FAST README

Table of Contents

Overview

This documents describes in details the various input/output options and input/output files for FAST. FAST is an application for efficiently running several gene based analysis methods simultaneously and efficiently on the same data set. The following methods are implemented using both linear regression (for quantitative traits) and logistic regression (for dichotomous traits):

1. GWiS (Huang et. al. PLoS Genet, 7(7):e1002177, Jul 2011)
2. Bimbam (Servin et. al. PLoS Genet, 3(7):e114, Jul 2007)
3. Vegas (Lui et. al. Am J Hum Genet, 87(1):139–145, Jul 2010)
4. MinSNP
5. MinSNP-Gene
6. Gates (Am J Hum Genet, 88(3):283–293, Mar 2011)
7. Single SNP Regression.
8. The above methods using summary data.

For ease of analysis of genome-wide data, a single chromosome can be run at a time so that all chromosomes can be run in parallel in a computer cluster.

Availability

FAST is downloadable from https://bitbucket.org/baderlab/fast/downloads/ . Additional files needed for running the software in "summary" mode (see options below) are also available at the same location.

Installation

GNU Scientific Library (GSL) is required to compile and use FAST. GSL can be downloaded from http://www.gnu.org/software/gsl/. After you download GSL, please refer to the INSTALL file within GSL directory for installation of GSL. Once GSL is installed, download FAST and decompress the archive. Go to the software directory and type "make clean" and then type "make". This should compile the code to produce a binary for the appropriate platform. The final executable is named FAST. Type "./FAST --help" to see all the options.

For more information on how to install GSL, see https://bitbucket.org/baderlab/fast/wiki/Installation

Organization of the application folders

```
|-- Code/
|   |-- source code files.
|
|-- Documentation/
|   |-- Readme.1.5.mc.pdf
|   |-- Readme.txt
|   |-- Readme.examples.txt
|
|-- Example/
|   |-- DATA.geno/
|   |   |-- Sample input files for example of mode = genotype
|   |-- DATA.summary/
|   |   |-- Sample input files for example of mode = summary
|   |-- OUT/
|   |   |-- output files are stored here.
|   |
|   |-- script run.Linear.geno.sh to run example for mode = genotype
|   |   for linear models with input files in FAST format.
|   |
|   |-- script run.Linear.impute2.geno.sh to run example for mode = genotype
|   |   for linear models with input files in IMPUTE2 format.
|   |
|   |-- script run.Logistic.geno.sh to run example for mode = genotype
|   |   for logistic models.
|   |
|   |-- script run.Summary.1.sh to run example for mode = summary
|   |   using pre-computed LD
|   |
|   |-- script run.Summary.2.sh to run example for mode = summary
|       while computing LD on the fly using haplotype and pre-computed
|
|-- FAST (main application)
|-- FAST.utils.sh (helper utility)
|
|-- GSL/ (needed to compile FAST)
|   |-- gsl-1.15.tar.gz (if user needs to install GSL which is a
|       pre-requisite for FAST)
|
|-- License.txt
|-- Makefile
|-- Readme.txt
|-- Readme.examples.txt
|
|-- Utils/
|   |-- GETREPORT/
|   |   |-- script to generate integrated report
|   |-- QQ/
|       |-- script to generate QQ plots for each method
|   |-- PLINK2FAST/
|       |--files to convert from plink tped format to FAST input format
|
|-- bin/
    |-- linux/
    |   |-- FAST (binary in linux)
    |-- macOS/
        |-- FAST (binary in macOS)
```

FAST options

Option	Value	Comments
--help	-	To see all the options described in this table.
--mode	'genotype' or 'summary'	MANDATORY option specifying the type of input data.
		mode = genotype
--impute2-file	File name	MANDATORY option for mode = genotype. It specifies the file containing the imputed genotype probabilities in IMPUTE2 format.
--impute2-info-file	File name	MANDATORY option for mode = genotype. It specifies the file containing the SNP imputation information in IMPUTE2 format.
--tped-file	File name	MANDATORY option for mode = genotype. It specifies the file containing the genotype dosages for a single chromosome.
--snpinfo-file	File name	MANDATORY option for mode = genotype. It specifies the file containing the SNP chromosome and positional base-pair information. Must be for a single chromosome sorted in increasing order of base-pair positions.
--mlinfo-file	File name	MANDATORY option for mode = genotype. It specifies the file containing the SNP allele and imputation quality information. Must be for a single chromosome in same order as the snpinfo-file.
--indiv-file	File name	MANDATORY option for mode=genotype. This file contains the individual IDs of the samples in the same order as the individuals in the tped-file.
--trait-file	File name	MANDATORY file for mode=genotype containing the phenotype and optionally, the covariates.
--scale-pheno	-	For mode=genotype, option to scale the phenotype to have unit variance
--quantile-pheno	-	For mode=genotype, option to use normal quantile transformation to transform the phenotype to normal distribution
--num-covariates	value	count of covariates (default = 0) for mode=genotype
		mode = summary
--summary-file	File name	MANDATORY option for mode = summary. It specifies the file containing the SNP regression coefficients and standard errors. Must be for a single chromosome sorted in increasing order of base-pair positions.
--multipos-file	File name	Option for mode = genotype. It specifies the file containing the SNPs mapped to multiple positions in the genome.
--ld-file	File name	Option for mode = genotype. It specifies the file containing the SNP-SNP LD per chromosome. Must be for a single chromosome. If this option is input, the --allele-file option must be also be input. Input when --compute-ld = 0 is specified.
--allele-file	File name	Option for mode = genotype. It specifies the file containing the SNP reference allele used in LD computation. Must be for a single chromosome in same order as the summary-file. If this

