Angela Andaleon
aandaleon@luc.edu

Ad_PX_Pipe Manual

4 December 2018

INTRODUCTION

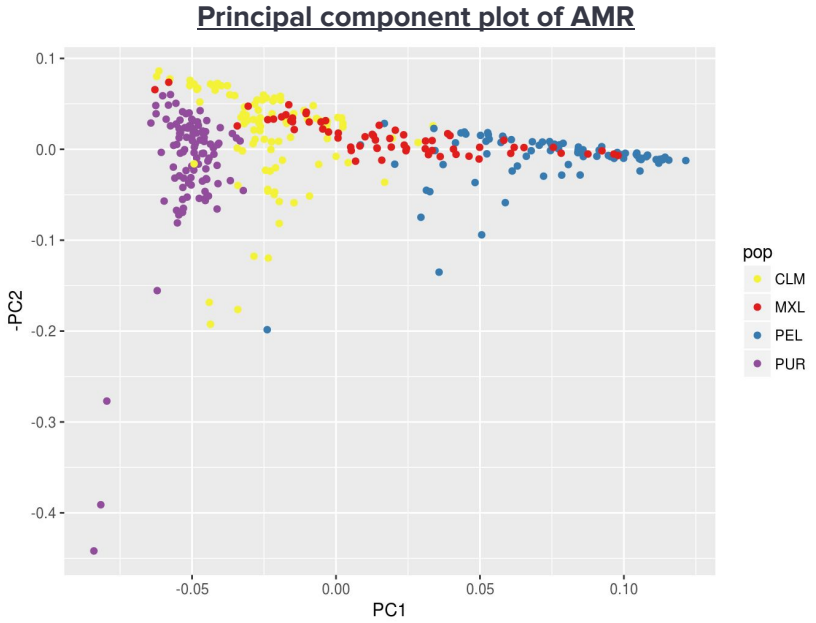
Overview

Ad_PX_pipe is a pipeline for performing a genome-wide association study and PrediXcan in admixed populations, finding independent and eQTL colocalized signals, and mapping admixture using local ancestry estimations. This manual is a supplement to Ad_PX_Pipe that better explains what each step and each script does, as well as giving more detail on the significance behind each step. It is expected that all scripts are run from the same directory.

It is assumed that the user knows the general purpose and execution of a genome-wide association study and PrediXcan, and I will give a brief introduction into colocalization, backward elimination modeling, and local ancestry mapping. However, these topics are much further better explained here, here, and here in their introductions.

Sample Cohort

The test data we're going to use is 1000G AMR, a cohort of 347 individuals from the Americas. Individuals include Mexican Ancestry from Los Angeles, USA (MXL); Puerto Ricans from Puerto Rico (PUR); Colombians from Medellin, Colombia (CLM); and Peruvians from Lima, Peru (PEL). These correspond with my study of the Hispanic Community Health Study/Study of Latinos, as these individuals also have multi-continental ancestry. These data are subset to only include 100,000 SNPs for speed.

