**fcGENE: A versatile tool for format conversions of SNP datasets**

Nab Raj Roshyara1,2, Markus Scholz1,2

1. Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstrasse 16-18, 04107 Leipzig, Germany
2. LIFE Center (Leipzig Interdisciplinary Research Cluster of Genetic Factors, Phenotypes and Environment), University of Leipzig, Philipp-Rosenthal Strasse 27, 04103 Leipzig, Germany

**Supplementary Material**

In this document, we present a number of additional commands supported by fcGENE, and explain their use in more detail. We also provide additional examples of fcGENE workflows.

**Commands summary**

Here, we provide an overview of implemented command options of fcGENE.

|  |  |
| --- | --- |
| **Type of file format** | **Command options to load data into fcGENE** |
| PLINK | --ped example.ped --map example.map |
| plink-formatted  covariate file | --covar examplecov.txt--covar-name pheno1,pheno2,covar\_A,covar\_B  --covar-type P,B,D,C  (These options are necessary while preparing files for SNPTEST) |
| PLINK-formatted dosage file | -- dosage example.dosage --fam example.fam --map example.map |
| PLINK-raw (recodeAD type ) | --recodeADexample.raw –map example.map --snpinfo example\_snpinfo.txt |
| PLINK-raw (recodeA type ) | --recodeA example.raw, --map example.map --snpinfo example\_snpinfo.txt |
| MaCH | --ped example.ped -- dat example.dat |
| mach references | --mach-hap mach\_ref.hap --mach-snp mach\_ref.snp  --force pheno=unaff,sex=m (optional ) |
| mach imputation output | --mach-geno example.geno --mach-info example.info |
| mach imputation output | --mach-mlgeno example.mlgeno --mach-mlinfo example.mlinfo |
| mach imputation output | --mach-mlprob example.mlprob --mach-mlinfo example.mlinfo |
| minimac imputation output | --minimac-mlprob example.mlprob --minimac-mlinfo example.mlinfo |
| IMPUTE | --gens example.gens |
| impute references | --impute-hap impute\_ref.hap --impute-legend impute\_ref.legend  --force pheno=unaff,sex=m (optional) |
| impute imputation output | --gens example.impute2 --info example.impute2\_info |
| BEAGLE | --bgl example.bgl |
| beagle references | --bgl beagle\_ref.bgl  --force pheno=unaff,sex=m (optional) |
| beagle imputation output | --bgl example.bgl.phased --bgl-rsq example.bgl.r2 (optional) |
| beagle imputation output | --bgl-gprobs example.bgl.gropbs --bgl-rsq example.bgl.r2 (optional) |
| BIMBAM | --wbg example.geno.txt |
| bimbam imputation output | --wbg example.best.guess.genotype.txt  --pos example.snpdata.txt |
| bimbam imputation output | --wgd example. genotype.probability.distribution.txt  --pos example.snpinfo.txt |

**Table 1:** Commands to read SNP data of different formats.

Table 1 summarizes command options necessary to upload genotype data of different formats into fcGENE. In the table, we used the name “example” as file name combined with different extensions specific for different data formats.

**Example:** To upload BEAGLE output, one can use the command

*./fcgene --bgl-gprobs example.bgl.gropbs*

After loading a dataset into fcGENE, it can be converted into the format of any of the programs mentioned in table 2. Data conversions are usually performed by combining an option of table 1 with an option of table 2.

