A guide for conducting GWAS QC & statistical steps

EHA

16 April 2020

This tutorial provides a guide to conduct GWAS, starting with quality control (QC) steps[1].

1 Software

Plink[2] is the most common tool for conducting QC and statistical for GWAS analysis. Documentation and list of option can be found here http://zzz.bwh.harvard.edu/plink/reference.shtml#options; https://www.cog-genomics.org/plink/.

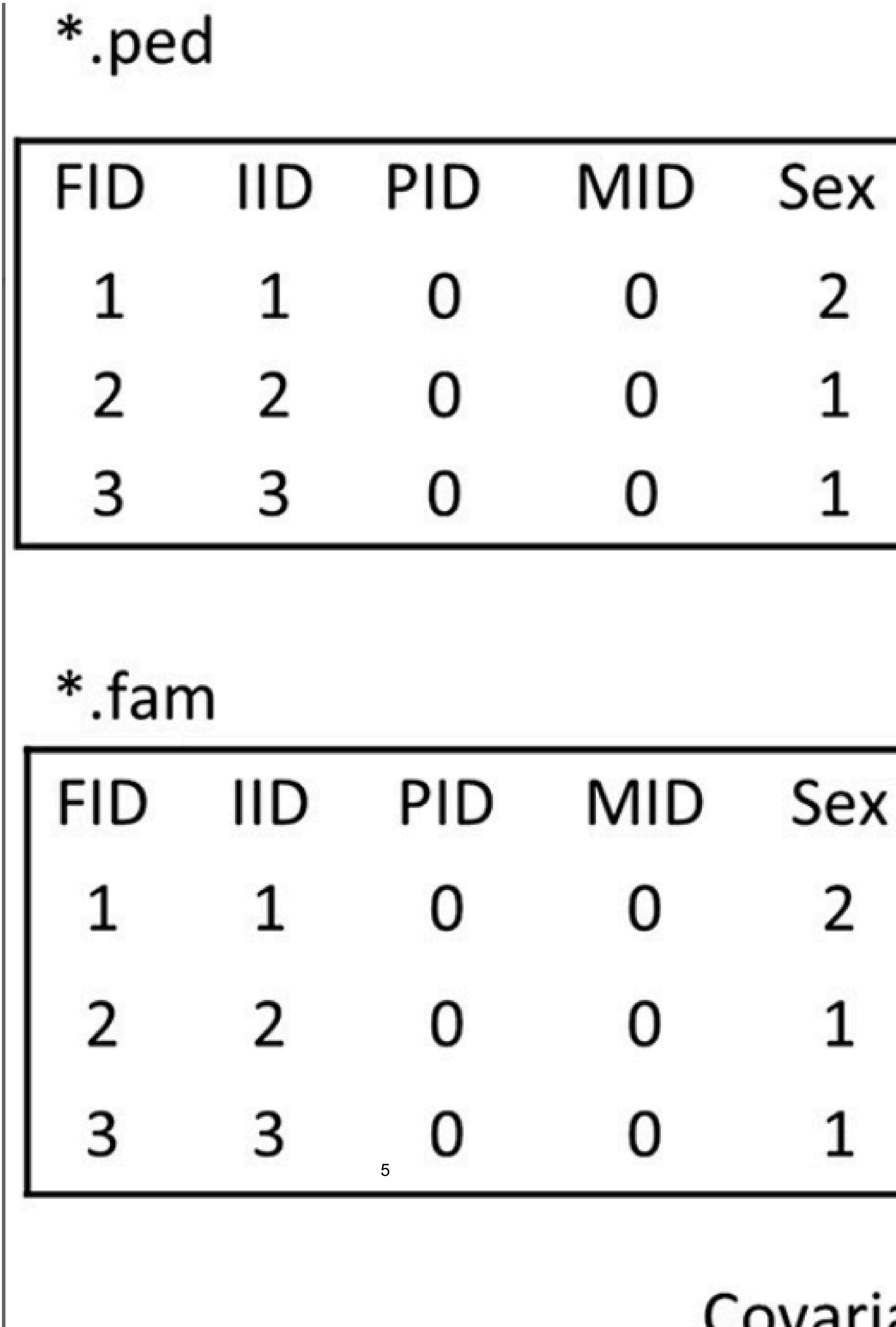
A list of free GWAS tools can be found here.

