Running Score Tests in MMAP

MMAP implements score tests as an extension of routines that estimate variance components within the mixed model. The variance component routines are used to compute the variance estimates under the null model and compute the residuals, which are then used in the score test. These score tests can be structured to include any number of random effects in the null model such as additive, dominance, epistasis, permanent environment, genomic and heteroscedastic errors in both population and family data. The simplest model for population data would be an error term and for family data a polygenic effect and error term.

MMAP can perform general SNP set testing but we assume rare variant testing for ease of presentation. MMAP uses that standard two step approach. First, the score statistics and covariance matrix are computed and save to a file. These results are then used for rare variant tests and meta analysis. Currently we implement a SKAT and burden test, but are working on additional tests. We also have an option to create a txt output file that can be can converted into an R object that can be used with the seqMeta R package [1] for variant or analysis. We are also working on an interface to rvtests [2] and raremetalworker [3].

Command line options

The pedigree, phenotype and trait options are as described previously.

--ped <pedigree filename>

--phenotype_filename <filename>

--trait <trait>

--covariates <covariates>

Genotype Data

--binary_genotype_filename <filename>

Binary formatted genotype file in marker-by-subject (MxS) format as described previously. Note that when converting a csv file to MMAP format non-autosome chromosome names should be X,Y,XY and MT, not a numerical value. If the binary genotype file was created from Plink files, the conversion is done automatically.

Missing genotype data

Missing genotypes are imputed to the average dosage based on the subjects in the analysis, called the *sample frequencies*, not the total subjects in the genotype file, called the *population frequencies*. To use population frequencies add the option **--use_population_frequencies**. The minor allele frequency may differ significantly between the two if the population comprises different genetic groups and the subjects being analyzed are from only one of them. For example, if the genotyped *sample* subjects are monomorphic for a rare effect allele, then the

SNP remains monomorphic after imputing missing genotypes, but if the allele is present in the genotyped *population* then missing data is imputed to a non-zero dosage.

X chromosome

The allele frequency on the X chromosome is computed using both males and females, but with males contributing a single allele in both numerator and denominator. Male genotypes are coded 0/2 and missing data for both sexes is assigned the mean dosage. The option **--x_male_coding_01** will code males as 0/1 with missing data in males assigned half the mean dosage. The male coding does not impact the value of the mean dosage as that is determined by allele frequencies. Male heterozygous genotypes are treated as missing data, thus, do not contribute to the allele frequency estimation and are assigned the appropriate missing value.

--binary_covariate_filename <filename>

Option allows the binary genotype file to be used as a covariate to facilitate including SNPs as covariates such as conditional and interaction analysis. **Note that when genotypes are entered as covariates in the model missing genotypes are NOT imputed to sample averages as is done for SNP analysis, thus subjects missing genotypes are dropped from the analysis.** [option to impute is being added]

Score Tests

To perform score tests additional options are added to the variance component estimation routines [see variance component documentation for details]. Currently the format follows the seqMeta package used by CHARGE that requires a SNP info file that groups the variants into genes or other clusters.

--score_test

Option to run score test and requires the option –snp_info_filename.

--snp_info_filename <filename>

A **tab** delimited file with marker and gene with header, say **Locus** and **SNPset.** The file is tab delimited to account for formats that where a *gene* comprises multiple gene names separated by a comma. The file header is actually not used at the moment, so can any two names can be used. Chromosome information for marker is assumed to be present in the binary genotype file. A snp info file compatible with the CHARGE exome chip analysis plan is available for download on the website. We may expand the allowable format to include chromosome and position information.

--gene_set <filename>

Specify a set of genes to include in the analysis for a subset analysis. The file has one gene name per row with no header.

--output_prepscores_file

Output a text file the can be converted into a seqMeta preScores R object with **mmap2seqmeta.R**

Output files

<trait>.<suffix>.skat.res.bin

Binary file containing scores and covariance matrix used for rare variant testing and meta analysis.

<trait>.<suffix>.prepscores.txt

Produced with the ---output_prepscores_file. A tab delimited file that contains the sample size, standard error, gene name, marker names, score test and LD matrix that can be converted using the script mmap2seqmeta.R into an R object that can be input into seqMeta for running rare variant tests.

Converting MMAP output to seqMeta R object

mmap2seqmeta.R <trait>.<suffix>.prepscores.txt <output filename> <R name>

R script to convert MMAP output file to an R object for input into seqMeta. < R name> is the name assigned when running prepScores that will be used in the skatMeta call. For example, <R Name><-prepScores(…) and

skatMeta(<R name>,SNPInfo=SNPInfo,snpName="Name",aggregateBy="gene")