		option is input, the --ld-file option must be also be input. Input when --compute-ld = 0 is specified.
--hap-file	File name	MANDATORY option for mode = summary when --compute-ld = 1. When computing LD on the fly, it specifies the input haplotype file. Must be for a single chromosome. If this option is input, the --pos-file option must be also be input.
--pos-file	File name	MANDATORY option for mode = summary when --compute-ld = 1. When computing LD on the fly, it specifies the haplotype start byte positions for each haplotype in the file specified with the --hap-file option. Must be for a single chromosome. If this option is input, the --hap-file option must be also be input.
--pheno-var	value	For mode=summary option specifying phenotype variance
--compute-ld	0/1	For mode=summary, option to compute LD on the fly when 1 (default) or use pre-computed LD when 0. If 1 is specified, --hap-file and --pos-file must also be present when mode=summary.
--n-sample	value	For mode=summary option specifying number of samples/individuals to be used in the analysis.
Input options common for both modes		
--chr	chromosome number	MANDATORY chromosome number for either mode. It should be same as the chromosome number specified in the above input files for either mode. 1 - 22 for autosomes and 23 for chromosome-X.
--out-file	Filename	Output files are prefixed with filename, default = FAST.result
--gene-set	filename	File containing gene names and boundaries for either mode. If not specified, FAST will perform single SNP analysis.
--maf-cutoff	value	value to filter snps with maf < maf-cutoff for gene-based analysis (default = 0.01)
--random-seed	value	Seed for permutations (default = 2)
--flank	value	gene flanking region in base-pairs (default = 20000 bp)
--max-perm	value	maximum no. of permutations (default = 1000000)
--n-perm-min	value	minimum no. of permutations (default = 100)
--max-missingness	value	Fraction indicating max allowed missingness per snp (default = 0.05)
--missing-val	value	For mode = genotype, value indicating missing genotypes (default = -1)
--eff-sample-size	value	min effective sample size per snp (default = 5)
--imputation-quality	value	min imputation quality per snp between 0 and 1.0 (default = 0.3)
--omit-strand-ambiguous	-	drop strand ambiguous snps (default = no)
--skip-perm	-	If specified no permutations will be performed.
--verbose	-	Detailed output for debugging (default = no)
--n-threads	value	No. of threads / multiple cores in permutations (default = 1)
--sigma-a	value	priors for additive effects in Bayes Factor computations (default

		= 0.2)
Linear Regression Based Analysis Options		
--linear-snp	-	single snp linear regression for all snps
--linear-snp-gene	-	single snp linear regression for only the snps in the genes
--linear-minsnp	-	linear regression based minsnp
--linear-minsnp-gene-perm	-	linear regression based minsnp-gene-perm
--linear-gwis	-	linear regression based GWiS (i.e. GWiS-Linear)
--linear-bf	-	linear regression based Bayes Factors
--linear-vegas	-	linear regression based Vegas
--linear-gates	-	linear regression based Gates
--linear-minsnp-perm	-	linear regression based minsnp with permutation pvalues
--linear-gwis-perm	-	linear regression based GWiS with permutation pvalues
--linear-bf-perm	-	linear regression based Bayes Factors with permutation pvalues
--linear-vegas-perm	-	linear regression based Vegas with permutation pvalues
Logistic Regression Based Analysis Options		
--logistic-snp	-	single snp logistic regression for all snps
--logistic-snp-gene	-	single snp logistic regression for only the snps in the genes
--logistic-minsnp	-	logistic regression based minsnp
--logistic-minsnp-gene-perm	-	logistic regression based minsnp--perm
--logistic-gwis	-	logistic regression based GWiS (i.e. GWiS-Logistic)
--logistic-bf	-	logistic regression based Bayes Factors
--logistic-vegas	-	logistic regression based Vegas
--logistic-gates	-	logistic regression based Gates
--logistic-minsnp-perm	-	logistic regression based minsnp with permutation pvalues
--logistic-gwis-perm	-	logistic regression based GWiS with permutation pvalues
--logistic-bf-perm	-	logistic regression based Bayes Factors with permutation pvalues
--logistic-vegas-perm	-	logistic regression based Vegas with permutation pvalues

File format for Input Genotype Data (mode=genotype)

Tip: To ensure proper running, please check that all input files are tab–delimited.

Input genotype data can be provided in either of the two ways: (1) IMPUTE2 format similar to the output of the imputation software IMPUTE2 (2) Format specific to FAST.

