Overview

About

OSCA (OmicS-data-based Complex trait Analysis) is a software tool written in C/C++ for the analysis of complex traits using multi-omics data. It is developed by Futao Zhang, Zhihong Zhu and Jian Yang at Institute for Molecular Bioscience, The University of Queensland. Bug reports or questions: Jian Yang <jian.yang@uq.edu.au>.

Functions currently supported are:

- Estimating the epigenetic (or transcriptomic) relationships between individuals from genome-wide DNA methylation (or gene expression) data.
- Estimating the proportion of phenotypic variance for a complex trait can be "explained" by all DNA methylation (or gene expression) probes.
- Mixed linear model based analysis to test for associations between DNA methylation (or gene expression) probes and complex traits.
- Estimating the joint "effects" of all methylation (transcription) probes on a phenotype (e.g. BMI) in a mixed linear model (analogous to BLUP). These estimated effects can be used to predict the phenotype in a new independent sample.
- eQTL/mQTL analysis with linear regression model. The model can adjust for the relevant covarites. The results are saved in BESD format.
- vQTL analysis to test the homogeneity of variance. Three algorithms have been implemented (Bartlett's test, Levene's test and Brown–Forsythe test, Fligner-Killeen test). The results are saved in BESD format.

Note: Although this software tool is designed for gene expression and DNA methylation data, it can be applied to any other source of omics data including microbiome, proteome and brain connectome.

Credits

Futao Zhang and Jian Yang developed the methods, software and webpage. Zhihong Zhu contributed to the development of the ORM and OREML methods. Ting Qi contributed to MeCS method. Huanwei Wang contributed to vQTL method. Zhili Zheng provided the template of the website.

Questions and Help Requests

Bug reports or questions to Jian Yang (jian.yang@uq.edu.au) at Institute for Molecular Bioscience, The University of Queensland.

Citation

Zhang F, Chen W, Zhu Z, Zhang Q, Deary IJ, Wray NR, Visscher PM, McRae AF, Yang J (2018) OSCA: a tool for omic-data-based complex trait analysis. bioRxiv 445163; doi: https://doi.org/10.1101/445163.

Download

Executable Files (version 0.43)

The Linux version is available at: osca_Linux.zip.

The MacOS version is available at: osca_Mac.zip.

The Windows version is available at: osca_Win.zip.

The example files are in example.zip.

The executable files (binary code) are release under MIT lincense.

Update log

6. Version 0.43 (18 January, 2019): added a function to run a variance quantitative trait locus (vQTL) analysis.
6. Version 0.42 (29 November, 2018): added an efficient function to run a mixed-linear-model-based cis-eQTL analysis.
5. Version 0.41 (22 October, 2018): added logistic regression in EWAS analysis.
4. Version 0.40 (14 August, 2018): added a new method called MOMENT for EWAS analysis and some functions for Data Management.
3. Version 0.39 (14 February, 2018): added a solution if the MLM estimation process in EWAS analysis does not work.
2. Version 0.38 (8 January, 2018): added meta analysis, eQTL analysis.
1. 21 July, 2017: first release.

Data Management

The DNA methylation (or gene expression) data are stored in a binary format for storage efficiency. We store the data in three separate files .oii (individual information, similar as a PLINK .fam file), .opi (probe information) and .bod (a binary file to store the DNA methylation or gene expression profiles).

To efficiently save the result of eQTL / vQTL analysis, OSCA also supports BESD format.

BOD format

 # BOD format: an efficient format to store omics data

myeed.oii

```
F01 I01 0   0   NA
F02 I01 0   0   NA
F03 I02 0   0   NA
...
```

Columns are family ID, individual ID, paternal ID, maternal ID and sex (1=male; 2=female; 0=unknown). Missing data are represented by "NA".

myeed.opi

```
1   probe101    924243  Gene01  +
1   probe102    939564  Gene01  -
1   probe103    1130681 Gene01  -
...
```

Columns are chromosome, probe ID (can be the ID of an exon or a transcript for RNA-seq data), physical position, gene ID and gene orientation.

myeed.bod

DNA methylation (or gene expression) data in binary format. Please do not try to open this file with a text editor.

Note: The text below are for advanced users only. We use the first 12 bytes of the .bod file to store the descriptive information for the data. The first 2 bytes are reserved for further extension of the software. The type of data is indicated by the 3rd byte (0 for gene expression data, 1 for DNA methylation data, and 2 for any other type of data). The type of value is indicated by the 4th byte (0 for DNA methylation beta value, 1 for DNA methylation m value and 2 for any other type of value). The number of individuals and number of probes are stored as integers (4 bytes each). For example, for a DNA methylation data set (0x01) of m values (0x01) on 1,342 (0x53e) individuals and 228,694 (0x37d56) CpG sites, the first 12 bytes of the .bod file are:

```
0101 0000 3e05 0000 567d 0300
```

The bytes afterwards are for the DNA methylation or gene expression data.

Make a BOD file

 # compile data in binary format from text format

```
 osca --efile myprofile.txt --methylation-beta --make-bod --out myprofile
```

--efile reads a DNA methylation (or gene expression) data file in plain text format.

--methylation-beta indicates methylation beta values in the file.

--make-bod saves DNA methylation (or gene expression) data in binary format.

--out saves data (or results) in a file.

myprofile.txt

```
FID  IID cg00000658  cg26036652  cg00489772  ...
F01 I01 0.909   0.845   0.41    ...
F01 I02 0.832   0.732   0.503   ...
...
```

This is a file with a header line that contains family ID, individual ID and names of probes. The column of family ID is optional. Please use the flag "**--no-fid**" for data without family ID.

```
osca --efile myprofile.txt  --methylation-beta --make-bod --no-fid --out myprofile
```

--no-fid indicates data without family ID.

```
osca --efile myprofile.txt  --methylation-m --make-bod --out myprofile
```

--methylation-m indicates DNA methylation m values in the file.