Example MMAP commands

Family data: Command run score test with null model BMI = mean + sex + age + sex*age + exm10 + exm20 + g + e, where sex, age, and sex-by-age interaction are fixed effects, SNPs exm10 and exm20 are also fixed effects and read from the binary genotype file exome.MxS.bin, g is the additive random effect with label A and modeled by covariance matrix kinship.bin. Estimation of the variances use EM-REML for 1 iteration, then AI-REML using DPOTRS and two threads. The reported likelihood includes the constant term. A text file that contains the seqMeta prepScores output is produced.

```
mmap  --score_test --snpinfo_filename snpinfo.txt --binary_genotype_filename
exome.MxS.bin --ped pedigree.csv --phenotype_filename phenotype.csv --trait
BMI  --estimate_variance_components --variance_component_filename kinship.bin
--num_em_reml_burnin 1 --use_em_ai_reml  --use_dpotrs  --
variance_component_label A --covariates sex age exm10 exm20 --
add_likelihood_constant  --interaction age*sex --binary_covariate_filename
exome.MxS.bin --file_suffix exome.conditional --num_mkl_threads 2 –
output_prepscores_file
```

Population data: Same model as above but excluding the polygenic component. The **–single_pedigree** is required when treating subjects as single cluster..

```
mmap  --read_snpinfo_file snpinfo.txt --binary_genotype_filename
exome.MxS.bin --ped pedigree.csv --phenotype_filename phenotype.csv --trait
BMI --covariates sex age exm10 exm20 --add_likelihood_constant –interaction
age*sex  --score_test_weight  wu  --binary_covariate_filename exome.MxS.bin -
-file_suffix exome.conditional –single_pedigree –output_prepscores_file
```

Rare Variant Tests

The input file(s) are of the form **<trait>.<suffix>.skat.res.bin**. When more than one file is present the results are meta analyzed. The csv output file contains the results of the analysis.

--burden_meta <input file(s)> –max_minor_allele_frequency <maf> --csv_output_filename <output file>

Perform burden test using MAF cutoff.

--skat_meta <input file(s)> --score_test_weight wu csv_output_filename <output file>

SKAT tests using Wu weights. Other weight functions to be added.

Output <trait>.<suffix>.skat.res.txt

The file contains the results of the score test using the Wu weights. P-values are computed using Kuonen's algorithm. The file format is similar to the output format of skatMeta in seqMeta.

Computational considerations

There are two main computations for the score test: (1) estimating the variance components and (2) computing the score statistics. For family studies that consist of many pedigrees with a small number of subjects, (1) is most efficient treating each pedigree as its own cluster, but (2) is most efficient treating the pedigrees as a single pedigree with the option **–single_pedigree** added to the command line. The reason is that the advantage of MKL for matrix multiplication increases as the matrices become larger. Thus, if running MMAP across many phenotypes, it might be useful to determine if **--single_pedigree** increases overall performance using a single phenotype. The same recommendation holds for determining if parallelization with threads should be used.

Comparing MMAP and seqMeta

Since prepScores fits the null model using coxme::lmekin with method="REML" the estimates of the model fit will be the same as MMAP and the log likelihoods will match if it the option *add_likelihood_constant* is included in the MMAP run. The skatMeta results should match between the programs (within numerical roundoff)

IMPORTANT: When comparing MMAP and seqMeta for family data the kinship matrix in prepScores must be ordered the same as the phenotype data, otherwise the results will not be correct.

Parallelization

The score test can be parallelized by breaking up the SNPInfo into smaller files. The seqMeta R objects would then be combined within R. Another option us to use the thread option available for the matrix operations.

Null models

Determining the appropriate null model variance components can be done prior to the score test calculation using goodness of fit testing. For example addition of a dominance, X chromosome and/or mitochondria random effect to the additive effect can be tested using likelihood ratio test by starting with the additive component then including the other components one at a time. If the X chromosome effect does not provide a better fit (choose p-value cutoff) then the component is not needed in the null even if X chromosome variants are tested. Likewise, if the is significant X chromosome contribution to the phenotype, then the X random effect can be included when autosome variants are tested. Dominance, X and MT contributions are second order, so in general, their estimates may be zero. Similar goodness of fit testing can be done for environmental random effects such as sib or nuclear family.

Simulated test data

The file **MMAP.score.test.tar.gz** contains files and scripts to run MMAP on a small simulated data set. The genotype file has no missing data, so the algorithm to fill in missing data is not being tested. More example data sets and a tutorial are being developed.

Website: http://edn.som.umaryland.edu/mmap/index.php

Future directions:

1. Implement SNP set score tests as described in [4]. These tests focus on common variants in genes, pathways etc, and implements non-linear kernel functions for IBS sharing that are not derived as products of the SNP data matrix.
2. Implement the model used by rvtests and raremetalworker that are being used in GIANT exome chip analysis. The input formats and output are different than seqMeta, so a comparison is useful.
3. GLMM for binary traits.

References

1. Voorman A, Brody J, Chen H, Lumley T. seqMeta: An R package for meta-analyzing region-based tests of rare DNA variants 2013. Available from: http://cran.r-project.org/web/packages/seqMeta/.
2. Zhan X. rvtests 2014. Available from: https://github.com/zhanxw/rvtests.

3. Abacasis G. RAREMETALWORKER 2013. Available from: http://genome.sph.umich.edu/wiki/RAREMETALWORKER.
4. Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser PA, Lin X. SNP Set Association Analysis for Familial Data. Genetic epidemiology. 2012. doi: 10.1002/gepi.21676. PubMed PMID: 22968922; PubMed Central PMCID: PMC3683469.