|  |  |
| --- | --- |
| **Type of program-files to be created** | **Command option** |
| PLINK (\*.ped ,\*.dat) | --oformat plink |
| PLINK (dose file: \*.dat and \*.fam , \*.map), | --oformat plink-dosage |
| PLINK (raw(recodeA type) file: \*.raw and \*.map), | --oformat plink-recodeA |
| PLINK-raw expected dose file: \*.raw and \*.map), | --oformat recodeA-dose |
| PLINK (raw(recodeAD type) file: \*.raw and \*.map), | --oformat plink-recodeAD |
| MaCH (\*.ped , \*.dat) | --oformat mach |
| minimac (\*.snps, \*.ped) | --oformat minimac |
| IMPUTE (\*.gens, \*.strand) | --oformat impute |
| BEAGLE (\*.bgl) | --oformat beagle |
| BIMBAM (\*.geno.txt, \*.pos.txt) | --oformat bimbam |
| PHASE/fastPHASE | --oforamt phase /--oformat fastphase |
| SNPTEST (\*.gens , \*.sample ) | --oformat snptest |
| EIGENSOFT | --oformat eigensoft |
| HAPLOVIEW | --oformat haploview |
| R compatible files: \*.geno.txt, \*.affs.txt | --oformat R or --oformat r |
| Standard text file with expected doses of minor allele | --oformat r-dose or -- oformat R-dose |
| GenABEL | --oformat genable |
| Pedigree information (pedinfo.txt) | --write-pedinfo |
| SNP information (snpinfo.txt) | --write-snpinfo |
| A list of individuals (pedlist.txt) | --write-pedlist |
| A list of SNP/rsid (snplist.txt) | --write-snplist |
| writing a file with p-values of exact HWE test | --hardy |
| Calculation of SNP-wise and/or sample-wise callrate | --crate |
| Calculation of allele frequencies | --freq |

**Table 2: Commands to generate file formats required for different GWA analysis tools**

**Example:** If genotype SNP data are given in MaCH-imputed mlprob format, we can use the following command to prepare inputs of PLINK.

*./fcgene --mach-mlprob example.mlprob --mach-mlinfo example.mlinfo - -oformat plink*

In Table 3, we summarized different optional commands related to quality control and data manipulation. Users can adjust for example different threshold values for different quality measures such as imputation quality, minor allele frequency (MAF), p-value of Hardy-Weinberg equilibrium (HWE), call rate etc. and apply these filters to their data. The sequential order of options is unimportant throughout. The following command can be used to filter MERLIN (MaCH-input) formatted data regarding different quality issues:

*./fcgene --ped example.ped --dat example.map\*

*--filter-snp hwe=1e-2,maf=0.1,crate=0.95 --filter-indiv crate=0.95\*

*--oformat mach --out mach/example\_qcd\_data*

|  |  |
| --- | --- |
| **Optional commands** | **Function of the command option** |
| --fitler-snp maf=0.1, crate=0.95, hwe=1e-2 | This command excludes SNPs that have smaller values of minor allele frequency, call rate, or p-value of Hardy-Weinberg equilibrium than the prescribed by the cutoff values defined by the commands maf=, crate= and hwe= respectively. Filters can be combined arbitrarily. |
| --filter-indiv crate=0.95 | Filters individuals with call rate lower than given cut-off value. |
| --pedinfo | By this command, one can provide an extra file containing pedigree information of individuals. |
| --snpinfo | This option is used to provide an extra file containing SNP annotations (e.g. rsid, base pair position). |
| --rsq 0.3 | This command can be used with MaCH- and MINIMAC-imputed data excluding SNPs with imputation quality (Rsq value) lower than given cutoff. Rsq is an estimation of the squared correlation between imputed and true genotypes |
| --thresh | While finding the best guess genotype for each individual at every SNP, one can use the threshold option --thresh maxProb to guarantee that the best guess genotype has a probability equal or greater than maxProb. If the maximum probability of predicted genotypes is less than maxProb, then the genotype is set to missing. This option improves the confidence of best guess genotypes. |
| --rsq-thresh | This option is used to filter poorly imputed SNPs on the basis of squared correlation used by BEAGLE as a measure imputation quality. |
| --info-thresh | This option is used for post-imputation filtering of IMPUTE-imputed data on the basis of the info score |
| --force | This option forces fcGENE to assign different phenotype information and SNP information to specified values or strings. e.g. --force option can assign to dummy sex status i.e. all individuals either to male or female |

**Table 3: Optional commands used in fcGENE: These options comprise specification and application of quality cut-offs as well as provision of supplementary data.** **A functional description of each of the commands is given in the second column.**

**File formats used by different GWA-tools:**

Genotype file formats used by different GWA tools are described in the following. We also describe how to load and create these formats with fcGENE.