IMPUTE2 format

genotype File (option --impute2-file): For 2 individuals at 5 SNPs whose genotypes are

SNP 1 : AA AA

SNP 2 : GG GT

SNP 3 : CC CT

SNP 4 : CT CT

SNP 5 : AG GG

The correct genotype file would be

SNP1	rs1	1000	A	C	1	0	0	1	0	0
SNP2	rs2	2000	G	T	1	0	0	0	1	0
SNP3	rs3	3000	C	T	1	0	0	0	1	0
SNP4	rs4	4000	C	T	0	1	0	0	1	0
SNP5	rs5	5000	A	G	0	1	0	0	0	1

So, at SNP3 the two alleles are C and T so the set of 3 probabilities for each individual correspond to the genotypes CC, CT and TT respectively.

imputation information File (option --impute2-info-file): Name of SNP-wise information file with one line per SNP and a single ***mandatory header*** line at the beginning. This file always contains the following columns:

1. SNP identifier
2. rsID
3. base pair position
4. expected frequency of allele coded '1' in genotype file
5. measure of the observed statistical information associated with the allele frequency estimate
6. average certainty of best-guess genotypes
7. internal "type" assigned to SNP (not used by FAST, set to 0)
8. info_typeX (not used by FAST, set to 0)
9. concord_typeX (not used by FAST, set to 0)
10. r2_typeX (not used by FAST, set to 0)

FAST format

tped File (option --tped-file): This file must be ***tab delimited*** and it contains the genotype dosage data. Each row represents a SNP, and each column an individual. Each genotype is a real value between 0 and 2 as output from a genotype imputation algorithm like Impute or Mach. Further, genotype data represented with two alleles - 'a/a', 'A/a', 'A/A' can be converted to dosage of any allele i.e, the count of allele 'a' so that 'a/a' becomes 0, 'A/a' becomes 1 and 'A/A' becomes 2.

Note **1.** ***Missing genotype values are indicated with a negative value (default = -1), see option "--missing-val".***

2. ***No header line is allowed in this file. The file corresponds to a single chromosome.***

For example, here are five individuals typed for 2 SNPs (one row = one snp, one column = one individual):

0.2	0.5	1.2	1.9	1.1
1.4	0.0	0.0	2.0	2.0
		…		

---

mlinfo File (option --mlinfo-file) : This file must be ***tab delimited*** and it should have exactly 6 columns –
(1) rs# or SNP identifier,
(2) allele1,
(3) allele2,
(4) frequency of allele1,
(5) minor allele frequency(MAF), and
(6) imputation quality for each SNP.

For imputed data, 'Qual' can represent the 'Rsq' metric ouput by Mach or 'Info' metric output by Impute algorithms. For genotype data that are converted to dosage data for analysis, the 'Qual' column can be all 1.0 representing perfect quality.

The format of this file is a ***mandatory header line*** followed by one row for each SNP.

Note ***The header line must start with a '#'. The file corresponds to a single chromosome.***

#SNP	*Allele1*	*Allele2*	*Freq*	*Maf*	*Qual*
rs1	A	G	0.3	0.3	0.9
rs2	T	C	0.8	0.2	0.5
			….		

---

snp info File (option --snpinfo-file): This file must be ***tab delimited.*** Each line of the file describes a single marker and must contain exactly 4 columns –
(1) rs# or SNP identifier,
(2) chromosome,
(3) genetic distance (morgans), and
(4) base-pair position (bp units).

The format of this file is a ***mandatory header line*** followed by one row for each SNP.

Note ***The header line must start with a '#'. The file corresponds to a single chromosome.***

#SNP	*Chr*	*GD*	*BP*
rs1	1	0	10000
rs2	1	0	10004
	….		

Other input files

Individual ID File (option --indiv-file): This file contains the unique individual ID's corresponding to each column of the tped file. The file contains a single column where each row contains a single ID. The count of individual IDs in this file must match the count of columns of the tped file. The order of individuals in this file must match the order in the tped file.

Note ***No header line is allowed in this file.***

indiv_1
indiv_2
indiv_3
indiv_4
indiv_5
….

Phenotype + Covariate File (option --trait-file) : This file must be ***tab delimited*** . This file describes the phenotypes and covariates for each individual.

The first six columns are mandatory:-
(1) Family ID
(2) Individual ID
(3) Paternal ID
(4) Maternal ID
(5) Sex (1=male; 2=female; other=unknown)
(6) Phenotype.

The Phenotype column can be ***optionally*** followed by more than one covariate column (when --num-covariates > 0).

Note **1.** ***The first line must be a header line starting with a '#'.***

2. ***Only a single phenotype column is permitted.***

3. ***All covariates specified will be used for analysis.***

4. ***Missing phenotype/covariate values must be specified with NA.***

Example: (Note, the columns Cov1, Cov2 are optional)

#Fam ID	*Ind ID*	*Dad ID*	*Mom ID*	*Sex*	*Phenotype*	*Cov1*	*Cov2*
fam_id1	ind_1	ind_3	ind_5	1	0.3833	10.344	10
fam_id2	ind_2	ind_4	ind_6	2	-0.2231	21.322	20
				….			