If the profile is of gene expression value.

```
osca --efile myprofile.txt  --gene-expression --make-bod --out myprofile
```

--gene-expression indicates gene expression profiles in the file.

For any other type of data:

```
osca --efile myprofile.txt --make-bod --out myprofile
```

A binary file also can be made from a transposed file in text format.

```
osca --tefile mytprofile.txt --methylation-beta --make-bod --no-fid --out myprofile
```

mytprofile.txt

```
FID  F01 F01 ...
IID I01 I02 ...
cg00000658  0.909   0.832   ...
```

```
cg26036652  0.845   0.732   ...
cg00489772  0.41    0.503   ...
...
```

The row of family ID is optional. Please use the flag "**--no-fid**" if there is no family ID in your data.

```
osca --tefile mytprofile.txt --methylation-beta --make-bod --no-fid --out myprofile
```

 # update probe annotation

Probe information are sometimes not available in the original DNA methylation (or gene expression) data. These information can be updated the command below.

```
osca --befile myprofile --update-opi annotated.opi
```

--befile reads a DNA methylation (or gene expression) data file in binary format.

--update-opi reads a fully annotated .opi file.

Manage BOD file(s)

 # make a subset

To extract a probe

```
osca --befile myprofile --probe cg00000658 --make-bod --out mysubprofile
```

--probe extracts a specified probe.

To exclude a probe

```
osca --befile myprofile --probe-rm cg00000658 --make-bod --out mysubprofile
```

--probe-rm excludes a specified probe.

To extract a subset of probes

```
osca --befile myprofile --extract-probe probe.list --make-bod --out mysubprofile
```

--extract-probe extracts a subset of probes.

probe.list

```
cg00000658
cg26036652
...
```

To exclude a subset of probes

```
osca --befile myprofile --exclude-probe probe.list --make-bod --out mysubprofile
```

--exclude-probe excludes a subset of probes.

To extract a subset of individuals

```
osca --befile myprofile --keep indi.list --make-bod --out mysubprofile
```

--keep extracts a subset of individuals.

indi.list

```
F01    I01
F01 I02
...
```

To exclude a subset of individuals

```
osca --befile myprofile --remove indi.list --make-bod --out mysubprofile
```

--remove excludes a subset of individuals.

To extract a subset of genes

```
osca --befile myprofile --genes gene.list --make-bod --out mysubprofile
```

--genes extracts a subset of genes.

gene.list

```
MAN1B1
EDEM2
...
```

To extract a gene

```
osca --befile myprofile --gene MAN1B1 --make-bod --out mysubprofile
```

--gene extracts a gene.

To extract a subset of probes based on the order

```
osca --befile myprofile --from-probe probe0 --to-probe probe1 --make-bod --out mysubprofile
```

--from-probe specifies the start probe.

--to-probe specifies the end probe.

To extract a subset of probes in a genomic region centred at a specific probe

```
osca --befile myprofile --probe cg00000658 --probe-wind 500 --make-bod --out mysubprofile
```

--probe-wind defines a window \<kb> centred on a specified probe.

To extract a subset of probes on a chromosome

```
osca --befile myprofile --chr 1 --make-bod --out mysubprofile
```

--chr specifies a chromosome to select probes.

To extract a subset of probes based on physical positions

```
osca --befile myprofile --from-probe-kb 100 --to-probe-kb 200 --make-bod --out mysubprofile
```

--from-probe-kb specifies the start physical position of the probes.

--to-probe-kb specifies the end physical position of the probes.

 # merge BOD files

```
osca --befile-flist mybod.flist --make-bod --out myprofile
```

--befile-flist reads a file to get the full paths of the binary files.

mybod.flist

```
path1/my_bod1
path2/my_bod2
...
```

 # Quality control

To remvoe probes with low variance

```
osca --befile myprofile --sd-min 0.02 --make-bod --out newprofile
```

--sd-min removes the probes with the standard deviation smaller than a specified threshold.

To remove probes with a missing rate threshold

```
osca --befile myprofile --missing-ratio-probe 0.01 --make-bod --out newprofile
```

--missing-ratio-probe specifies a missing proportion threshold to remove probes.

 # Quality control of methyaltion data

To remove constitutively methylated/unmethylated probes

```
osca --befile myprofile --upper-beta 0.8 --lower-beta 0.2 --make-bod --out newprofile
```

--upper-beta removes the DNA methylation probes with the mean beta value larger than a specified threshold.

--lower-beta removes the DNA methylation probes with the mean beta value smaller than a specified threshold.

To QC with detection p-values

```
osca --befile myprofile --detection-pval-file dpval.txt --dpval-mth 0  --dpval-thresh 0.05 --make-bod --out newprofile
```

--detection-pval-file reads a file that contains DNA methylation detection p-values.

--dpval-mth specifies a method to do quality control with the detection p-values, 0 for removing the probes with one or more detection p-values violating a threshold (default as 0.05), and 1 for dropping the samples violating a proportion threshold (default as 1%) and simultaneously dropping the probes violating a proportion threshold (default as 1%).

--dpval-thresh specifies a threshold of detection p-value.

```
osca --befile myprofile --detection-pval-file dpval.txt --dpval-mth 1  --ratio-probe 0.01 --ratio-sample 0.01 --dpval-thresh 0.05 --make-bod --out newprofile
```

--ratio-probe specifies a proportion threshold to remove probes.

--ratio-samplespecifies a proportion threshold to remove individuals.

the file "dpval.txt" is in the same format with a transposed profile file. Option "**--no-fid**" is also valid here.

 # adjust the covariates for each probe

```
osca --befile myprofile --covar my.cov --qcovar my.qcov --adj-probe --make-bod --out newprofile
```

--adj-probe adjusts the covariates for each probe.

 # standardize each probe

```
osca --befile myprofile --std-probe --make-bod --out newprofile
```

--std-probe standardizes each probe.

```
osca --befile myprofile --rint-probe --make-bod --out newprofile
```

--rint-probe normalizes each probe by Rank-based Inverse Normal Transformation.

