JAMP

Joint Genetic Association of Multiple Phenotypes

Manual, version 1.0
15/06/2015
D Posthuma
AE van Bochoven
ctglab.nl/software
danielle.posthuma@vu.nl

JAMP is a free, open source tool to run multivariate GWAS. It combines information from multiple phenotypes to obtain one combined P-value per SNP. As JAMP uses permutation to determine P-values, raw genotype data is required as input.

If you use JAMP, please refer to

Posthuma D, de Leeuw C, van der Sluis S, Bochoven A. JAMP: Joint Genetic Association of Multivariate Phenotypes. *submitted*

http://ctglab.nl/software

Download

JAMP is written in Python and runs on Mac OSX, Windows and Linux. It can be downloaded from http://ctglab.nl/software. After download, copy all the JAMP files to your working directory. JAMP is a command line tool, you need to open a terminal window and type commands at the prompt to perform analyses with JAMP.

JAMP can be invoked by typing ./jamp, from the directory where JAMP is installed.

The following files are available for download:

jamp	*[the program]*
JAMP_manual.pdf	*[the manual]*
example.bed	*[example binary pedigree file]*
example.bim	*[example binary map file]*
example.fam	*[example fam file]*
pheno_b.txt	*[a phenotype file containing 4 binary traits]*
pheno_q.txt	*[a phenotype file containing 4 quantitative traits]*
pheno_bq.txt	*[a phenotype file containing 8 traits: 4 binary and 4 quantitative]*

The example input files are based on the CEU hapmap genotypes (10 SNPs) with randomly generated phenotypes.

Besides JAMP you need to have PLINK from Shaun Purcell installed, which can be downloaded freely from here: http://pngu.mgh.harvard.edu/~purcell/plink. In addition, python v2.6 or higher is required and if not already installed can be downloaded from: http://python.org/getit/.

Running JAMP from any directory

You can skip this section if you only want to run JAMP from the folder where you saved the source code.

If you want to invoke JAMP from any directory you need to add the path to JAMP to PATH. How to do this, depends on the environment you are working in. Alternatively you can add (a link to) the executable to a folder that is already in your PATH. Typing PATH shows the current PATH settings.

If you are using a bash shell in a MAC/UNIX/Linux environment, you need to modify the file .profile (for MacOSX) or .bashrc (for some other linux versions) (in your home directory). Simply add the following line to this script:

alias jamp= [insert path to the jamp executable]

This allows invoking jamp from any directory by typing jamp at the prompt.

Input Files

JAMP uses the same input files as PLINK, which can be in plain text rectangular (.ped and .map) format, in transposed or long format, or in the more efficient and widely used binary format (.bed, .bim. and .fam files). Please consult the PLINK documentation (http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#plink) if you are unfamiliar with these file formats.

JAMP expects that multivariate phenotypes are available in a separate, alternate phenotype file, formatted as specified by PLINK, with the exception that the phenotype file is not allowed to have a header row for use in JAMP. The first two columns in the phenotype file must contain Family ID and Individual ID, the consecutive columns contain the multiple phenotypes. Also, the delimiter must be a space, not a tab.

IMPORTANT: *The alternate phenotype file should not contain a header and should have space as a delimiter.*

Quickstart: running JAMP

To obtain a combined P-value based on 100 permutations of all phenotypes for each SNP, provide PLINK input files and a phenotype file without a header and with spaces as delimiter, then type:

```
./jamp --bfile example --assoc --pheno pheno_b.txt --all-pheno --jperm 100
```

This will invoke JAMP, which first calls PLINK to run genetic association tests on all SNPs available in the input files and for all phenotypes in pheno_b.txt. It will then permute and calls PLINK 100x to run association on the permuted phenotypes. The --all-pheno option must always be provided to ensure PLINK runs association on all phenotypes, and is the JAMP default. If you only wish to analyze a subset of your phenotypes it is required to generate a new alternate phenotype file that includes only those phenotypes that you wish to analyze.