The phenotype can be either a quantitative trait or a binary affection status column: FAST will automatically detect which type. Quantitative traits with decimal points must be coded with a period/full-stop character and not a comma, i.e. 5.123 not 5,123. For dichotomous trait, any two integer values (e.g. 0/1 or 1/2) must be used.

If Sex needs to be specified as a covariate, it must also be specified in one of the covariate columns, e.g.

Fam ID	*Ind ID*	*Dad ID*	*Mom ID*	*Sex*	*Phenotype*	*Cov1*	*Sex*
fam_id1	ind_1	ind_3	ind_5	1	0.3833	10.344	1
fam_id2	ind_2	ind_4	ind_6	2	-0.2231	21.322	2
				….			

File format for Input Summary Data (mode=summary)

Tip: To ensure proper running, please check that all input files are tab–delimited.

summary data file (option --summary-file) : This file contains the meta-analysis information for each SNP. ***The file must be tab delimited. The first line is a mandatory header line and must start with a '#'.*** Each subsequent row provides the information for each SNP and must have the following 10 columns :–

(1) Chromosome
(2) rs# or SNP identifier,
(3) Allele 1 (allele coded as 0)
(4) Allele 2 (allele coded as 1)
(5) allele frequency of Allele 2
(6) number of samples without missing data
(7) SNP base pair position
(8) Single SNP regression coefficient (beta)
(9) Single SNP regression standard error (se)
(10) Single SNP regression pvalue

Example:

#chr	*snp*	*Allele1*	*Allele2*	*Af*	*Nsample*	*bp*	*beta*	*se*	*pvalue*
10	rs1	A	T	0.3	2000	123456	0.34	0.12	0.108
10	rs2	G	C	0.2	1998	123478	1.4	0.2	0.045
				….					

Note **1.** ***The first line must be a header line starting with a '#'.***
2. ***The file corresponds to a single chromosome.***
3. The SNP alleles can be coded as A/G/T/C or 1/2/3/4.

A simpler summary data file (option --summary-file) : This file provides simpler format that contains the meta-analysis information for each SNP. ***The file must be tab delimited. The first line is a mandatory header line and must start with a '#'.*** Each subsequent row provides the information for each SNP and must have the following 4 columns :–

(1) Chromosome
(2) rs# or SNP identifier,
(3) SNP base pair position
(4) Single SNP regression pvalue

Example:

#chr	*snp*	*bp*	*pvalue*
10	rs1	123456	0.108
10	rs2	123478	0.045

Note **1.** ***The first line must be a header line starting with a '#'.***
2. ***The file corresponds to a single chromosome.***
3. The SNP alleles can be coded as A/G/T/C or 1/2/3/4.
4. The sign of the regression coefficients are unavailable in simpler file format, which slightly decreases the accuracy of GWiS, so use the detailed summary file whenever possible.

LD file (option --ld-file) : This file specifies the pair-wise LD information between SNPs. ***The file must be tab delimited. The first line is a mandatory header line and must start with a '#'.*** Each line contains mandatory 7 columns:-

(1) Chromosome of SNP 1
(2) base pair position of SNP 1
(3) rs# or SNP identifier for SNP 1

(4) Chromosome of SNP 2
(5) base pair position of SNP 2
(6) rs# or SNP identifier for SNP 2
(7) Correlation between SNP 1 and SNP 2 (a value between -1 and +1).

#CHR1	*BP1*	*SNP1*	*CHR2*	*BP2*	*SNP2*	*LD*
1	12345	rs1	1	12346	rs2	0.342
1	12345	rs1	1	12347	rs3	-0.59
				….		

Note **1.** ***The file corresponds to a single chromosome.***

2. ***The file must be sorted first in ascending order of the base pair position of SNP 1 and then in ascending order of the base pair position of SNP 2.***

---

allele info file (option --allele-file)**:** This file specifies the reference and alternate alleles used in computing the LD in the LD file. The file must be tab delimited. ***The first line is a mandatory header line and must start with a '#'.*** Each line contains mandatory 3 columns:-

(1) rs # or SNP identifier
(2) Allele 1
(3) Allele 2

Example:

snp	*Allele1*	*Allele2*
rs1	A	T
rs2	G	C

Note ***The file corresponds to a single chromosome.***

Haplotye file (option --hap-file)**:** This file specifies the reference haplotypes for computing LD on the fly. The file must be tab delimited**.** Each line contains a single SNP with the columns:-

(1) chromosome
(2) rs # or SNP identifier
(3) base pair position
(4) Allele1
(5) Allele2
(6) String of 0 and 1, where 0 represents Allele1 and 1 represents Allele2

Example:

22	rs1	12345	A	T	1	0	1	0	1 …
22	rs2	18345	T	G	0	0	1	0	0 …
		…							

Note, no header line present. Each column (column 6 onwards) is a haplotype.