 # other options

Transform methylation beta value to methylation m value and vice versa

```
osca --befile myprofile --m2beta --make-bod --out newprofile
```

--m2beta calculates the methylation beta value from the methylation m value.

```
osca --befile myprofile --beta2m --make-bod --out newprofile
```

--beta2m calculates the methylation m value from the methylation beta value.

Query a BOD file

 # make a text format file

```
osca --befile myprofile --make-efile --out myprofile.txt
```

--make-efile saves the DNA methylation (or gene expression) data in text format.

```
osca --befile myprofile --make-tefile --out mytprofile.txt
```

--make-tefile saves the DNA methylation (or gene expression) data in transposed text format.

NOTE: The options in Manage BOD file(s) can also be applied.

 # Calculate the variance and the mean of probes

```
osca --befile myprofile --get-variance --get-mean --out newprofile
```

--get-variance calculates the variance of each probe.

--get-mean calculates the mean of each probe.

newprofile.var.txt

```
cg00000957 0.000812127
cg00001349  0.00560701
...
```

newprofile.mean.txt

```
cg00000957 0.901574
cg00001349  0.860279
...
```

BESD format

Results from eQTL analysis will be saved in SMR BESD format see SMR website for more information.To save the eQTL results more efficiently, we extended the BESD format.

Note: The texts below are for advanced users only.

BESD dense format 1: the first 64 Bytes are reserved for the descriptive information which starts with 0x00000005. The data onwards are a vector of effect sizes followed by a vector of SEs of each probe across all the snps. The effect sizes and SEs are saved in single float-precision (4B).

BESD dense format 2 (only supported in OSCA): the first 64 Bytes are reserved for the descriptive information which starts with 0x00000004. The data onwards are the effect sizes followed by the SEs of each SNP across all the probes. The effect sizes and SEs are saved in single float-precision (4B).

BESD sparse format: the first 64 Bytes are reserved for the descriptive information which starts with 0x00000001. The effect sizes and the SEs as saved in the CSC format

---

Query a BESD file

This feasuer is memory efficent even to query with huge dense BESD file.

 # Command line options for SNPs

To query the eQTL resutls for a single SNP, we could use this command

```
osca --beqtl-summary myeqtl --query 5.0e-8 --snp rs123 --out myquery
```

--query saves in text format a subset of the eQTL summary dataset based on the specified eQTL p-value threshold. The default value is 5.0e-8.

--snp specifies a single SNP.

myquery.txt

```
SNP    Chr BP  A1  A2  Freq    Probe   Probe_Chr   Probe_bp    Gene    Orientation b   se  p
rs01    1   1001    A   G   0.23    cg01    1   1101    gene1   +   -0.033  0.006   3.8e-08
rs01    1   1001    A   G   0.06    cg02    1   1201    gene2   -   0.043   0.007   8.1e-10
......
```

To query the eQTL resutls excluding a single SNP, we could use this command

```
osca --beqtl-summary myeqtl --query 5.0e-8 --snp-rm rs123 --out myquery
```

--snp-rm specifies a single SNP to exclude.

To query the eQTL resutls extracting a subset of SNPs , we could use this command

```
osca --beqtl-summary myeqtl --query 5.0e-8 --extract-snp mysnp.list --out myquery
```

--extract-snp extracts a subset of SNPs for analysis.

To query the eQTL resutls excluding a subset of SNPs , we could use this command

```
osca --beqtl-summary myeqtl --query 5.0e-8 --exclude-snp mysnp.list --out myquery
```

--exclude-snp excludes a subset of SNPs from analysis.

To query eQTL resutls for a range of SNPs in a genomic region

```
osca --beqtl-summary myeqtl --query 5.0e-8 --from-snp rs123 --to-snp rs456 --out myquery
```

--from-snp specifies the start SNP.

--to-snp specifies the end SNP.

NOTE : All SNPs should be on the same chromosome.

To query eQTL results for all SNP on a chromosome

```
osca --beqtl-summary myeqtl --query 5.0e-8 --snp-chr 1
```

--snp-chr specifies a chromosome to select SNPs.

NOTE : The probes in the result could be on the other chromosomes if there are *trans*-eQTLs.

To query SNPs based on physical positions

```
osca --beqtl-summary myeqtl --query 5.0e-8 --snp-chr 1 --from-snp-kb 100 --to-snp-kb 200 --out myquery
```

--from-snp-kb specifies the start physical position of the region.

--to-snp-kb specifies the end physical position of the region.

NOTE : You will need to specify a chromosome (using the '--snp-chr' option) when using this option.

To query based on a flanking region of a SNP

```
osca --beqtl-summary myeqtl --query 5.0e-8 --snp rs123 --snp-wind 50 --out myquery
```

--snp-wind defines a window centred on a specified SNP.

 # Command line options for probes

To query based on a single probe

```
osca --beqtl-summary myeqtl --query 5.0e-8 --probe cg123 --out myquery
```

--probe specifies a single probe.

To query excluding a single probe

```
osca --beqtl-summary myeqtl --query 5.0e-8 --probe-rm cg123 --out myquery
```

--probe-rm specifies a single probe to exclude.

To query the eQTL resutls extracting a subset of probes , we could use this command

```
osca --beqtl-summary myeqtl --query 5.0e-8 --extract-probe myprobe.list --out myquery
```

--extract-probe extracts a subset of probes for analysis.

To query the eQTL resutls excluding a subset of probes , we could use this command

```
osca --beqtl-summary myeqtl --query 5.0e-8 --exclude-probe myprobe.list --out myquery
```

--exclude-probe excludes a subset of probes from analysis.

To query based on a range of probes

```
osca --beqtl-summary myeqtl --query 5.0e-8 --from-probe cg123 --to-probe cg456 --out myquery
```

--from-probe specifies the start probe.

--to-probe specifies the end probe.

NOTE : All probes should be on the same chromosome.