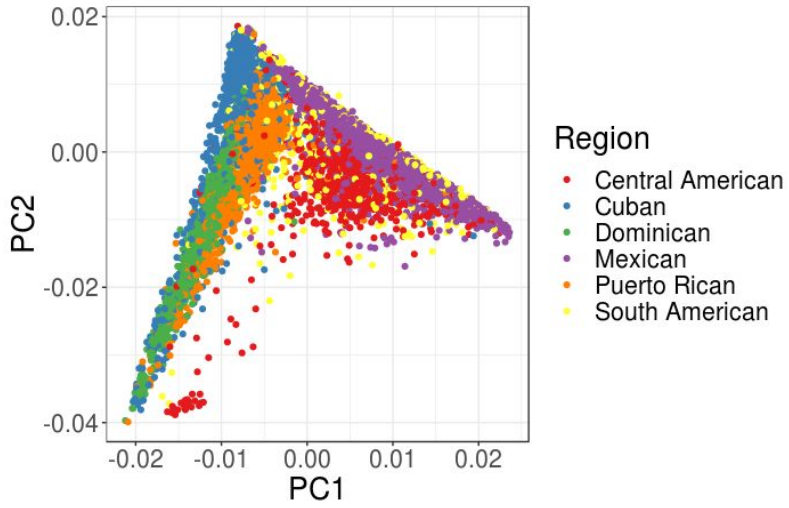


WALKTHROUGH

(0) We will produce randomized phenotypes and covariates, as the 1000G data does not have any phenotypes associated with it, but this does not apply in the real life data you will study. Raw data that you will analyze is normally messy, including low accuracy SNPs or individuals that are outliers in the phenotype (ex. TRIG > 1,500 mg/dL), so we must perform quality control to ensure that our findings are accurate and not simply an artifact of messy data. (1) We would normally perform quality control in PLINK using Ryan's gwasqc_pipeline, but the test data has already been filtered. Please use the gwasqc_pipeline for your own raw data.

A primary concern with admixed populations is population structure due to the differences in allele frequencies continentally, which becomes complicated in individuals with multi-continental ancestry such as AMR. We calculate principal components to control for this population structure, because without these data, the phenotypic findings at the end are heavily confounded, inflated, and unreliable. Additionally, another source of confounding is relatedness within cohorts, which we may not be able to remove completely without sacrificing a large portion of our cohort. GWAS softwares such as GEMMA can account for this relatedness computationally with the input of the cohort's relationship matrix, a measurement of the relatedness between all members of a cohort. (2) We calculate principal components and a relationship matrix in KING, a software optimized for structured populations.

Principal component plot of an admixed Hispanic cohort

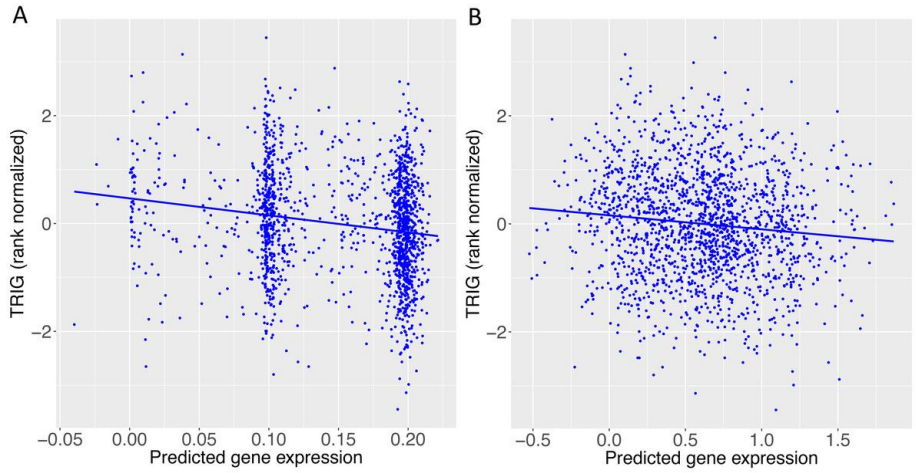


Another issue with raw data is the lack of coverage across the genome, as many commercial genotyping arrays capture less than 500k or 1m SNPs across the genome. We can infer the missing data in between our known data by performing imputation, which uses whole-genome sequences to "fill in" the gaps where SNPs are known to be inherited together most of the time. (3) In regular data, we would then impute data with the Michigan Imputation Server to increase the number of SNPs in linkage disequilibrium with our data. The instructions for this are included in the powerpoint in the repository. This imputation produces another type of genotype format called a VCF (variant call format), so we have to convert it to a useable format for our other softwares. (4) We convert our genotypes to PrediXcan-style dosages with pre-built scripts.

(5) This begins our genome-wide association study using GEMMA. GEMMA, standing for Genome-wide Efficient Mixed Model Association, performs a genome-wide association study while also accounting for relatedness and given covariates, making it ideal for related and structure cohorts. (5a) We convert from the PrediXcan dosage format to the similar BIMBAM format, which is the genotype input for GEMMA. (5b) The next script makes a covariance file from known covariates and a user-determined number of principal components from KING. It is important that you include covariates that may confound your phenotype, such as lipid-lowering medications when studying cholesterol levels. (5c) We then run all these data across all 22 chromosomes in a loop.