Note ***The file corresponds to a single chromosome.***

Haplotye index file (option --pos-file)**:** This file specifies the reference haplotypes start byte positions for computing LD on the fly. The file must be tab delimited**.** Each line contains a single SNP with the columns:-

(1) chromosome
(2) rs # or SNP identifier
(3) start of haplotype position in bytes in the haplotype file.
(4) Allele1
(5) Allele2

Example:

```
22      rs1     0       A       T
22      rs2     4096    T       G
                ...
```

Note, no header line present.

Note ***The file corresponds to a single chromosome.***

Multipos File: This file maintains a pre-computed list of problematic SNPs that are mapped to multiple loci and will be skipped during analysis. For build 37, it is available from https://bitbucket.org/baderlab/fast/wiki/RefHaps

---

Pre-computed haplotype files from 1000 Genomes reference panels

Pre-computed haplotype files and their corresponding index files (for input with options --hap-file, --pos-file) from 1000 Genomes released on May 2012 are available for download from
https://bitbucket.org/baderlab/fast/wiki/RefHaps
for use with mode=summary. The following reference populations are available : ASW, CEU
Please use the correct reference panel with same ethnicity as the study sample for gene-based analysis.

File format for input files specifying gene list

Tip: To ensure proper running, please check that all input files are tab–delimited.

Gene-set File: This file specifies the gene boundary information for each gene to be used in the analysis. ***The first line is a mandatory header line and must start with a '#'.*** Each line contains 5 mandatory columns:-

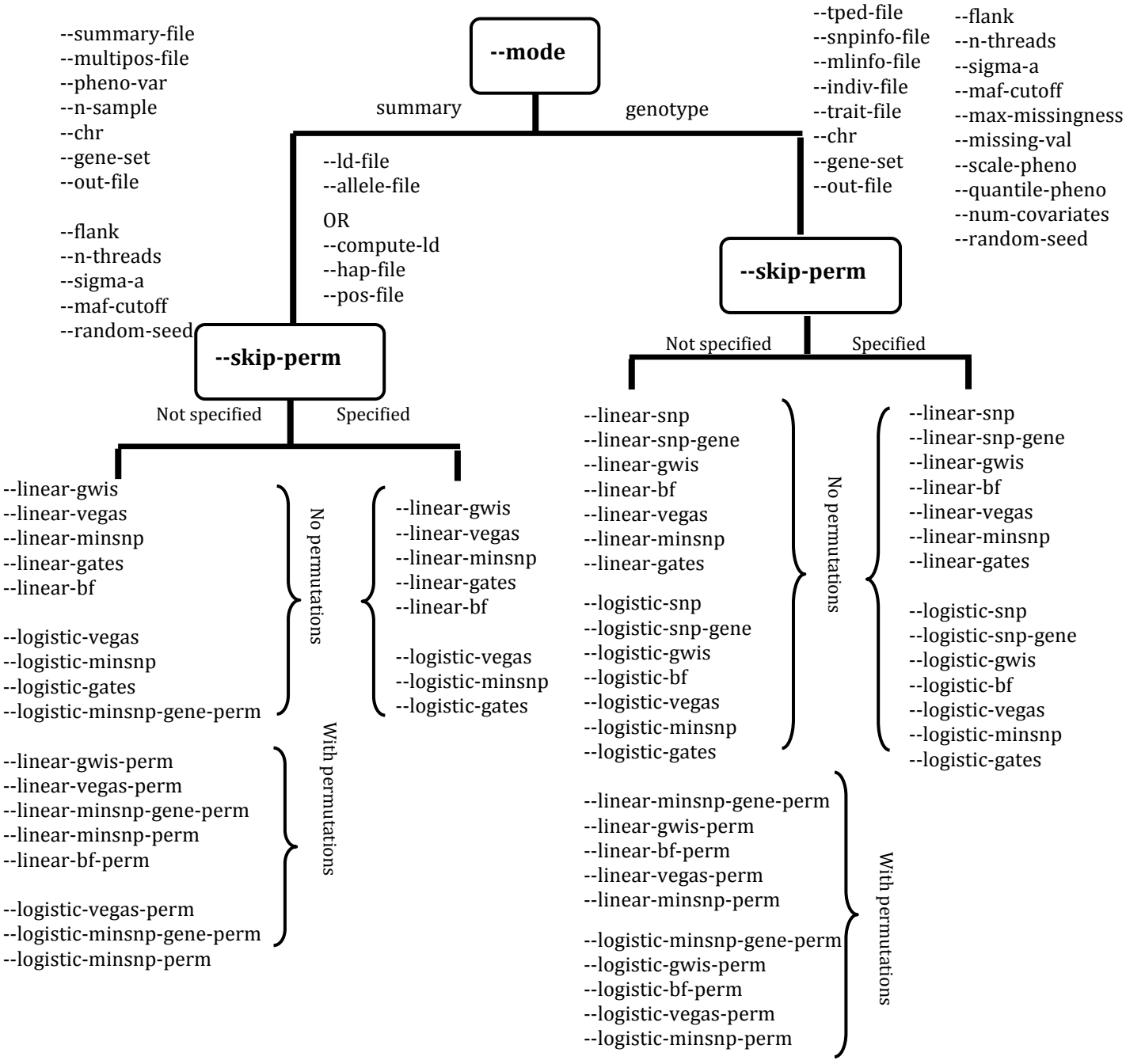
(1) Gene id
(2) Gene name
(3) Chromosome
(4) Gene start position in base-pairs
(5) Gene end position in base-pairs

e.g.

