Genome Wide Association Analysis (GWAS) pipeline

By Dina Gamaleldin Mansour Aly

Introduction

Genome Wide Association Studies (GWAS) (1), are genotype-phenotype association studies that determine the way and extent a certain phenotype is associated with the individual genetic makeup. GWAS includes both genetic dataset and clinical dataset. Clinical dataset usually has clinical information about the individuals enrolled in the GWAS; physical examination data (body weight, height, blood pressure, body temperature, heart rate, performance status, family history of diseases, medication history, individual clinical history), laboratory diagnostic data (complete blood count, random blood sugar, kidney function tests, hepatic function tests, other important specific disease markers that are analyzed in the blood), investigational data (ultrasonography, x-rays, computed tomography) and surgical data (any data about previous operations). The genetic dataset includes the raw genotyping sequencing data. High-throughput sequencing platforms enabled gene sequencing timely and affordable that accelerated the initiation of many genetic projects in the field of medicine. The harvest of the GWAS pave a way to Precision medicine. Genome wide association data analysis is a rate limiting step in the interpretation of genotyping data in the field of population genetics.

Quality control of genotype data

Raw sequencing data is subjected to quality control and imputation using an array of well-developed and optimized software before the data analysis starts. Quality control of the raw genotype data is a very important step; during this step samples are being filtered based on individual characteristics; genotyping rate, sex check, population stratification, identity by descent and heterozygosity and variant sequencing quality; genotyping rate, Hardy_Weinberg equilibrium, minor allele frequency and minor allele count. QC_Protocols (2) using PLINK

produce efficient and reproducible QC that can be customized to suit different genotype datasets. This crucial step can be slow and time consuming depending on the size of the dataset; number of individuals genotyped and number of variants per individuals. Quality control filtering cleans the data from individuals that can cause errorness in the analysis; duplicated individual samples, first degree relatives, individuals from different ancestry and individuals with low genotyping rate. Rare variants are those single nucleotide polymorphism (SNPs) that deviate significantly from Hardy-Weinberg equilibrium and can cause false positives.

Imputation of genotype data

Genotype imputation is a very professional step where the genotypes of the samples that weren't assayed are statistically inferred. Imputation servers; Haplotype Reference Consortium (HRC) (3) is a large reference panels that has 64970 human haplotypes and 39,235,157 snps from 20 cohorts mainly from European ancestry. HRC server offers an imputation platform for genotype data. Genotype data that passed the quality control, is separated to chromosome files each file has the SNPs of single chromosome to be imputed and they are usually in VCF format. The imputed chromosome files along with a quality report are downloaded from the server.

Clinical data preparation

Clinical data is usually raw data from hospital databases; ANDIS (4). The clinical data are checked for their units, tested for normality, any calculated data is done (Body Mass Index (BMI), HOMA_IR, HOMA_B). Some clinical data files need to be formatted so that all the clinical data for each individual enrolled in the GWAS are in one file, this makes further manipulation of the dataset easier.

Association analysis

The association analysis is usually performed using specific designed software programs. PLINK (5,6) and SNPTEST (7-10) are well-known and accredited software programs that are

used for GWAS. Linear Regression model is the statistical model build in the software. The most important step in using these software is the input file formats; binary files for PLINK while GEN and SAMPLE files for SNPTEST. The output files usually have all the required values to report association of SNPs with clinical variable and its direction; beta, P-values, odd_ratios along with standard error and 95% confidence intervals. The output files must be filtered for certain parameters; minor allele frequency (maf), Hardy_Weinberg deviation and GWAS level of significance (5x10e-8).

GWAS pipeline

The GWAS pipeline is designed to perform association analysis for quality controlled and imputed genotype data and one phenotype variable (previously checked for it units and normality). The GWAS pipeline applies SNPTEST for the association analysis.

GWAS pipeline input files

Input files are chromosome VCF (11) files that passed quality control and imputed, and sample file (12) that has the phenotype data for each individual in the imputed chromosome files. The individuals order in the sample file must be in the same order as the chromosome files. This can be checked by making a file of only a subset of SNPs form on imputed chromosome VCF file using VCFTOOLs (13) and then preparing a .gen file and .sample file by GTOOLs (14). The individual clinical data can be added to the generated sample file using Rscripts or BASH scripts. The phenotype (clinical variable must be checked for its units, calculations and normality, the GWAS pipeline presumes that the sample file has the final phenotype to run the analysis).