Though we have known SNP-level data, we would like to know how they affect biological mechanisms, such as gene expression, which is a much more feasible target for precision medicine. We do not have the cohort's gene expression profile available (because that's expensive and difficult), but we can predict the "missing" data similarly to how we imputed our genotype data - using reference models in PrediXcan. (6) We start the imputed transcriptome based association study by calculating predicted gene expression in 44 GTEx tissues and 5 MESA models. (7a) We then convert these predicted expression data into pseudo-genotypes to input into GEMMA similar to BIMBAM, and (7b) run all populations and tissues in a loop in GEMMA similar to the GWAS, accounting for relatedness and covariates.

Predicted gene expression vs. phenotype



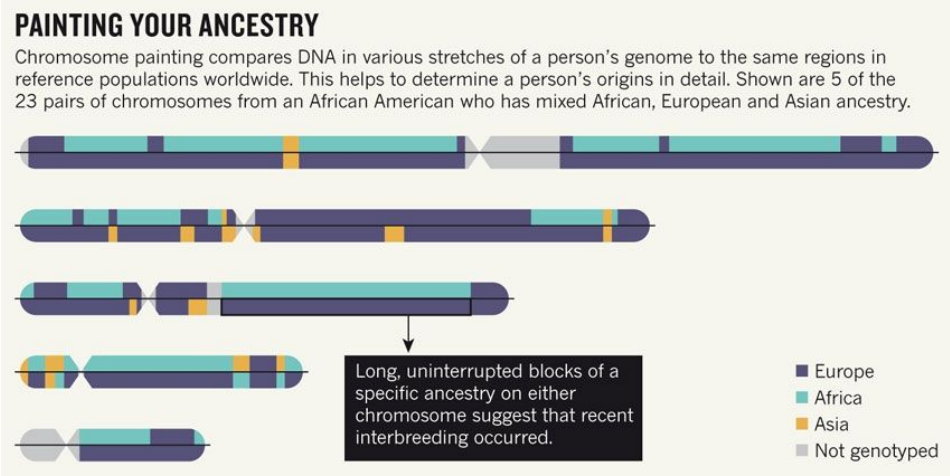
(8) We then extract the significant SNPs with the threshold determined by the user. After determining significant SNPs, we seek to find independently associated loci instead of cutting off independence at an arbitrary distance. (9) We calculate independent significant SNPs in a joint analysis in GCTA-COJO, (9a) starting by converting the GWAS output into GCTA-COJO format, (9b) then running GCTA.

Additionally, with our GWAS results, we seek to find if our SNPs also have biological significance. One form of SNPs with biological significance is an expression quantitative trait loci, which links variations in gene expression levels to genotypes. (10) After we have our significant gene results, we will then perform colocalization testing between GWAS results and eQTL data using COLOC in a COLOC wrapper. COLOC estimates if GWAS and eQTL signals are linked and colocalized (P4 > 0.5), if they are independent from each other (P3 > 0.5), or neither.

Genes are not independent of each other and may have correlated expressions, but we seek to find independent, possibly causal gene signals. (11) We perform backward elimination modelling of all significant genes using stepwise regression to find statistically significant predictors for our phenotype.

A unique aspect of admixed populations is their multi-continental ancestry, which creates a mosaic of ancestries along the genome. For African-American cohorts, this is usually a two-way admixture between European and West African populations, and for Hispanic populations such as AMR, this is usually a three-way admixture between Native American, European, and West African populations. (12) We infer this “mosaic” using local ancestry analyses in RFMix, starting with making reference populations in PLINK, which will be determined mainly by the population in question. For African-American populations, this will be CEU and YRI. For Hispanic populations, this will be IBS, NAT, and YRI. (13) We infer haplotypes using HAPI-UR from our genotypes. (14) Using these haplotypes, we calculate local ancestry estimations in RFMix.

Chromsome painting of multi-continental ancestry



These different blocks of local ancestry may contribute differently to the phenotype, with previous studies finding Hispanic-specific or African-specific SNPs contributing the most to the phenotype. Using our local ancestry estimates, we can also test if the presence of local ancestry is also significantly associated with our phenotype. (15) We convert these local ancestry estimations for use in GEMMA to perform admixture mapping.