#Gene ID	*Gene Name*	*Chr*	*Start*	*End*
GeneID:347688	TUBB8	10	82997	85178
GeneID:439945	LOC439945	10	116561	122386
		....		

FAST Options Flow Chart

The following diagram shows the interplay between the various options of FAST. The input options for each mode are also shown, the mandatory options are in bold. The –skip-perm option is useful when all methods are specified with –perm suffix and the user chooses to avoid all permutations for quickly exploring the genes.



FAST analysis options and output file format for various methods

Note: FAST will append the chr number to the output file name prefix provided as input with the '--out-file' option; e.g. --out-file outfile with --chr 10 options will result in output file prefix **outfile.chr10**. So for GWiS method with linear regression, FAST will generate output file named **outfile.chr10.GWiS.Linear.txt.** Similarly for all other methods.

Method	Input Option & conditions	Output File Name & Format
GWiS-Linear Regression GWiS-Linear Regression with permutations	mode=genotype/summary --linear-gwis --linear-gwis-perm	Filename: Out.chrXX.BIC.Linear.txt <u>Multiple lines are present for each multi-SNP model in a gene :-</u> 1. The first line of each model has SNP.name=NONE indicating the null (intercept only model when no covariates, or covariates only model when covariates are present). 2. Next follows one or more lines for each SNP added to the model. 3. The line with SNP.name=SUMMARY indicates end of the model for the gene. This line also prints the values of K, SSM, BIC, f.stat n.stop, n.better and pval for the final K-snp model in the gene. <u>Format</u> Chr : Chromosome GeneID : Unique gene id Name : Gene name Start : Gene start in bp End : Gene end in bp Length : Gene length in bp SNPs : No. of snps in the gene Tests : Effective no. of snps in the gene SNP.name : SNP entering the model SNP.pos : SNP position in bp SNP.MAF : SNP minor allele frequency SNP.qual : SNP imputation quality K : Current model size SSM : Sum of the squares of the model BIC : BIC increment for the snp F.stat : Current model F-statistic R2 : Multiple R2 of the snp with the others in the model. n.stop : No of permutations executed n.better : No of permutations with better BIC score pval : Gene pvalue
GWiS-Logistic Regression GWiS-Logistic with permutations	mode=genotype --logistic-gwis --logistic-gwis-perm	Filename: Out.chrXX.BIC.Logistic.txt <u>Multiple lines are present for each multi-SNP model in a gene :-</u> 1. The first line of each model has SNP.name=NONE indicating the null (intercept only model when no covariates, or covariates only model when covariates are present). 2. Next follows one or more lines for each SNP added to the model. 3. The line with SNP.name=SUMMARY indicates end of the model for the gene. This line also prints the values of K, SSM, BIC, chi2, n.stop, n.better and pval for the final K-snp model in the gene. <u>Format</u> Chr : Chromosome GeneID : Unique gene id Name : Gene name Start : Gene start in bp End : Gene end in bp

		Length : Gene length in bp SNPs : No. of snps in the gene Tests : Effective no. of snps in the gene SNP.name : SNP entering the model SNP.pos : SNP position in bp SNP.MAF : SNP minor allele frequency SNP.qual : SNP imputation quality K : Current model size BIC : BIC increment for the snp chi2 : Current model chi squared n.stop : No of permutations executed n.better : No of permutations with better BIC score pval : Gene pvalue
minSNP –Linear Regression minSNP Gene Perm –Linear Regression	mode=genotype/summary (no permutations) --linear-minsnp (with permutations) --linear-minsnp-perm --linear-minsnp-gene-perm	Filename : Out.chrXX.minSNP.Linear.txt (for minSNP) : Out.chrXX.minSNP_Gene.Linear.txt (for minSNP_Gene) minsnp-perm assigns the gene the permutation pvalue of the best SNP in the gene s.t. the permutations are performed only for the best SNP. minsnp-gene-perm assigns the gene the permutation pvalue of the best SNP in the gene s.t. each permutation step involves the best SNP in the entire gene for the permuted trait. One line for each snp mapped to a gene. <u>Format</u> Chr : Chromosome GeneID : Unique gene id Name : Gene name Start : Gene start in bp (includes flank) End : Gene end in bp (includes flank) Length : Gene length in bp (includes flank) SNPs : No. of snps in the gene Tests : Effective no. of snps in the gene SNP.name : SNP entering the model SNP.pos : SNP position in bp SNP.MAF : SNP minor allele frequency SNP.qual : SNP imputation quality chi2 : SNP chi squared statistic n.tot : No of permutations executed n.better : No of permutations with better chi2 pval : p-value isBest : 0/1 indicating if this SNP has best chi2 in the gene. n.tot and n.better are output only when permutations are performed for minSNP or for minSNP Gene Perm. pval contains SNP parametric pvalue for minSNP when permutations are disabled.
minSNP Logistic	mode=genotype/summary (no permutations) --logistic-minsnp	Filename : Out. chrXX.minSNP.Logistic.txt (for minSNP) : Out. chrXX.minSNP_Gene.Logistic.txt (for minSNP_Gene) minsnp-perm assigns the gene the permutation pvalue of the best SNP in the gene s.t. the permutations are performed only for the best SNP. minsnp-gene-perm assigns the gene the permutation pvalue of the