To query based on a chromosome

```
osca --beqtl-summary myeqtl --query 5.0e-8 --probe-chr 1
```

--probe-chr specifies a chromosome to select probes.

NOTE : The SNPs in the result could be on the other chromosomes if there are *trans*-eQTLs.

To query based on physical positions of the probes

```
osca --beqtl-summary myeqtl --query 5.0e-8 --probe-chr 1 --from-probe-kb 1000 --to-probe-kb 2000 --out myquery
```

--from-probe-kb specifies the start physical position of the probes.

--to-probe-kb specifies the end physical position of the probes.

NOTE : You will need to specify a chromosome (using the '--probe-chr' option) when using this option.

To query based on a flanking region of a probe

```
osca --beqtl-summary myeqtl --query 5.0e-8 --probe cg123 --probe-wind 1000 --out myquery
```

--probe-wind defines a window centred on a specified probe.

To query based on a gene

```
osca --beqtl-summary myeqtl --query 5.0e-8 --gene gene1 --out myquery
```

--gene specifies a single gene to select probes.

 # Command line option for cis-region

```
osca --beqtl-summary myeqtl --query 5.0e-8 --probe cg123 --cis-wind 2000 --out myquery
```

 # File-list options

To query based on a list of SNPs

```
osca --beqtl-summary myeqtl --extract-snp snp.list --query 5.0e-8 --out myquery
```

To query based on a list of probes

```
osca --beqtl-summary myeqtl --extract-probe probe.list --query 5.0e-8 --out myquery
```

To qurey based on a list of genes

```
osca --beqtl-summary myeqtl --genes gene.list --query 5.0e-8 --out myquery
```

--genes extracts a subset of probes which tag the genes in the list.

gene.list

```
gene1
gene2
gene3
...
```

Manage a BESD file

 # make a subset of data

All the data management options in Query a BESD file can be here.

for example, To extract a subset of SNPs and/or probes

```
osca --beqtl-summary myeqtl --extract-snp mysnp.list --extract-probe myprobe.list  --make-besd --out mybesd
```

 # tansform to SMR compatible format

```
osca --beqtl-summary myeqtl --make-besd --to-smr --out mybesd
```

--to-smr tranforms BESD file to SMR compatible format.

 # shrink a BESD file

To remove probes without any value across all the SNPs and SNPs without any value across all the probes.

```
osca --beqtl-summary myeqtl --make-besd --besd-shrink --out mybesd
```

--besd-shrink removes probes and SNPs that have no value.

ORM

estimate ORM

```
osca --befile myprofile --make-orm --out myorm
```

```
osca --befile myprofile --make-orm-bin --out myorm
```

--make-orm or **--make-orm-bin** estimates the omics relationship matrix (ORM) between pairs of individuals from a set of probes and save the lower triangle elements of ORM to binary files. This format is compatible with the GRM in the GCTA software.

```
osca --befile myprofile --make-orm-gz --out myorm
```

--make-orm-gz estimates the ORM and save the lower triangle elements of ORM to compressed plain text files.

```
osca --befile myprofile --make-orm --orm-alg 1 --out myorm
```

--orm-alg specifies the algorithm to estimate the ORM. 1 for standardized data of each probe, 2 for centred data of each probe and 3 for iteratively standardizing probes and individuals. The default option is 1.

Note that although we describe the options above using DNA methylation and gene expression data, all the options can be applied to any other source of omics data mentioned above.

Principal Component Analysis

```
osca --orm myorm --pca 20 --out mypca
```

--orm reads the ORM binary files.

--pca conducts principal component analysis and saves the first n (default as 20) PCs.

mypca.eigenval

```
122.014
95.3064
70.8055
...
```

mypca.eigenvec

```
R06C01 R06C01  0.012637    -0.00328532 0.0263842   -0.00595246
R05C02  R05C02  0.0106692   -0.0184545  -0.0159826  0.013951
R04C02  R04C02  0.0024727   -0.0118185  -0.0256503  -0.0106235
...
```

Users can also manipulate the ORM in the analysis.

```
osca --orm myorm --keep indi.list --pca 20 --out mypca
```

```
osca --orm myorm --remove indi.list --pca 20 --out mypca
```

```
osca --orm myorm --orm-cutoff 0.05 --pca 20 --out mypca
```

--orm-cutoff removes one of a pair of individuals with estimated omics relationships larger than the specified cut-off value.

```
osca --multi-orm myorm.flist --pca 20 --out mypca
```

--multi-orm reads multiple ORMs in binary format.

OREML

```
osca --reml --pheno my.phen --out myreml
```

--reml performs REML (restricted maximum likelihood) analysis. This option is usually followed by the option --orm (one ORM) or --merge-orm (multiple ORMs) to estimate the variance explained by the probes that were used to estimate the omics relationship matrix.

--pheno reads phenotype data from a plain text file. Missing value should be represented by "NA".

my.phen

```
R06C01 R06C01  32.6332
R05C02  R05C02  23.9411
R04C02  R04C02  29.7441
...
```

NOTE: current version only supports single trait analysis in one run.

```
osca --reml --orm myorm --pheno my.phen --out myreml
```

```
osca --reml --orm myorm --pheno my.phen --keep indi.list --out myreml
```

```
osca --reml --orm myorm --pheno my.phen --orm-cutoff 0.05 --out myreml
```

With multiple ORMs

```
osca --reml --multi-orm myorm.flist --pheno my.phen --out myreml
```

```
osca --reml --multi-orm myorm.flist --pheno my.phen --reml-alg 0 --out myreml
```

--reml-alg specifies the algorithm to do REML iterations, 0 for average information (AI), 1 for Fisher-scoring and 2 for EM. The default option is 0, i.e. AI-REML, if this option is not specified.

```
osca --reml --multi-orm myorm.flist --pheno my.phen --reml-maxit 100 --out myreml
```

--reml-maxit specifies the maximum number of iterations. The default number is 100 if this option is not specified.

```
osca --reml --orm myorm --pheno my.phen --covar my.covar --out myreml
```

--covar reads discrete covariates from a plain text file.

my.covar

```
R06C01 R06C01  F   0
R05C02  R05C02  F   1
R04C02  R04C02  M   1
...
```

```
osca --reml --orm myorm --pheno my.phen --qcovar my.qcovar --out myreml
```

--qcovar reads quantitative covariates from a plain text file.

my.qcovar

```
R06C01 R06C01  25
R05C02  R05C02  16
R04C02  R04C02  30
...
```

```
osca --reml --orm myorm --pheno my.phen --reml-est-fix --out myreml
```

--reml-est-fix displays the estimates of fixed effects on the screen.

```
osca --reml --orm myorm --pheno my.phen --reml-no-lrt --out myreml
```

--reml-no-lrt turns off the LRT.