While permuting JAMP keeps all phenotypic scores from one individual together in order to retain the phenotypic structure in the original data. JAMP thus corrects for the correlational structure between phenotypes and does not make any assumptions on the multivariate nature of the phenotypic data. The phenotypes can be binary, quantitative or a combination of these.

Crude permutation controlling family wise error rate

The command --jperm invokes crude permutation and runs the same number of permutations for each SNP. For example, the command --jperm 1000 runs 1000

permutations for all SNPs and provides an empirical P value (EMP_P) based on 1000 permutations.

The empirical P-value is calculated as follows: JAMP first calculates and stores the **Σ**-$\log_{10}$(P) across all phenotypes for each SNP based on the original dataset. Then for each permutation JAMP also calculates the **Σ**-$\log_{10}$(P). When finished permuting, JAMP obtains the empirical P-value (EMP_P) for each SNP by dividing the number of times the **Σ**-$\log_{10}$(P) from the permuted analyses exceeds or equals the **Σ**-$\log_{10}$(P) from the original analysis (hits, *H*) by the number of permutations run (*M*):

$$EMP_P = \frac{\boldsymbol{H}}{\boldsymbol{M}}$$

As the same number of permutations for every SNP is run, JAMP calculates an additional empirical P-value (EMP_{P_COR}) that controls for the family wise error due to testing multiple SNPs. This is achieved by comparing every observed **Σ**-$\log_{10}$(P) with the maximum **Σ**-$\log_{10}$(P) obtained across all SNPs for each permutation.

The empirical P valued based on the **Σ**-$\log_{10}$(P) test statistic, tests the hypothesis that the multivariate pattern of P-values of all phenotypes is significantly different than what is expected under the null hypothesis of no association. A significant P value is thus suggestive of multivariate association with a SNP.

In addition to this test, JAMP produces a second empirical P-value (EMP_{Pmin}) to test the hypothesis that at least one of the phenotypes is significantly associated with a SNP, given the multivariate nature of the phenotypes. For each SNP, the smallest P value from the original range of P-values from all phenotypes is evaluated against the smallest P-value from all P-values for the multiple phenotypes obtained in each permutation, thus correcting for the multivariate nature of the data. In addition, the original smallest P-value is evaluated against the smallest of the smallest P-values across all SNPs, providing the EMP_{Pmin_CORR}, which is corrected for testing multiple SNPs.

The output of --jperm thus produces two P-values per tested hypothesis: one empirical P-value which is corrected for testing multiple phenotypes, but uncorrected for testing multiple SNPs (and which needs to be evaluated against a generally accepted genome-wide significance level based on Bonferroni correction for multiple testing) and one empirical P-value which is corrected testing both multiple phenotypes and multiple SNPs which can be evaluated against a nominal significance level of 0.01 or 0.05. This latter P-value corrects for multiple testing conditional on the genomic data and tends to be less conservative compared to Bonferroni. As it is usually sufficient to show that the corrected P-value is < .05 or .01, only 100-1000 permutations are needed with the crude permutation scheme.

When finished, a file called jamp.empp is created, containing the following columns:

CHR	The name of the chromosome
SNP	The SNP-id
NPHENO	The number of phenotypes for which the analysis was run
P_P1	The P-value of the 1st phenotype, as produced by PLINK
P_P2	The P-value of the 2nd phenotype, as produced by PLINK
..	..
P_Pn	The P-value of the nth phenotype, as produced by PLINK

SUMLOGP	The **Σ**-$\log_{10}$(P) across all phenotypes for one SNP
NPERMS	The number of permutations run for each SNP
EMP_P	The empirical P-value of the multivariate SNP association
EMP_P_COR	EMP_P corrected for the family-wise error rate of testing multiple SNPs
EMP_Pmin	The empirical P-value of the test that at least one phenotype is significantly associated given the multivariate nature of the phenotypes
EMP_Pmin_COR	EMP_Pmin corrected for the family-wise error rate of testing multiple SNPs

Note that the --jperm [number] option not only specifies how many permutations need to be carried out, but it also specifies the seed numbers, as JAMP takes the number of the permutation as seed number. This can come in handy if you wish to reproduce exactly the same results. However if you wish to split up the permutations in two batches of 500 each, you need to ensure that you do not obtain two batches of exactly the same permutations. The command --jperm 1-500 specifies that 500 permutations will be run with permutation (seed) numbers 1-500 whereas the command --jperm 501-1000 specifies a different set of 500 permutations with seed numbers 501-1000.