SCRIPTS

00_simulate_pheno_covar.R

Purpose: make randomized phenotypes and covariates
Simple input: PLINK binary genotype file
Output: Phenotype file and covariate file

Example:

```
Rscript 00_simulate_pheno_covar.R --bfile AMR
```

Options

- (--bfile) PLINK binary genotype file

02_related_matrix_PCs.R

Purpose: make relatedness matrix and calculate PCs
Simple input: PLINK binary genotype file
Output: GEMMA relationship matrix and principal components file

Example:

```
plink --bfile AMR --chr 22 --make-bed --out AMR_chr22; Rscript 02_relate_matrix_PCs.R
--bfile AMR_chr22
```

Options

- (--bfile) PLINK binary genotype file
- (--king; default = /home/angela/px_his_chol/KING_LAPACK/king-offline) KING executable

05a_PrediXcan_dosages_to_GEMMA.py

Purpose: convert PrediXcan dosages to GEMMA BIMBAM genotypes
Simple input: Path to PrediXcan dosages
Output: 2 folders with information for 22 chromosomes: BIMBAM/ (GEMMA genotype file) and anno/ (SNP annotations)

Example:

```
python 05a_PrediXcan_dosages_to_GEMMA.py --dosage_path dosages/ --dosage_suffix
.txt.gz
```

Options

- (--dosage_path) Path to folder containing dosages and samples.txt
- (--chr; optional; default = 1:22) Path to chromosome to analyze. If no input, analyzes all 22 pairs of chromosomes

- (--dosage_suffix; default = .maf0.01.r20.8.dosage.txt.gz) Suffix of dosages

05b_make_GEMMA_covars.R

Purpose: make covariance file from known covariate and KING PCs
Simple input: Covariate file
Output: GEMMA covariate file

Example:

```
Rscript 05b_make_GEMMA_covars.R --covar covar_woIID.txt --pcs_file kingpc.ped
--pcs_num 5 --output GEMMA_covars.txt
```

Options

- (--covar) Covariance file (w/o IDs)
- (--pcs_file; default = kingpc.ped) Principal components file created by KING
- (--pcs_num; default = 5) Number of principal components to include in covariates file
- (--output; default = GEMMA_covars.txt) Name of output file to be used in GEMMA

05c_GEMMA_loop.sh

Purpose: run GEMMA in either a loop of chromosomes (SNPs) or tissues (genes)
Simple input: GEMMA genotype file, GEMMA phenotype file, GEMMA relatedness matrix, output prefix
Output: GEMMA LMM output

Examples:

- `bash 05c_GEMMA_loop.sh -g BIMBAM/chr -p pheno_woIID.txt -a anno/anno -k relatedness_woIID.txt -c GEMMA_covars.txt -o AMR_`
- `bash 05c_GEMMA_loop.sh -g pred_exp_GEMMA/ -p pheno_woIID.txt -a anno/anno -k relatedness_woIID.txt -c GEMMA_covars.txt -o AMR_ -h`

Options

- All options are detailed here
- (-h) Activate “hacked” GEMMA for use with PrediXcan data

06_make_pred_exp.py

Purpose: calculate predicted gene expressions in PrediXcan using GTEx and MESA models
Simple input: Dosages path
Output: 49 predicted expression files (44 GTEx, 5 MESA)

Examples:

```
python 06_make_pred_exp.py --dosages_path dosages/ --output_prefix pred_exp/
```

Options

- (--PrediXcan_path; default = /usr/local/bin/PrediXcan.py) Path to PrediXcan executable
- (--dosages_path; default = dosages/) Path to PrediXcan dosages
- (--MESA_prefix; default = /home/lauren/files_for_revisions_plosgen/en_v7/dbs/) Prefix of all MESA models
- (--MESA_suffix; default = _imputed_10_peer_3_pcs_2.db) Suffix of all MESA models
- (--GTEx_prefix; default = /home/wheelerlab3/Data/PrediXcan_db/GTEx-V6p-HapMap-2016-09-08/TW_) Prefix of all GTEx models
- (--GTEx_suffix; default = _0.5.db) Suffix of all GTEx models
- (--output_prefix; default = pred_exp/) Prefix of PrediXcan output, preferably a folder name