EWAS

MOMENT

 # MOMENT (Multi-cOmponent MIm-based association ExcludiNg the Target)

```
osca --moment --befile myprofile --pheno my.phen --out my
```

--moment initiates a multi-component MLM based association analysis excluding the target probe (the probe to be tested for association) and the probes in its flanking region from the ORM (the size of flanking region can be modified by --moment-wind). The results will be saved in a plain text file with *.mlma. as the filename extension.

```
osca --moment --befile myprofile --pheno my.phen --moment-wind 100 --out my
```

--moment-wind specifies a flanking region to exclude probes from the ORM. The default value is 100 Kb (i.e. a 100 Kb region centred at the probe to be tested).

my.mlma

```
Chr    Probe   bp  Gene    Orientation b   se  p
1   cg00003287  201346149   TNNT2   -   -0.156  0.597   0.794
1   cg00008647  207082900   IL24    +   0.032   0.354   0.926
1   cg00009292  50882082    DMRTA2  -   0.120   1.182   0.919
...
```

This is a text file with headers. Columns are chromosome, probe, probe BP, gene, orientation, effect size, standard error and p-value.

NOTE: Sometimes the MLM estimation process in the EWAS analysis does not work (the REML iteration cannot converge or an estimate hits the boundaries of parameter space) especially when the sample size is small. In such case we switch MLMA to standard PC based linear regression analysis. We have implemented a strategy to identify the number of PCs iteratively until the genomic inflation factor (lambda value) falls into a user-defined range, or the number of PCs reaches the lower (0) or upper boundary (half of the sample size). The range size for lambda can be specified by the option **--lambda-range**

```
osca --moment --befile myprofile --pheno my.phen --lambda-range 0.05 --out my
```

--lambda-range specifies a range for the lambda value. The default value is 0.05. This option only works when the MLM estimation process fails.

NOTE: The analysis becomes slower as the number of PCs becomes larger. The option **--fast-linear** can accelerate the analysis by pre-adjusting both the trait and gene expression (or DNA methylation) value by the PCs. In this case, all the missing gene expression (or DNA methylation) values will be replaced by the mean. It should be noted that if the proportion of missing values in the gene expression (or DNA methylation) data is large, this accelerated analysis may lead to substantial differences in results.

```
osca --moment --befile myprofile --pheno my.phen --fast-linear --out my
```

--fast-linear runs a fast linear regression analysis. This flag can also be used in the Fast Linear Regression analysis module

---

MOA

 # MOA (MLM-based Omic Association)

```
osca --moa --befile myprofile --pheno my.phen --out my
```

If you have already computed the ORM

```
osca --moa --befile myprofile --pheno my.phen --orm myorm --out my
```

--moa initiates an MLM based association analysis including the target probe (the probe to be tested for association) in the ORM. The results will be saved in a plain text file with **.mlma** as the filename extension.

Linear Regression

```
osca --befile myprofile --pheno my.phen --linear --out my
```

```
osca --befile myprofile --pheno my.phen --qcovar my.qcovar --covar my.covar --linear --out my
```

--linear saves linear regression statistics to a plain text file.

my.linear

```
probeChr   ProbeID Probe_bp Gene    Orientation   BETA    SE  P   NMISS
1   cg00003287  201346149   TNNT2   -   -0.0594 0.531   9.11e-01    1337
1   cg00008647  207082900   IL24    +   0.5003  0.263   5.72e-02    1337
1   cg00009292  50882082    DMRTA2  -    1.0512  1.026   3.06e-01    1337
...
```

This is a text file with headers. Columns are chromosome, probe, probe BP, gene, orientation, effect size, standard error, p-value and number of non-missing individuals.

Fast Linear Regression

The options above fit the covariates in the model as y=xb+Cβ+e. This is equivalent to $y'=y-C\beta^{\wedge}$, $x'=x-C\beta^{\wedge}$, $y'=x'b$. This equivalent transformation avoids the repeated computing of the inverse of a (p +1)*(p + 1) matrix where p is the number of covariates.

```
osca --linear --befile myprofile --pheno my.phen --qcovar my.qcovar --covar my.covar --fast-linear --out my
```

--fast-linear runs a fast linear regression analysis.

Logistic Regression

```
osca --befile myprofile --pheno my.phen --logistic --out my
```

```
osca --befile myprofile --pheno my.phen --qcovar my.qcovar --covar my.covar --logistic --out my
```

--logistic outputs logistic regression analysis result to a plain text file.

my.logistic

```
probeChr   ProbeID Probe_bp Gene    Orientation   OR    SE  P   NMISS
1    cg00297950  110282525   GSTM3   -   0.010953    1.30438 5.386424e-04    1318
1    cg00299820  11943154    NPPB    -   0.0085006   2.77659 8.596523e-02    1318
1    cg00305285  1017115 RNF223  -   8.22097 2.99934 4.824397e-01    1318
...
```

This is a text file with headers. Columns are chromosome, probe, probe BP, gene, orientation, odd ratio, standard error, p-value and number of non-missing individuals.

NOTE: It is allowed to use characters, strings, or numbers as phenotype values in logistic regression but "NA" or "na" will be recognised as a missing value.