2 Data format

PLINK can either read text-format files or binary files.

- Text PLINK data consist of two files: one contains information on the individuals and their genotypes (.ped); the other contains information on the genetic markers (.map)
- Binary PLINK data consist of three files, a binary file that contains individual identifiers (IDs) and genotypes (.bed), and two text files that contain information on the individuals (.fam) and on the genetic markers (.bim)

Analysis using covariates often requires a fourth file, containing the values of these covariates for each individual.



*.ped

FID	IID	PID	MID	Sex
1	1	0	0	2
2	2	0	0	1
3	3	0	0	1

*.fam

FID	IID	PID	MID	Sex
1	1	0	0	2
2	2	0	0	1
3	3	0	0	1

Covari

3 QC steps of genetic data

The seven QC steps consist of filtering out of SNPs and individuals based on the following[3]:

Individual and SNP missingness

Individual□level missingness: This is the number of SNPs that is missing for a specific individual. High levels of missingness can be an indication of poor DNA quality or technical problems.

Step	Command	Function	Thresholds and explanation
1- Missingness of SNPs and individuals	□□geno	Excludes SNPs that are missing in a large proportion of the subjects. In this step, SNPs with low genotype calls are removed.	We recommend to first filter SNPs and individuals based on a relaxed threshold (0.2; >20%), as this will filter out SNPs and individuals with very high levels of missingness. Then a filter with a more stringent threshold can be applied (0.02).
	□□mind	Excludes individuals who have high rates of genotype missingness. In this step, individual with low genotype calls are removed.	Note, SNP filtering should be performed before individual filtering.

Sex discrepancy (inconsistencies in assigned and genetic sex of subjects)

Sex discrepancy: This is the difference between the assigned sex and the sex determined based on the genotype. A discrepancy likely points to sample mix□ups in the lab. Note, this test can only be conducted when SNPs on the sex chromosomes (X and Y) have been assessed.

Step	Command	Function	Thresholds and explanation
2- Sex discrepancy	□□check-sex	Checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates.	Can indicate sample mix□ups. If many subjects have this discrepancy, the data should be checked carefully. Males should have an X chromosome homozygosity estimate >0.8 and females should have a value <0.2.

Minor allele frequency (MAF)

Minor allele frequency (MAF): This is the frequency of the least often occurring allele at a specific location. Most studies are underpowered to detect associations with SNPs with a low MAF and therefore exclude these SNPs.

Step	Command	Function	Thresholds and explanation
3- Minor allele frequency (MAF)	□□maf	Includes only SNPs above the set MAF threshold.	SNPs with a low MAF are rare, therefore power is lacking for detecting SNP□phenotype associations. These SNPs are also more prone to genotyping errors. The MAF threshold should depend on your sample size, larger samples can use lower MAF thresholds. Respectively, for large (N = 100.000) vs. moderate samples (N = 10000), 0.01 and 0.05 are commonly used as MAF threshold.

Deviations from Hardy–Weinberg equilibrium (HWE)

The Hardy–Weinberg (dis)equilibrium (HWE) law: This concerns the relation between the allele and genotype frequencies. It assumes an indefinitely large population, with no selection, mutation, or migration. The law states that the genotype and the allele frequencies are constant over generations. Violation of the HWE law indicates that genotype frequencies are significantly different from expectations (e.g., if the frequency of allele A = 0.20 and the frequency of allele T = 0.80; the expected frequency of genotype AT is 2*0.2*0.8 = 0.32) and the observed frequency should not be significantly different. In GWAS, it is generally assumed that deviations from HWE are the result of genotyping errors. The HWE thresholds in cases are often less stringent than those in controls, as the violation of the HWE law in cases can be indicative of true genetic association with disease risk.