GWAS pipeline scripts

The GWAS pipeline is designed to analyze GWAS datasets for a single cohort (GWAS pipeline_1.sh) and two cohorts (GWAS_pipeline_2.sh). Each GWAS pipeline consists of a

bash script and an R script (15). The bash script takes seven command line arguments; STUDY_NAME (the name of the GWAS study), SAMPLE (sample file), EXCLUSION_FILE (list of individual identity numbers (IDs) to exclude from the analysis), COV_NAMES (covariate names space delimited), PHENO_NAME (name of the clinical variable as written in the sample file), PHENO_VARIABLE_TYPE (binary, continuous as in the sample file), METHOD (method of counting for the genotype uncertainity; score, threshold, expected, em, ml). The path for the SNPTEST, R and raw GWAS files are to be added to the bash script before the analysis to run. The bash script runs SNPTEST and then sorts and filters the SNPTEST output. The Rscript takes the filtered SNPTEST output files from the bash script and plots the Manhattan and QQplot for the phenotype variable. The R packages used are "qqman" (16) and "plyr" (17).

GWAS pipeline output files

The GWAS pipeline outputs in the same directory as the analysis directory; a SNPTEST output file and log file for each chromosome, Sorted_SNPTEST_output.csv (all chromosome output files are written and sorted for P-values in ascending order), Top_100_GWAS_snps.txt (the top 100 snps with the lowest P-values, Filtered_out_snps.txt (list of the snps that were filtered based on MAF and HWE), GWAS_plot.txt (this file has the SNP ids, chromosome number, base pair position and P-values for all snps in the GWAS study after filtration, "GWAS_manhattan_plot.pdf", and "GWAS_qqplot.pdf".

GWAS pipeline limitations

The GWAS pipeline analyzes one phenotype in a single run; it doesn't analyze more than one phenotype per run. The path for the analysis tools are user supplied as well as the command line arguments. The GWAS pipeline filters based on the SNPTEST out file columns which may different between versions of SNPTEST.

GWAS pipeline efficiency

The GWAS pipeline performs multiple steps automatically without the user attendance. Once started the GWAS pipeline works out all the association analysis and filters the output files for snps "maf and hwe" and then sorts the filtered files for the P-values to get the top 100 snps with the lowest p-values. The GWAS pipeline plots the results; Manhattan plots and QQplots. The output files can be used for annotation of the snps using specific software; ENSEMBL, vVariant Effect Predictor (VEP) (18). The GWAS pipeline saves a lot of time compared with each individual step performed alone. The time taken by GWAS pipeline to complete the analysis of one phenotype depends on the number of individuals and the SNPs in the data files; usually about 12 hours.

References

1- Price AL, Spencer CCA, Donnelly P. Progress and promise in understanding the genetic basis of common diseases, 2015; Proc. R. Soc. B 282: 20151684. http://dx.doi.org/10.1098/rspb.2015.1684.

2- Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols; 5(9):1564-1573. doi:10.1038/nprot.2010.116.

3- McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation, 2016; Nature Genetics 48: 1279–1283, doi:10.1038/ng.3643.

4- ANDiS : All New Diabetics in Scania, Skåne Hospital.

5- PLINK : Package : PLINK v1.90b4.1

Authors : Shaun Purcell, Christopher Chang

URL : www.cog-genomics.org/plink/1.9/

6- Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.

7- SNPTEST : SNPTEST is v2.5.2

8- J. Marchini, B. Howie, S. Myers, G. McVean and P. Donnelly (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics 39 : 906-913.

9- The Wellcome Trust Case Control Consortium (2007) Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447;661-78. PMID: 17554300 DOI: 10.1038/nature05911.

10- J. Marchini and B. Howie (2010) Genotype imputation for genome-wide association studies. Nature Reviews Genetics.

11- The Variant Call Format (VCF) Version 4.2 Specification: The master version of this document can be found at https://github.com/samtools/hts-specs. This printing is version 084587e from that repository, last modified on the date shown above.

12- Sample file: http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html.

13- VCFTOOLs : The Variant Call Format and VCFtools, Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group, Bioinformatics, 2011.

14- GTOOLs : http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html

15- R studio : RStudio is an integrated development environment (IDE) for R. https://www.rstudio.com/products/RStudio/

16- Turner, S.D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. biorXiv DOI: 10.1101/005165.

17- Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

18- McLaren W, Gil L, Hunt S, Riat H, Ritchie G, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. 2016, Genome Biology; 17:122, doi: 10.1186/s13059-016-0974-4.