**PLINK -format files: example.ped and example.dat**

* example.ped:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fam\_1 | ind\_1 | 0 | 0 | 1 | 1 | T | T | T | T | G | G |
| fam\_2 | ind\_2 | 0 | 0 | 1 | 1 | G | G | C | T | G | A |
| fam\_3 | ind\_3 | 0 | 0 | 1 | 2 | T | G | T | T | A | G |
| fam\_4 | ind\_4 | 0 | 0 | 2 | 2 | G | G | C | C | 0 | 0 |

* example.map

|  |  |  |  |
| --- | --- | --- | --- |
| 22 | snp1 | 0.02 | 16212142 |
| 22 | snp2 | 0.13 | 20278224 |
| 22 | snp3 | 0.13 | 20304703 |

* fcGENE commands:
  + - To read files:

*./fcgene --ped example.ped --map example.map*

* + - To create this format:

*--oformat plink*

**PLINK -formatted dosage file: example.dosage, example.fam , example.map**

* example.dosage

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SNP | A1 | A2 | fam\_1 | ind\_1 | fam\_2 | ind\_2 | fam\_3 | ind\_3 | fam\_4 | ind\_4 |
| snp1 | T | G | 0 | 1 | 0.05 | 0 | 0.02 | 0.97 | 0 | 0.03 |
| snp2 | C | T | 0.88 | 0.03 | 1 | 0 | 1 | 0 | 1 | 0 |
| snp3 | G | A | 0.78 | 0.2 | 1 | 0 | 0.91 | 0 | 1 | 0 |

* example.fam

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fam\_1 | ind\_1 | 0 | 0 | 1 | 1 |
| Fam\_2 | ind\_2 | 0 | 0 | 1 | 1 |
| Fam\_3 | ind\_3 | 0 | 0 | 1 | 2 |
| Fam\_4 | ind\_4 | 0 | 0 | 2 | 2 |

* fcGENE commands:
  + To read files:

*./fcgene --dosage example.dose --fam example.fam\*

*–map example.map (optional)*

* + To create this format:

*--oformat plink-dosage*

**PLINK -formatted raw file (created with --recodeA in plink): example.raw, example\_snpinfo.txt**

* example.raw :

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| FID | IID | PAT | MAT | SEX | PHENOTYPE | snp1\_T | snp2\_C | snp3\_A |
| fam\_1 | ind\_1 | 0 | 0 | 1 | 1 | 2 | 0 | 0 |
| fam\_2 | ind\_2 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| fam\_3 | ind\_3 | 0 | 0 | 1 | 2 | 1 | 0 | 1 |
| fam\_4 | ind\_4 | 0 | 0 | 2 | 2 | 0 | 2 | NA |

* example.map: see plink-formatted map file above.
* example\_snpinfo.txt :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| nchr | snpid | rsid | bp | cm\_pos | allele1 | allele2 |
| 22 | snp1 | snp1 | 16212142 | 0.02 | T | G |
| 22 | snp2 | snp2 | 20278224 | 0.13 | T | C |
| 22 | snp3 | snp3 | 20304703 | 0.13 | G | A |

**Remark:** “example\_snpinfo.txt” file can be loaded by fcGENE with option “--snpinfo”. Since plink-formatted raw file has no allele information, it is necessary to provide at least the allele information. Therefore, “example\_snpinfo.txt “ should contain at least rsid, allele1 and allele2.

* fcGENE commands:
  + To read files:

*./fcgene --recodeA example.raw --map example.map\*

*--snpinfo example\_snpinfo.txt\*

*--pedinfo example\_pedinfo* (optional)

* + To create this format:

*--oformat plink-recodeA*

**PLINK-formatted raw dose file: example.raw**

This type of file format has the same form as previously mentioned PLINK raw files but provides expected allele doses of reference allele instead of genotypes resulting in a numbers between 0 and 2. By default, minor-allele is taken as reference allele. One can force fcGENE to change the reference allele with command option “--force ref-allele=”. Possible options for forcing reference allele are “--force ref-allele=minor-allele” (this is default), “--force ref-allele=major-allele”, “--force ref-allele=allele1” and “--force ref-allele=allele2”.