Step	Command	Function	Thresholds and explanation
4- Hardy–Weinberg equilibrium (HWE)	□□hwe	Excludes markers which deviate from Hardy–Weinberg equilibrium.	Common indicator of genotyping error, may also indicate evolutionary selection. For binary traits we suggest to exclude: HWE p value <1e−10 in cases and <1e−6 in controls. Less strict case threshold avoids discarding disease□associated SNPs under selection. For quantitative traits, we recommend HWE p value <1e□6.

Heterozygosity rate

Heterozygosity: This is the carrying of two different alleles of a specific SNP. The heterozygosity rate of an individual is the proportion of heterozygous genotypes. High levels of heterozygosity within an individual might be an indication of low sample quality whereas low levels of heterozygosity may be due to inbreeding.

Step	Command	Function	Thresholds and explanation
5- Heterozygosity		Excludes individuals with high or low heterozygosity rates.	Deviations can indicate sample contamination, inbreeding. We suggest removing individuals who deviate ±3 SD from the samples' heterozygosity rate mean.

Relatedness

Relatedness: This indicates how strongly a pair of individuals is genetically related. A conventional GWAS assumes that all subjects are unrelated (i.e., no pair of individuals is more closely related than second□degree relatives). Without appropriate correction, the inclusion of relatives could lead to biased estimations of standard errors of SNP effect sizes. Note that specific tools for analysing family data have been developed.

Step	Command	Function	Thresholds and explanation
6- Relatedness	□□genome	Calculates identity by descent (IBD) of all sample pairs.	Use independent SNPs (pruning) for this analysis and limit it to autosomal chromosomes only.

Step	Command	Function	Thresholds and explanation
	□□cluster □□mds□plot k	Produces a k□dimensional representation of any substructure in the data, based on IBS.	K is the number of dimensions, which needs to be defined (typically 10). This is an important step of the QC that consists of multiple proceedings but for reasons of completeness we briefly refer to this step in the table. This step will be described in more detail in section “controlling for population stratification.”

Pruning: This is a method to select a subset of markers that are in approximate linkage equilibrium. In PLINK, this method uses the strength of LD between SNPs within a specific window (region) of the chromosome and selects only SNPs that are approximately uncorrelated, based on a user□specified threshold of LD.

Population stratification (Ethnic outliers)

Population stratification: This is the presence of multiple subpopulations (e.g., individuals with different ethnic background) in a study. Because allele frequencies can differ between subpopulations, population stratification can lead to false positive associations and/or mask true associations. An excellent example of this is the chopstick gene, where a SNP, due to population stratification, accounted for nearly half of the variance in the capacity to eat with chopsticks.

Because of population substructure or cryptic relatedness, which could cause spurious associations. If information on a large number of genetic markers is available, adjusting the analysis results by using the method of genomic control (GC) is possible.[4–6].LD Score regression approach, that quantifies the contribution of each by examining the relationship between test statistics and linkage disequilibrium (LD). The LD Score regression intercept can be used to estimate a more powerful and accurate correction factor than genomic control. It was found that polygenicity accounts for the majority of the inflation in test statistics in many GWAS of large sample size[7].

Step	Command	Function	Thresholds and explanation
7-Population stratification	□□genome	Calculates identity by descent (IBD) of all sample pairs.	Use independent SNPs (pruning) for this analysis and limit it to autosomal chromosomes only.

Step	Command	Function	Thresholds and explanation
	□□min	Sets threshold and creates a list of individuals with relatedness above the chosen threshold. Meaning that subjects who are related at, for example, pi□hat >0.2 (i.e., second degree relatives) can be detected.	Cryptic relatedness can interfere with the association analysis. If you have a family□based sample (e.g., parent□offspring), you do not need to remove related pairs but the statistical analysis should take family relatedness into account. However, for a population based sample we suggest to use a pi□hat threshold of 0.2, which in line with the literature (Anderson et al., 2010; Guo et al., 2014).