EWAS simulation

The phenotypes are simulated based on a set of real DNA methylation (or gene expression) data and a simple model y= sum($x_i b_i$)+ε, where y is a vector of phenotypes, x_i is a vector of raw DNA methylation (or gene expression) profile or standardized profile of the i-th "causal" probe, b_i is the effect of the i-th causal probe and ε ~ N(0,var(sum($x_i b_i$))(1/h^2-1)) is a vector of residual effect.

```
osca --simu-qt --simu-hsq 0.1 --befile myprofile --simu-causal-loci mycausal.list --out mypheno
```

--simu-qt smulates a quantitative trait.

--simu-hsq specifies the proportion of variance in phenotype explained by the causal probes. The default value is 0.1.

--simu-causal-loci reads a list of probes as causal probes. If the effect sizes are not specified in the file, they will be generated from a standard normal distribution.

mycausal.list

```
cg04584301 1.55182
cg04839274  -0.106226
cg16648571  0.0257417
...
```

This is a text file with no headers. Columns are probe ID and effect size.

```
osca --simu-qt --simu-hsq 0.1 --simu-eff-mod 0 --befile myprofile --simu-causal-loci mycausal.list --out mypheno
```

--simu-eff-mod specifies whether or not to standardize the causal probe, 0 for standardized profile and 1 for raw profile. The default value is 0.

```
osca --simu-cc 100 300 --simu-hsq 0.1 --simu-k 0.1 --befile myprofile --simu-causal-loci mycausal.list --out mypheno
```

--simu-cc simulates a case-control trait and specifies the number of cases and the number of controls.

--simu-k specifies the disease prevalence. The default value is 0.1 if this option is not specified.

Prediction Analysis

```
osca --reml --orm myorm --pheno my.phen --reml-pred-rand --out myblp
```

--reml-pred-rand predicts the random effects by the BLUP (best linear unbiased prediction) method. This option is to estimate the aggregated effect of all the probes (used to compute the ORM) to the phenotype of an individual. The aggregated omics effects of all the individuals will be saved in a plain text file ***.indi.blp**.

myblp.indi.blp

```
R06C01 R06C0   1.02275 0.07065 1.05692 1.05692
R05C02  R05C02  0.18653 0.27059 0.86650 0.86650
R04C02  R04C02  -0.1982 -0.1673 -0.9209 -0.9209
...
```

This is a text file with no headers. Columns are family ID, individual ID, an intermediate variable, the aggregated omics effect, another intermediate variable and the residual effect.

```
osca --befile myprofile --blup-probe myblp.indi.blp --out myblp
```

--blup-probe calculates the BLUP solutions for the probe effects.

myblp.probe.blp

```
cg04584301 -0.000654646
cg04839274  0.000602484
cg16648571  -4.70356e-05
...
```

This is a text file with no headers. Columns are probe ID and BLUP of the probe effect.

```
osca --befile myprofile --score myblp.probe.blp --out myscore
```

```
osca --befile myprofile --score myblp.probe.blp 1 2 --out myscore
```

--score reads score files for probes and generates predicted omics profiles for individuals. (Note that this option largely follows the --score option in PLINK.) It allows users to specify the column numbers for probe ID and score (the default values are 1 and 2 as shown in the example above).

myscore.profile

```
FID    IID PHENO   CNT SCORE
131000028   422572  -9  20000   -7.269198e-07
131000031   243421  -9  20000   9.322096e-06
131000179   338728  -9  20000   -1.250443e-05
...
```

This is a text file with headers. Columns are family ID, individual ID, phenotype, Number of non-missing probes and score.

```
For example
In the score file:
cg04584301  -0.065
cg04839274  0.060
cg16648571  -1.03

In the DNA methylation data:
FID IID cg04584301  cg04839274  cg16648571
R05C02  R05C02  0.18653 0.27059 0.86650

The score should be:
( 0.18653*(-0.065) + 0.27059*0.060 + (-1.03)*0.86650 ) / 3 = (-0.8883841) / 3 = -0.296
```

```
osca --befile myprofile --score myblp.probe.blp --score-has-header --out myscore
```

--score-has-header indicates probe score file has headers.

eQTL/mQTL Analysis

This is a module in OSCA to perform a standard eQTL mapping analysis. We describe the module for eQTL mapping but it can be applied to genetic association analysis for all kinds of molecular traits (e.g. DNA methylation, histone modification, metabolites and Microbiome)

Linear-regression-based

 # Basic option

```
osca --eqtl --bfile mydata --befile myprofile --out myeqtl
```

--eqtl enables the eQTL mapping analysis.

--bfile reads individual-level SNP genotype data (in PLINK binary format), i.e. .bed, .bim, and .fam files.

By default, the results will be saved in BESD format .

 # Multi-threading computation

```
osca --eqtl --bfile mydata --befile myprofile --task-num 100 --task-id 1 --thread-num 10 --out myeqtl
```

--task-num specifies the total number of tasks to partition the computation by probes.

--task-id specifies the task IDs form 1 to the total task number.

--thread-num specifies the number of threads for parallel computing. The default value is 1.

 # Using other data management options

```
osca --eqtl --bfile mydata --befile myprofile --extract-snp mysnp.list --maf 0.01 --task-num 100 --task-id 1 --thread-num 10 --out
 myeqtl
```

--maf filters SNPs based on the specified MAF threshold.

 # Saving the result in SMR BESD format.

```
osca --eqtl --bfile mydata --befile myprofile --to-smr --task-num 100 --task-id 1 --thread-num 10 --out myeqtl
```

--to-smr saves the result in SMR compatible format.

 # Fitting covariates

```
osca --eqtl --bfile mydata --befile myprofile --covar mycovar --qcovar myqcovar --task-num 100 --task-id 1 --thread-num 10 --out my
eqtl
```

 # runing cis-eQTL analysis

```
osca --eqtl --bfile mydata --befile myprofile --cis --task-num 1000 --task-id 1 --thread-num 10 --out myeqtl
```

```
osca --eqtl --bfile mydata --befile myprofile --cis --cis-wind 2000 --task-num 100 --task-id 1 --thread-num 10 --out myeqtl
```

--cis runs a cis-eQTL analysis.