07a_convert_PrediXcan_to_GEMMA.py

Purpose: convert predicted expression to GEMMA-style pseudo-genotypes
Simple input: Predicted expression prefix
Output: 49 pseudo-genotypes (44 GTEx, 5 MESA)

Examples:

```
python 07a_convert_PrediXcan_to_GEMMA.py --pred_exp_prefix pred_exp/ --output_prefix pred_exp_GEMMA/
```

Options

- (--pred_exp_prefix; default = pred_exp/) Prefix of PrediXcan predicted expression
- (--output_prefix; default = pred_exp_GEMMA/) Prefix of pseudo-genotypes

08_sig_SNP_sig_gene.py

Purpose: find significant SNPs from PrediXcan in R
Simple input: Input prefix
Output: File of significant SNPs (sig_snps.txt) and file of significant genes (sig_genes.txt)

Examples:

```
python 08_sig_SNP_sig_gene.py --SNP_sig 5e-4 --gene_sig 0.05 --input_prefix AMR_
```

Options

- (--SNP_sig; default = 5e-8) Significance threshold for SNPs
- (--gene_sig; default = 9.654e-6) Significance threshold for genes
- (--input_prefix) Prefix for input, not including output/

09a_GEMMA_to_GCTA-COJO.py

Purpose: make GWAS output into GCTA-COJO format
Simple input: PLINK binary genotype .fam file, GWAS prefix, and output prefix
Output: GCTA-COJO input .ma (stands for meta-analysis)

Examples:

```
python 09a_GEMMA_to_GCTA-COJO.py --fam AMR.fam --GWAS_prefix AMR_ --output_prefix AMR
```

Options

- (--fam) .fam file path
- (--GWAS_prefix) Prefix of GWAS results files (not including output/)
- (--output_prefix) Prefix of output file for GCTA-COJO input

DATA FORMATS

PLINK binary genotype file: .bim

Rows: SNPs
Columns (no header): Chromosome number, rs id, distance in centimorgans, base pair positions, effect allele, reference allele
Delimiter: Tab-delimited

Excerpt:

```
1       rs141149254     0       54490   A       G
1       rs62637815      0       59040   C       T
1       rs3131979       0       726944  C       G
1       rs61770163      0       732032  C       A
1       rs144022023     0       732801  G       A
```

PLINK binary genotype file: .bim

Rows: Individuals
Columns (no header): Family ID, individual ID, paternal ID, maternal ID, sex code, phenotype value
Delimiter: Tab-delimited

Excerpt:

```
PR01 HG00551 0 0 0 1
PR02 HG00553 0 0 0 1
PR02 HG00554 0 0 0 1
PR03 HG00637 0 0 0 1
PR03 HG00638 0 0 0 1
```

Phenotype file

Rows: Individuals
Columns (no header): Phenotype
Delimiter: Tab-delimited

Excerpt:

```
-0.445778264836677
-1.2058565689643
0.04112631384569
0.639388407571143
-0.786554355912735
```

Covariate file

Rows: Individuals
Columns (no header): Intercept, covariate
Delimiter: Tab-delimited

Excerpt:

```
1	1
1	0
1	0
1	0
1	0
```

GEMMA relationship matrix

Rows: Individuals
Columns (no header): Individuals
Delimiter: Tab-delimited

Excerpt:

```
0.5	-0.0344	0.0262	-0.0742	-5e-04
-0.0344	0.5	0.0307	-0.0227	0.0041
0.0262	0.0307	0.5	0.0119	0.0232
-0.0742	-0.0227	0.0119	0.5	-0.0041
-5e-04	0.0041	0.0232	-0.0041	0.5
```

Principal components file

Rows: Individuals
Columns (no header): Family ID, individual ID, paternal ID, maternal ID, sex code, phenotype value, PC1, PC2...
Delimiter: Tab-delimited

Excerpt:

```
CLM01 HG01119 0 0 0 1 0.0149 0.0368 -0.0242
CLM02 HG01121 0 0 0 1 0.0107 0.0332 0.0416
CLM02 HG01122 0 0 0 1 -0.0042 0.0068 0.0465
CLM03 HG01112 0 0 0 1 -0.0675 0.0730 0.0109
CLM03 HG01113 0 0 0 1 -0.0091 0.0273 -0.0379
```