4 Statistical tests

Binary outcome measure

The association between SNPs and a binary outcome (value 1 = unaffected and value 2 = affected; 0 and −9 represent missing)., that can be tested either by X2 test of association that does not allow the inclusion of covariates or with logistic regression analysis will which allows the inclusion of covariates.

Quantitative outcome measure

The association between SNPs and quantitative outcome (i.e.,values other than 1, 2, 0, or missing)., that can be tested either by an asymptotic version of the usual Student's t test that does not allow the inclusion of covariates or with liner regression analysis will which allows the inclusion of covariates.

Correction for multiple testing

What is the Multiple Testing Problem? If you run a hypothesis test, there's a small chance (usually about 5%) that you'll get a bogus significant result. If you run thousands of tests, then the number of false alarms increases dramatically. For example, let's say you run 10,000 separate hypothesis tests (which is common in fields like genomics). If you use the standard alpha level of 5% (which is the probability of getting a false positive), you're going to get around 500 significant results — most of which will be false alarms. This large number of false alarms produced when you run multiple hypothesis tests is called the multiple testing problem. (Or multiple comparisons problem).

Three widely applied alternatives for determining genome□wide significance are the use of Bonferroni correction, Benjamini–Hochberg false discovery rate (FDR), and permutation testing.

- Single step methods like the Bonferroni correction and sequential methods like Holm's method control the Family-wise Error Rate(FWER). The FWER is just a term for all of those false positives you get with multiple tests. Usually used when it's important not to make any Type I Errors at all.
- The Benjamini-Hochberg procedure and Storey's positive FDR control the False Discovery rate. These procedures limit the number of false discoveries, but you'll still get some, so use these procedures if a small number of Type I errors is acceptable.

An unfortunate “side effect” of controlling for multiple comparisons is that you’ll probably increase the number of false negatives — that is, there really is something significant happening but you fail to detect it.

5 Meta-analysis

Combining results from genome-wide association studies (GWASs) of multiple cohorts to to obtain enough power. The combination of GWAS from different cohorts increases the power of analysis, this is the theoretical base of the genotype imputation. Since imputations strongly rely on linkage disequilibrium (LD), the similarity of LD patterns in the study population and the population that is used as the imputation reference is crucial for precise imputations.

Imputation

Meta-analysis equires all cohorts to have the same single-nucleotide polymorphisms (SNPs) in their GWASs. Imputations[8] allow to harmonize the set of SNPs analysed in studies using different SNP arrays, thus facilitating further meta-analysis. Closely linked SNPs are inherited together, so you can predict the genotype of a SNP from the genotypes of nearby SNPs. The technique allows geneticists to accurately evaluate the evidence for association at genetic markers that are not directly genotyped.

The quality of imputation relies on the quality of the reference panel as well as the quality of the study data. To ensure high data quality, there are a number of steps that were taken before imputation begins[9] and further before starting any metaanalysis of GWAS data[10].

References

1. Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research* **27,** 1–10 (2018).

2. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81,** 559–575 (2007).

3. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nature Protocols* **5,** 1564–1573 (2010).

4. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).

5. Devlin, B., Roeder, K. & Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology* **60,** 155–166 (2001).

6. Tsepilov, Y. A. *et al.* Development and application of genomic control methods for genome-wide association studies using non-additive models. *PLoS ONE* **8,** 6–14 (2013).

7. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47,** 291–295 (2015).

8. Suber, P. Gonçalo Abecasis, Serena Sanna, Cristen Willer, Yun Li. *Genomics* (2006). doi:10.1146/annurev.genom.9.081307.164242.Genotype

9. Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in Genetics* **5,** 1–15 (2014).

10. Winkler, T. W. & Day, F. R. Quality control and conduct of genome-wide association meta- analyses. **9,** 1192–1212 (2014).