--cis-wind specifies a window (in Kb unit) to store all the SNPs within the window of the probe in either direction. The default value is 2000Kb.

Mixed-linear-model-based

 # runing cis-eQTL analysis

```
osca --eqtl --bfile mydata --befile myprofile --mlm --cis --task-num 100 --task-id 1 --thread-num 10 --out myeqtl
```

--mlm runs a mixed-linear-model-based eQTL analysis on mixed linear model.

```
osca --eqtl --bfile mydata --befile myprofile --mlm --cis --grm mydata --task-num 100 --task-id 1 --thread-num 10 --out myeqtl
```

--grm reads the GRM generated by GCTA **--make-grm** option.

vQTL Analysis

This is a module in OSCA to perform a variance quantitative trait locus (vQTL) mapping analysis and save the results in BESD format. We have provided 4 methods to run vQTL analysis, namely Bartlett's test, Levene's test with mean, Levene's test with median, and Fligner-Killeen test. In the output file, apart from the vQTL statistic and p-value, we also provide the vQTL effect (i.e., the effect of a SNP on differences in phenotype variance among genotype groups), computed from z-statistic using the method described in (Zhu et al. 2016

Nature Genetics). Note that the vQTL z-statistic can be computed from p-value and the sign of the slope of regressing phenotype variance against genotypes.

 # Basic option

```
osca --vqtl --bfile mydata --pheno mypheno --out myvqtl
```

--vqtl to turn a vQTL analysis.

--bfile to input SNP genotype data.

--pheno to input phenotype data (see OREML for the input format).

myvqtl.vqtl

```
Chr    SNP    bp    statistic    df    beta    se    P    NMISS
21    rs144022851    14589985    2.98808    1    -0.246533    0.142619    8.387947e-02    1319
21    rs146286292    14592960    2.99122    1    -0.244279    0.141241    8.371710e-02    1319
21    rs2847443    14595264    2.40285    1    -0.209187    0.134949    1.211141e-01    1319
21    rs192179023    14600255    0.00383907    1    0.0357931    0.577678    9.505945e-01    1319
21    rs9984084    14602180    5.21829    2    0.0696413    0.0389252    7.359741e-02    1319
...
```

This is a text file with headers. Columns are chromosome, SNP, SNP BP, the statistic, degree of freedom, effect size, standard error, p-value and number of non-missing individuals.

 # Specifying the method to test variance heterogeneity

```
osca --eqtl --bfile mydata --pheno mypheno --vqtl-mtd 1 --out myvqtl
```

--vqtl-mtd to specify the method to test variance heterogeneity. 0 for Bartlett's test, 1 for Levene's test with mean, 2 for Levene's test with median, 3 for Fligner-Killeen test. The default option is 0.

myvqtl.vqtl

```
Chr    SNP    bp    F-statistic    df1    df2    beta    se    P    NMISS
21    rs144022851    14589985    2.03194    1    1317    -0.203255    0.142671    0.154261    1319
21    rs188453282    14591213    1.48959    1    1317    1.21945    0.999626    0.222499    1319
21    rs146286292    14592960    1.84845    1    1317    -0.192008    0.141303    0.174196    1319
21    rs2847443    14595264    1.83381    1    1317    -0.182687    0.134978    0.17591    1319
21    rs192179023    14600255    0.00102531    1    1317    0.018494    0.577679    0.974461    1319
21    rs9984084    14602180    1.73809    2    1316    0.0526684    0.0389454    0.176259    1319
...
```

This is a text file with headers. Columns are chromosome, SNP, SNP BP, F-statistic, K-1 degrees of freedom, N-K degrees of freedom, effect size, standard error, p-value and number of non-missing individuals.

 # Including data management options

```
osca --vqtl --bfile mydata --pheno mypheno --extract-snp mysnp.list --maf 0.01 --out myvqtl
```

 # Running vQTL analysis for molecular phenotypes

```
osca --vqtl --bfile mydata --befile myprofile --out myvqtl
```

--befile to input molecular phenotypes (e.g. DNA methylation or gene expression measures in BOD format).

 # Running vQTL analysis for molecular phenotypes in cis-regions

```
osca --vqtl --bfile mydata --befile myprofile --cis-wind 2000 --out myvqtl
```

--cis-wind to define a window centred around the probe to select SNPs for the vQTL analysis. The default value is 2000Kb.

Note that vQTL results for molecular traits will be saved in BESD format which can be tansformed to text format using OSCA query.

 # Multi-threading computation

```
osca --vqtl --bfile mydata --befile myprofile --cis-wind 2000 --task-num 1000 --task-id 1 --thread-num 10 --out myvqtl
```

Citation

Huanwei Wang, Futao Zhang, Jian Zeng, Yang Wu, Kathryn E. Kemper, Angli Xue, Min Zhang, Joseph E. Powell, Michael E. Goddard, Naomi R. Wray, Peter M. Visscher, Allan F. McRae, Jian Yang (2019) Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. bioRxiv 519538; doi: https://doi.org/10.1101/519538

Meta-analysis

Meta-analysis

 # Meta for GWAS summary data without sample overlap

```
osca --gwas-flist mygwas.flist --meta --out mymeta
```

--meta implements the conventional inverse-variance-weighted meta-analysis meta-analysis assuming all the cohorts are independent. Pleae refer to de Bakker PI et al.2008 Hum Mol Genet for the details.

--gwas-flist reads a file to get file paths of the GWAS summary data.

mygwas.flist

```
Height.01.COJO
Height.02.COJO
Height.03.COJO
...
```

This file has no header.

The input format of the GWAS summary data follows that for GCTA-COJO analysis (http://cnsgenomics.com/software/gcta/#COJO).