PrediXcan dosage

Rows: SNPs
Columns (no header): Chromosome number, rs id, base pair positions, effect allele, reference allele, minor allele frequency, individual 1 dosage, individual 2 dosage...
Delimiter: Tab-delimited

Excerpt:

```
22 rs3001810 16058766 G A 0.3285 1 0
22 rs1807458 16071624 G A 0.2767 1 0
22 rs2334338 16143946 G A 0.0634 0 1
22 rs2019546 16155259 G A 0.1499 0 1
22 rs372779614 16212480 T C 0.04899 0 0
```

GEMMA BIMBAM genotype

Rows: SNPs
Columns (no header): rs id, effect allele, reference allele, individual 1 dosage, individual 2 dosage...
Delimiter: Tab-delimited

Excerpt:

```
rs3001810       G       A       1       0
rs1807458       G       A       1       0
rs2334338       G       A       0       1
rs2019546       G       A       0       1
rs372779614     T       C       0       0
```

SNP annotation

Rows: SNPs
Columns (no header): rs id, base pair positions, chromosome number
Delimiter: Tab-delimited

Excerpt:

```
rs3001810       16058766        22
rs1807458       16071624        22
rs2334338       16143946        22
rs2019546       16155259        22
rs372779614     16212480        22
```

GEMMA covariate file

Rows: Individuals
Columns (no header): intercept, covariate 1, covariate 2...
Delimiter: Tab-delimited

Excerpt:

```
1       1       0.0149
1       0       0.0107
1       0       -0.0042
1       0       -0.0675
1       0       -0.0091
```

Predicted expression file

Rows: Individuals
Columns (w/ header): FID, IID, gene 1, gene 2...
Delimiter: Tab-delimited

Excerpt:

```
FID     IID     ENSG00000000457.8       ENSG00000000460.12
PR01    HG00551 0.0     0.0
```

```
PR02    HG00553 -0.0606493223073        0.0
PR02    HG00554 0.0     0.0
PR03    HG00637 0.0     0.0
```

Pseudo-genotype

Rows: Individuals
Columns (w/o header): Gene, allele 1 (NA), allele 0 (NA), predicted expression 1, predicted expression 2...
Delimiter: Tab-delimited

Excerpt:

```
ENSG00000000457.8       NA      NA      0.0     -0.060649322307299997   0.0
ENSG00000000460.12      NA      NA      0.0     0.0
ENSG00000000938.8       NA      NA      0.0     0.0
ENSG00000001036.8       NA      NA      0.0     0.0
ENSG00000001084.6       NA      NA      0.0     0.0
ENSG00000001167.10      NA      NA      0.0     0.0
```

Significant SNP file

Rows: SNPs
Columns (w/ header): chromosome number, rs id, base pair position, number of missing individuals, effect allele, other allele, allele frequency, effect size, standard error of effect size, l_remle, l_mle, P (we will use), p_lrt, p_score
Delimiter: Tab-delimited

Excerpt:

```
chr     rs      ps      n_miss  allele1 allele0 af      beta    se      l_remle l_mle   p_wald
p_lrt   p_score
1       rs72895329      53720574        0       G       A       0.124   3.835443e-01
1.079198e-01    1.000000e-05    1.000000e-05    4.332836e-04    3.956725e-04    5.833772e-04
1       rs861335        145195645       0       C       T       0.058   5.803832e-01
1.607578e-01    6.307488e-02    7.157456e-02    3.520143e-04    2.965795e-04    3.874091e-04
1       rs4658546       242070190       0       G       A       0.385   2.817338e-01
7.126564e-02    1.443027e-01    1.547113e-01    9.385176e-05    8.157362e-05    1.300028e-04
3       rs995540        131433077       0       C       T       0.343   2.640302e-01
7.440052e-02    1.000000e-05    1.000000e-05    4.416252e-04    4.034767e-04    5.329731e-04
```

Significant gene file

Rows: Genes
Columns (w/ header): chromosome number (NA), gene id, base pair position (NA), number of missing individuals, effect allele (NA), other allele (NA), allele frequency, effect size, standard error of effect size, l_remle, l_mle, P (we will use), p_lrt, p_score
Delimiter: Tab-delimited