* example.raw :

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| FID | IID | PAT | MAT | SEX | PHENOTYPE | snp1\_T | snp2\_C | snp3\_A |
| fam\_1 | ind\_1 | 0 | 0 | 1 | 1 | 1.89 | 0.19 | 0.18 |
| fam\_2 | ind\_2 | 0 | 0 | 1 | 1 | 0.15 | 1.3 | 0.96 |
| fam\_3 | ind\_3 | 0 | 0 | 1 | 2 | 1.23 | 0.04 | 1.11 |
| fam\_4 | ind\_4 | 0 | 0 | 2 | 2 | 0.35 | 1.95 | NA |

* example.map: see plink-formatted map file above.
* fcGENE commands:
  + To create this format

*--oformat recodeA-dose*

**PLINK -formatted raw file (created with --recodeAD in plink ): example.raw**

* example.raw:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| FID | IID | PAT | MAT | SEX | PHENOTYPE | snp1\_T | snp1\_HET | snp2\_C | snp2\_HET |
| fam\_1 | ind\_1 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 0 |
| fam\_2 | ind\_2 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| fam\_3 | ind\_3 | 0 | 0 | 1 | 2 | 1 | 1 | 0 | 0 |
| fam\_4 | ind\_4 | 0 | 0 | 2 | 2 | 0 | 0 | NA | NA |

* example.map: see plink formatted map file above.
* example\_snpinfo.txt : Given previously.
* fcGENE command:

*./fcgene --recodeAD example.raw --map example.map --snpinfo example\_snpinfo.txt\*

*--pedinfo example\_pedinfo* (optional)

**PLINK -formatted covariate file: example.cov.txt**

* example.cov.txt

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| FID | IID | pheno1 | pheno2 | covar\_A | covar\_B |
| fam\_1 | ind\_1 | 0 | 0 | 0 | 0.61 |
| fam\_2 | ind\_2 | 0 | 0 | 0 | 0.44 |
| fam\_3 | ind\_3 | 0 | 0 | 0 | 0.39 |
| fam\_4 | ind\_4 | 1 | 0 | 0 | 0.29 |

* fcGENE command: fcGENE can read this file and convert it into SNPTEST format using the following options:

*--covar example\_plink\_covariate.txt\*

*--covar-name pheno1,pheno2,covar\_A,covar\_B\*

*--covar-type P,B,D,C*

Here, command option *--covar* reads plink-formatted covariate file. Similarly, using command option *--covar-name*, one can determine the name of covariates and command option *--covar-type* helps to determine the type of covariates. The letter *P* stands for binary phenotypes taking values of 0 and 1 respectively, *B* stands for Binary, *D* and *C* symbolize discrete and continues variables.

**Remarks**: More information on plink-formatted data can be found at

<http://pngu.mgh.harvard.edu/~purcell/plink/dataman.shtml>

**MaCH (Merlin) format files: example.ped and example.dat**

* example.ped

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fam\_1 | ind\_1 | 0 | 0 | m | T | T | T | T | G | G |
| fam\_2 | ind\_2 | 0 | 0 | m | G | G | C | T | G | A |
| fam\_3 | ind\_3 | 0 | 0 | m | T | G | T | T | A | G |
| fam\_4 | ind\_4 | 0 | 0 | f | G | G | C | C | 0 | 0 |

or

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| fam\_1 | ind\_1 | 0 | 0 | m | T/T | T/T | G/G |
| fam\_2 | ind\_2 | 0 | 0 | m | G/G | C/T | G/A |
| fam\_3 | ind\_3 | 0 | 0 | m | T/G | T/T | A/G |
| fam\_4 | ind\_4 | 0 | 0 | f | G/G | C/C | 0/0 |
|  |  |  |  |  |  |  |  |