Height.01.COJO

```
SNP    A1  A2  freq    b    se  p    n
rs1001    A   G   0.8493  0.0024  0.0055  0.6653  129850
rs1002    C   G   0.03606 0.0034  0.0115  0.7659  129799
rs1003    A   C   0.5128  0.045   0.038   0.2319  129830
......
```

Columns are SNP, the effect (coded) allele, the other allele, frequency of the effect allele, effect size, standard error, p-value and sample size. The headers are not keywords and will be omitted by the program. Important: “A1” needs to be the effect allele with “A2” being the other allele and “freq” needs to be the frequency of “A1”. **NOTE: For a case-control study, the effect size should be log(odds ratio) with its corresponding standard error**.

 # Meta for eQTL summary data

```
osca --besd-flist mybesd.flist --meta --out mymeta
```

--besd-flist reads a file to get the full paths of the BESD files.

mybesd.flist

```
path1/my_besd1
path2/my_besd2
path3/my_besd3
...
```

This file has no header. The eQTL summary data should be in BESD formta.

MeCS

 # MeCS for eQTL summary data in correlated samples

MeCS is a method that only requires summary-level cis-eQTL data to perform a meta-analysis of cis-eQTLs from multiple cohorts (or tissues) with sample overlaps. It estimates the proportion of sample overlap from null SNPs in the cis-regions and meta-analysed the eQTL effects using a generalized least squares approach. The method can be applied to data from genetic studies of molecular phenotypes (e.g. DNA methylation and histone modification).

NOTE: Only the information in the cis-region would be used.

```
osca --besd-flist mybesd.flist --mecs --out mymecs
```

Options Reference

--befile	reads a DNA methylation (or gene expression) data file in binary format
--befile-flist	reads a file to get the full paths of the binary files
--beta2m	calculates the methylation m value from the methylation beta value
--blup-probe	calculates the BLUP solutions for the probe effects
--chr	specifies a chromosome to select probes
--covar	reads discrete covariates from a plain text file
--detection-pval-file	reads a file that contains DNA methylation detection p-values
--dpval-mth	specifies a method to do quality control with the detection p-values
--dpval-thresh	specifies a threshold of detection p-value
--efile	reads a DNA methylation (or gene expression) data file in plain text format
--exclude-probe	excludes a subset of probes
--extract-probe	extracts a subset of probes
--from-probe	specifies the start probe
--from-probe-kb	specifies the start physical position of the probes
--gene	extracts a gene
--genes	extracts a subset of genes
--gene-expression	indicates gene expression profiles in the file
--get-mean	calculates the mean of each probe
--get-variance	calculates the variance of each probe
--keep	extracts a subset of individuals
--linear	saves linear regression statistics to a plain text file
--lower-beta	removes the DNA methylation probes with the mean beta value smaller than a specified threshold
--lxpo	specifies a percentage of probes to exclude from calculating the ORM
--m2beta	calculates the methylation beta value from the methylation m value
--make-bod	saves DNA methylation (or gene expression) data in binary format
--make-efile	saves the DNA methylation (or gene expression) data in text format
--make-orm	estimates the omics relationship matrix (ORM) and save the lower triangle elements of ORM to binary files
--make-orm-bin	estimates the omics relationship matrix (ORM) and save the lower triangle elements of ORM to binary files
--make-orm-gz	estimates the omics relationship matrix (ORM) and save the lower triangle elements of ORM to compressed plain text files
--make-tefile	saves the DNA methylation (or gene expression) data in transposed text format
--merge-orm	reads multiple ORMs in binary format
--methylation-m	indicates methylation m values in the file
--methylation-beta	indicates methylation beta values in the file
--mlma	initiates an MLM based association analysis including the target probe (the probe to be tested for association) in the ORM
--mlma-loco	initiates an MLM based association analysis with the chromosome where the target probe is located excluded from the ORM
--mpheno	reads a list of comma-delimited trait numbers if the phenotype file contains more than one trait
--missing-ratio-probe	specifies a missing proportion threshold to remove probes
--no-fid	indicates data without family ID
--orm	reads the ORM binary files
--orm-alg	specifies the algorithm to estimate the ORM
--orm-cutoff	removes one of a pair of individuals with estimated omics relationships larger than the specified cut-off value
--out	saves data (or results) in a file
--pca	conducts principal component analysis and saves the first n (default as 20) PCs
--pheno	reads phenotype data from a plain text file
--probe	extracts a specified probe
--probe-wind	defines a window centred on a specified probe
--probe-rm	excludes a specified probe
--qcovar	reads quantitative covariates from a plain text file
--ratio-probe	specifies a proportion threshold to remove probes
--ratio-sample	specifies a proportion threshold to remove individuals
--reml	performs REML (restricted maximum likelihood) analysis
--reml-alg	specifies the algorithm to do REML iterations
--reml-est-fix	displays the estimates of fixed effects on the screen
--reml-maxit	specifies the maximum number of iterations
--reml-no-lrt	turns off the LRT
--reml-pred-rand	predicts the random effects by the BLUP (best linear unbiased prediction) method
--remove	excludes a subset of individuals
--score	reads score files for probes and generates predicted omics profiles for individuals
--score-has-header	indicates probe score file has headers
--simu-causal-loci	reads a list of probes as causal probes
--simu-cc	simulates a case-control trait and specifies the number of cases and the number of controls
--simu-eff-mod	specifies whether or not to standardize the causal probe
--simu-hsq	specifies the proportion of variance in phenotype explained by the causal probes
--simu-k	specifies the disease prevalence. The default value is 0.1 if this option is not specified
--simu-qt	simulates a quantitative trait
--std	removes the probes with the standard deviation smaller than a specified threshold
--to-probe	specifies the end probe
--to-probe-kb	specifies the end physical position of the probes
--update-opi	reads a fully annotated .opi file
--upper-beta	removes the DNA methylation probes with the mean beta value larger than a specified threshold