Excerpt:

```
chr     rs      ps      n_miss  allele1 allele0 af      beta    se      l_remle l_mle   p_wald
p_lrt   p_score
-9      ENSG00000002822.11      -9      0       NA      NA      0.022   -3.990223e+00
1.324256e+00    6.902094e-02    7.676000e-02    2.779816e-03    2.456470e-03    2.820869e-03
-9      ENSG00000035862.8       -9      0       NA      NA      -0.005  1.852481e+01
8.176046e+00    1.112456e-01    1.194347e-01    2.409748e-02    2.285608e-02    2.497196e-02
```

```
-9      ENSG00000055147.13      -9      0       NA      NA      0.000   9.172650e+01
4.378480e+01    9.871645e-02    1.076464e-01    3.691813e-02    3.495499e-02    3.702170e-02
-9      ENSG00000067064.6       -9      0       NA      NA      0.005   1.640739e+01
7.927578e+00    1.314638e-01    1.380397e-01    3.924134e-02    3.867524e-02    4.394364e-02
```

GCTA-COJO: .ma

Rows: SNPs
Columns (w/ header): rs id, effect allele, other allele, effect allele frequency, effect size, standard error of effect size, P, sample size
Delimiter: Tab-delimited

Excerpt:

```
rs      allele1 allele0 af      beta    se      p_wald  347
rs141149254     G       A       0.102   1.896115e-01    1.170418e-01    1.061550e-01    347
rs62637815      T       C       0.174   -1.737071e-01   8.690511e-02    4.642662e-02    347
rs3131979       G       C       0.305   -4.147885e-02   1.053819e-01    6.941201e-01    347
rs61770163      A       C       0.125   6.299383e-02    1.130819e-01    5.778516e-01    347
```

GCTA-COJO: .jma.cojo

Rows: SNPs
Columns (w/ header): chromosome number, rs is, base pair position, reference allele, effect allele frequency, effect size, standard error of effect size, P, sample size, frequency of the effect allele in the reference sample, effect size, standard error and p-value from a joint analysis of all the selected SNPs, LD correlation between the SNP i and SNP i + 1 for the SNPs on the list.
Delimiter: Tab-delimited

Excerpt:

```
Chr     SNP     bp      refA    freq    b       se      p       n       freq_geno       bJ
bJ_se   pJ      LD_r
20      rs61437950      52287257        C       0.33    0.314649        0.0730558
2.1703e-05      347     0.670029        0       0       1       0
```

SOFTWARE MANUALS AND CITATIONS

PLINK - http://zzz.bwh.harvard.edu/plink/index.shtml

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

KING - http://people.virginia.edu/~wc9c/KING/manual.html

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26(22):2867-2873

Michigan Imputation Server - https://imputationserver.sph.umich.edu/index.html#!pages/help

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze S, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. Next-generation genotype imputation service and methods. Nature Genetics 48, 1284–1287 (2016).

GEMMA - http://www.xzlab.org/software/GEMMAmanual.pdf

Xiang Zhou and Matthew Stephens (2012). Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 44: 821-824.

PrediXcan - https://github.com/hakyimlab/PrediXcan/tree/master/Software

Gamazon ER[†], Wheeler HE[†], Shah KP[†], Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ, Im HK. (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. doi:10.1038/ng.3367.

GCTA-COJO - https://cnsgenomics.com/software/gcta/#COJO

Yang et al. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44(4):369-375.

Yang J, Lee SH, Goddard ME and Visscher PM. GCTA: a tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011 Jan 88(1): 76-82.

COLOC - https://cran.r-project.org/web/packages/coloc/coloc.pdf

Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. (2014) Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet 10(5): e1004383. https://doi.org/10.1371/journal.pgen.1004383

HAPI-UR - https://code.google.com/archive/p/hapi-ur/

Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of Many Thousands of Genotyped Samples. The American Journal of Human Genetics 91, 238–251 (2012).

RFMix - https://sites.google.com/site/rfmixlocalancestryinference/

Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. The American Journal of Human Genetics 93, 278–288 (2013).