minSNP_Gene Perm – Logistic Regression	(with permutations) --logistic-minsnp-perm --logistic-minsnp-gene-perm	best SNP in the gene s.t. each permutation step involves the best SNP in the entire gene for the permuted trait. One line for each snp mapped to a gene. <u>Format</u> Chr : Chromosome GeneID : Unique gene id Name : Gene name Start : Gene start in bp End : Gene end in bp Length : Gene length in bp SNPs : No. of snps in the gene Tests : Effective no. of snps in the gene SNP.name : SNP entering the model SNP.pos : SNP position in bp SNP.MAF : SNP minor allele frequency SNP.qual : SNP imputation quality chi2 : SNP chi squared statistic n.tot : No of permutations executed n.better : No of permutations with better chi2 pval : p-value isBest : 0/1 indicating if this SNP ha best chi2 in the gene. n.tot and n.better are output only when permutations are performed for minSNP or for minSNP Gene Perm. pval contains SNP parametric pvalue for minSNP when permutations are disabled.
Bimbam –Linear Regression	mode=genotype (no permutations) --linear-bf (with permutations) --linear-bf-perm	Filename : Out.chrXX.BF.Linear.txt One line for each gene in the chromosome. <u>Format</u> Chr : Chromosome GeneID : Unique gene id Name : Gene name Start : Gene start in bp End : Gene end in bp Length : Gene length in bp SNPs : No. of snps in the gene Tests : Effective no. of snps in the gene BF_sum : Linear regression based Bayes Factor sum for the gene. n.tot : No of permutations executed n.better : No of permutations with better BF_sum pval : pvalue
Bimbam – Logistic	mode=genotype (no permutations) --logistic-bf (with permutations)	Filename : Out.chrXX.BF.Logistic.txt One line for each gene in the chromosome. <u>Format</u> Chr : Chromosome GeneID : Unique gene id Name : Gene name Start : Gene start in bp End : Gene end in bp

Regression	--logistic-bf-perm	Length : Gene length in bp SNPs : No. of snps in the gene Tests : Effective no. of snps in the gene BF_sum : Logistic regression based Bayes Factor sum for the gene. n.tot : No of permutations executed n.better : No of permutations with better BF_sum pval : p-value
Vegas –Linear Regression	mode=genotype/summary (no permutations) --linear-vegas (with permutations / simulations) --linear-vegas-perm	Filename : Out.chrXX.Vegas.Linear.txt One line for each gene in the chromosome. <u>Format</u> Chr : Chromosome GeneID : Unique gene id Name : Gene name Start : Gene start in bp End : Gene end in bp Length : Gene length in bp SNPs : No. of snps in the gene Tests : Effective no. of snps in the gene Vegas : Linear regression based Vegas score for the gene. n.tot : No of permutations executed n.better : No of permutations with better Vegas score. pval : p-value
Vegas –Logistic Regression	mode=genotype/summary (no permutations) --logistic-vegas (with permutations) --logistic-vegas-perm	Filename : Out.chrXX.Vegas.Logistic.txt One line for each gene in the chromosome. <u>Format</u> Chr : Chromosome GeneID : Unique gene id Name : Gene name Start : Gene start in bp End : Gene end in bp Length : Gene length in bp SNPs : No. of snps in the gene Tests : Effective no. of snps in the gene Vegas : Logistic regression based Vegas gene score. n.tot : No of permutations executed n.better : No of permutations with better Vegas score. pval : p-value
Gates Linear/Logistic Regression	mode=genotype/summary --linear-gates --logistic-gates	Filename : Out.chrXX.Gates.Linear.txt/ Out.chrXX.Gates.Logistic.txt One line for each gene in the chromosome. <u>Format</u> Chr : Chromosome GeneID : Unique gene id Name : Gene name Start : Gene start in bp End : Gene end in bp Length : Gene length in bp SNPs : No. of snps in the gene Tests : Effective no. of snps in the gene Gates : Gates score for the gene

		pval : pvalue

Tip When a method is specified with the –perm suffix, permutations are performed when mode=genotype, simulations are performed when mode=summary.

Tip If you only have a few genes on a chromosome, use the option ----linear-snp-gene or --logistic-snp-gene. This will limit the single SNP computations to only these genes.