IMPORTANT: *The permutation number is used as seed number - handy for exact reproduction, yet take note (i.e. provide different seed numbers) when running parallel jobs that later need to be merged.*

When running parallel permutations that later need to be merged, it is generally practical to use the option --out which adds a prefix to the JAMP output files. For example:

```
./jamp --bfile example --assoc --pheno pheno_b.txt --all-pheno --jperm 1-500 --out run1
```

and

```
./jamp --bfile example --assoc --pheno pheno_b.txt --all-pheno --jperm 501-1000 --out run2
```

will generate run1.jamp.empp and run2.jamp.empp

The empirical P-value based on all 1000 permutations can easily be obtained afterwards with the command jmerge:

```
./jamp --jmerge run1.jamp.empp run2.jamp.empp --out all
```

Or, if you have a long list of runs you can use a wildcard:

```
./jamp --jmerge run*.jamp.empp --out all.empp
```

This will generate an output file called all.empp.jamp.merged that includes an empirical P value based on the total number of permutations.

If JAMP is used to run multiple permutations simultaneously (for example using a cluster), JAMP will start each set of permutations with running PLINK on the actual data. In some cases (i.e. when the original analysis takes a long time) it may be convenient to provide

JAMP with output files from the actual run to avoid running the same analyses multiple times and to save computing time. The command --jstart invokes this behavior. For example

```
./jamp --bfile example --assoc --pheno phenol_b.txt --all-pheno --jperm 1-500 --out run1 --jstart run0
```

Will cause jamp to search for the following files
run0.jamp.chr_snp
run0.npheno_pheno_x
run0.jamp.sumlogp

JAMP will then skip running PLINK on the actual data and will start permutation right away. These files from the original run can be obtained by running jamp with zero permutations:

```
./jamp --bfile example --assoc --pheno pheno_b.txt --all-pheno --jperm 0 --out run0
```

IMPORTANT: *When using --jstart it is important that the provided files are based on exactly the same dataset as specified with the --bfile option*

Supported PLINK options

JAMP currently supports the following options for association in PLINK:

--assoc
--linear
--logistic
--trend
--model
--dosage

JAMP currently does not support the --mh, --adjust or any of the family based association options from PLINK.

In theory all other options that are used in PLINK can be added on the command line when calling JAMP. However, since JAMP uses permutation, it is often a good idea to pre-run options that require some time, especially when running 100 or 1000 permutations. In particular options intended to clean the datafiles are advised to be used with --make-bed in PLINK prior to running JAMP (e.g. --extract --remove --hwe --maf).

Some PLINK options require special attention when running JAMP:

--out [prefix]	The PLINK option --out changes all prefixes in PLINK, and also in JAMP
--sex	If you wish to correct for sex, you have to put the sex codes in a covariate file and use --covar, do not use the --sex option with JAMP
--covar [file.txt]	When you use the --covar option, JAMP will permute all covariates with the phenotypes, i.e. the relation between the covariates and the phenotypes is retained and the analyses on the permuted datasets are carried out using the same corrected phenotypes as in the original analyses. The file

	containing covariates should be in the PLINK covariate file format, except that it should not contain a header (whereas PLINK accepts both with and without a header), and have spaces and not tabs. Note that the JAMP output will also contain p-values for the covariates.
--adjust	JAMP currently does not support taking the GC corrected P-values, if you use --adjust and --assoc, JAMP will work with the output from --assoc. Adding --adjust will only slow down the permutation procedure.