has provided valuable contributions to the GWAS protocol. We specifically thank Carolina Medina-Gomez, Lizbeth Herrera, Marjolein Peters and Dr Karol Estrada who have developed and implemented improved methods for QC, imputation, meta-analysis and graphical representation of GWAS, on which these guidelines are based.We also thank BC Platforms Ltd (Finland) and its BCSNPmax module used to illustrate the imputation procedures. We thank Simon Roberts for his assistance in improving the English in the manuscript, Professor Stuart Ralston for his comments and Professor Tim Spector for his support. Supported by the IBMS-ECTS Young Investigators.

Conflict of Interest

Dr Gavin Lucas is a partner in Clear Genetics. Dr Nerea Alonso and Dr Pirro Hysi declare no potential conflict of interest.

References

1. Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. *Pharmacogenomics* 2009;**10**:191–201.
2. Olivier M. A haplotype map of the human genome. *Physiol Genomics* 2003;**13**:3–9.
3. International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;**437**:1299–1320.
4. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;**449**:851–861.
5. Barsh GS, Copenhaver GP, Gibson G, Williams SM. Guidelines for genome-wide association studies. *PLoS Genet* 2012;**8**:e1002812.
6. Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. *Stat Sci* 2009;**24**:561–573.
7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–575.
8. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genomics Inform* 2012;**10**:117–122.
9. DiStefano JK, Taverna DM. Technological issues and experimental design of gene association studies. *Methods Mol Biol* 2011;**700**:3–16.
10. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT *et al.* Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* 2011;Chapter 1:Unit1.19.
11. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc* 2010;**5**:1564–1573.
12. Wittke-Thompson JK, Pluzhnikov A, Cox NJ. Rational inferences about departures from Hardy–Weinberg equilibrium. *Am J Hum Genet* 2005;**76**:967–986.
13. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* 2013;**14**:379–389.
14. Sale MM, Mychaleckyj JC, Chen WM. Planning and executing a genome wide association study (GWAS). *Methods Mol Biol* 2009;**590**:403–418.
15. Gauderman WJ. Sample size requirements for association studies of gene–gene interaction. *Am J Epidemiol* 2002;**155**:478–484.
16. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012;**44**:955–959.
17. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet* 2009;**10**:387–406.
18. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;**39**:906–913.
19. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;**5**:e1000529.
20. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;**11**:499–511.
21. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3. (Bethesda.)* 2011;**1**:457–470.
22. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* 2011;**9**:179–181.
23. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 2013;**10**:5–6.
24. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010;**34**:816–834.
25. Browning SR. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* 2006;**78**:903–913.
26. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007;**81**:1084–1097.
27. Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet Epidemiol* 2007;**31**:365–375.
28. Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR *et al.* Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am J Hum Genet* 2008;**83**:112–119.
29. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 2010;**42**:348–354.
30. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–560.
31. Kavvoura FK, Ioannidis JP. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum Genet* 2008;**123**:1–14.
32. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;**26**:2190–2191.
33. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;**21**:263–265.
34. Albagha OM, Visconti MR, Alonso N, Langston AL, Cundy T, Dargie R *et al.* Genome-wide association study identifies variants at CSF1, OPTN and TNFRSF11A as genetic risk factors for Paget's disease of bone. *Nat Genet* 2010;**42**:520–524.
35. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;**316**:1331–1336.
36. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010;**26**:2336–2337.
37. Albagha OM, Wani SE, Visconti MR, Alonso N, Goodman K, Brandi ML *et al.* Genome-wide association identifies three new susceptibility loci for Paget's disease of bone. *Nat Genet* 2011;**43**:685–689.
38. Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* 2009;**5**:e1000534.
39. Estrada K, Styrkarsdottir U, Evangelou E, Hsu YH, Duncan EL, Ntzani EE *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat Genet* 2012;**44**:491–501.
40. Horne BD, Anderson JL, Carlquist JF, Muhlestein JB, Renlund DG, Bair TL *et al.* Generating genetic risk scores from intermediate phenotypes for use in association studies of clinically significant endpoints. *Ann Hum Genet* 2005;**69**:176–186.
41. Belsky DW, Moffitt TE, Sugden K, Williams B, Houts R, McCarthy J *et al.* Development and evaluation of a genetic risk score for obesity. *Biodemography Soc Biol* 2013;**59**:85–100.
42. Panagiotou OA, Ioannidis JP. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol* 2012;**41**:273–286.