Additional Output files

Out.chrXX.allSNP.Linear.txt : This file lists the single SNP linear regression results for each SNP.

Format

SNP.id : SNP name
pos : SNP position in base pairs
A1 : Allele 1
A2 : Allele 2
Beta : Regression coefficient
Se : Regression standard error
Chi2 : Chi square
logBF : Log Bayes Factor
MAF : Minor allele frequency
Qual : SNP imputation quality
eSampleSize : Effective sample size for the SNP computed as (#samples) x Qual x 2 x MAF x (1-MAF)
nGenes : No of genes to which this SNP belongs
Fmiss : Fraction of samples with missing values
pvalue : SNP parametric pvalue

Out.chrXX.allSNP.Logistic.txt : This file lists the single SNP logistic regression results for each SNP.

Format

SNP.id : SNP name
Chr : Chromosome
pos : SNP position in base pairs
A1 : Allele 1
A2 : Allele 2
Beta : Regression coefficient
Se : Regression standard error
Wald : Wald-statistic
logBF : Log Bayes Factor
MAF : Minor allele frequency
Qual : SNP imputation quality
eSampleSize : Effective sample size for the SNP computed as (#samples) x Qual x 2 x MAF x (1-MAF)
nGenes : No of genes to which this SNP belongs
Fmiss : Fraction of samples with missing values
pvalue : SNP parametric pvalue

Out.chrXX.geneSNP.txt : This file lists the mapping of each SNP and gene. A SNP can appear multiple times in this file if it belong to multiple overlapping genes.

Format

SNP.name : SNP name
SNP.chr : chromosome
SNP.bp : SNP position in base pairs
GeneID : Unique gene id
Gene.name : Gene name
Gene.start : Gene start in bp
Gene.end : Gene end in bp
MAF : SNP minor allele frequency
Qual : SNP quality

eSampleSize : SNP effective sample size

Examples

The 'Example' folder contains examples of running FAST with

1. genotype data using linear model (script to run is run.Linear.geno.sh)
2. genotype data using logistic model (script to run is run.Logistic.geno.sh)
3. summary data with pre-computed LD (script to run is run.Summary.1.sh)
4. summary data with computing LD on the fly (script to run is run.Summary.2.sh)

Each script is a shell script that specifies the various options described in this document for running FAST. The input files are stored in Example/DATA.geno and Example/DATA.summary. The output files are stored in OUT. Each script can be run from the command line by simply './script-name' (e.g. ./run.Summary.1.sh')

Preparing an integrated report file and QQ plots combining the output from several methods in FAST

Note: You will need to have perl and R installed in your system. Also you need to install the module 'Statistics::Distributions' available from http://search.cpan.org/~mikek/Statistics-Distributions-1.02/Distributions.pm.

The script FAST.utils.sh provides options to generate a report. **Must run ./install.sh first.**

The options for FAST.utils.sh are:

-p <prefix-to-file> <type>	QQPlot with 2 arguments, do QQ plots for each method
-r <prefix-to-file> <type> <pvalue-cutoff-for-gene> < pvalue-cutoff-for-SNP>	GetReport with 4 arguments, print combined report

type = Linear | Logistic | Summary

<pvalue-cutoff-for-gene>: p-value cut off for gene-based tests. If permutation is not performed, this will be the percentage of 'significant' genes, ranked by the test statistics.

<p-value cut off for SNP>: p-value cut off for SNP-based tests. If permutation is not performed, this will be the percentage of 'significant' genes, ranked by the test statistics.

To combine the results from different methods for analysis and interpretation, you can run the script ./FAST.utils.sh with appropriate options :

Run with the '-r' option as

./FAST.utils.sh -r <prefix-to-file> <type> <pvalue-cutoff-for-gene> < pvalue-cutoff-for-SNP>

e.g ./FAST.utils.sh -r ./Example/OUT/output Summary 0.001 0.0001

where <type> can be Linear or Logistic or Summary, and <prefix-to-file> is the path to the prefix of the output files generated by FAST, **EXCLUDING the chromosome number 'chrxxx'**. It will generate a integrated report file, with suffix report.txt, from the FAST output files for each method combining all the chromosomes that you have run with FAST.

OUTPUT: The output of getReport.pl is a list of genes that reach the significance threshold specified in the command line parameters. A gene will be reported if it has significant p value from at least one of the FAST methods. This list is followed by a list of significant SNPs in the non-transcribed regions. getReport.pl reports all available results from the FAST output files.

EXAMPLE:

The following examples takes the FAST output files using the linear model, and reports genes that have gene-based p-value < 0.01 and SNP-based p value < 0.001. Note that only the prefix of the FAST output files **EXCLUDING the chromosome number 'chrxxx'** is needed.

./FAST.utils.sh -r ./Example/OUT/output Linear 0.01 0.001

./FAST.utils.sh -r ./Example/OUT/output Logistic 0.01 0.001

./FAST.utils.sh -r ./Example/OUT/output Summary 0.01 0.001

These will generate output files : report.Linear.txt , report.Logistic.txt and report.Summary.txt