* example.dat

|  |  |
| --- | --- |
| M | Snp1 |
| M | snp2 |
| M | snp3 |

* fcGENE commands:
  + To read files:

*./fcgene --ped example.ped --dat example.map*

* + To create files:

*--oformat mach*

**MaCH-formatted references: mach\_ref.snps, mach\_ref.hap**

* mach\_ref.snps

|  |
| --- |
| Snp1 |
| Snp2 |
| Snp3 |

* mach\_ref.hap

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| fam\_1->ind\_1 | HAPLO1 | T | T | G |
| fam\_1->ind\_1 | HAPLO2 | G | T | G |
| fam\_2->ind\_2 | HAPLO1 | G | T | A |
| fam\_2->ind\_2 | HAPLO2 | G | C | G |
| fam\_3->ind\_3 | HAPLO1 | T | T | G |
| fam\_3->ind\_3 | HAPLO2 | G | T | A |
| fam\_4->ind\_4 | HAPLO1 | T | C | A |
| fam\_4->ind\_4 | HAPLO2 | T | T | G |

* fcGENE commands:
  + To read files:

*./fcgene --mach-hap mach\_ref.hap --mach-snp mach\_ref.snp\*

*--force pheno=unaff,sex=m* (optional)

* + To create files: not applicable

**MaCH imputation outputs:**

* example.geno and example.info files:
  + example.geno

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| fam\_1->ind\_1 | GENO | T/T | T/T | G/G |
| fam\_2->ind\_2 | GENO | G/G | T/C | G/A |
| fam\_3->ind\_3 | GENO | T/G | T/T | G/A |
| fam\_4->ind\_4 | GENO | G/G | C/C | G/G |

* + example.info

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SNP | Al1 | Al2 | Freq1 | MAF | Quality | Rsq |
| snp1 | T | G | 0.375 | 0.375 | 1 | 1 |
| snp2 | T | C | 0.625 | 0.375 | 1 | 1 |
| snp3 | G | A | 0.6825 | 0.3175 | 0.89 | 0.6022 |

* + fcGENE commands:
    - To read files:

*./fcgene --geno example.geno -- info example.info*

* + - To create files: not applicable
* example.mlgeno and example.mlinfo files:
  + example.mlgeno

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| fam\_1->ind\_1 | ML\_GENO | G/G | C/C | G/G |
| fam\_2->ind\_2 | ML\_GENO | T/T | T/C | G/A |
| fam\_3->ind\_3 | ML\_GENO | T/G | C/C | G/A |
| fam\_4->ind\_4 | ML\_GENO | T/T | T/T | G/G |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SNP | Al1 | Al2 | Freq1 | MAF | Quality | Rsq |
| snp1 | T | G | 0.4901 | 0.4901 | 0.7088 | 0.6784 |
| snp2 | T | C | 0.4596 | 0.4596 | 0.8106 | 0.7682 |
| snp3 | G | A | 0.6686 | 0.3314 | 0.6021 | 0.0853 |
|  |  |  |  |  |  |  |

* + example.mlinfo
  + fcGENE commands:
    - To read files:

*./fcgene --mach-mlgeno example.geno --mach-mlinfo example.info*

*--rsq 0.3* (optional) *--maf-thresh 0.1* (optional)

* + - To create files with fcGENE: not applicable
* example.mlprob and example.mlinfo files:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| fam\_1->ind\_1 | ML\_PROB | 0.006 | 0.14 | 0.081 | 0.407 | 0.365 | 0.478 |
| fam\_2->ind\_2 | ML\_PROB | 0.301 | 0.516 | 0.393 | 0.535 | 0.395 | 0.5 |
| fam\_3->ind\_3 | ML\_PROB | 0.05 | 0.364 | 0.065 | 0.397 | 0.271 | 0.544 |
| fam\_4->ind\_4 | ML\_PROB | 0.383 | 0.472 | 0.886 | 0.11 | 0.519 | 0.403 |

* + example.mlprob
  + **exa**
  + example.mlinfo (see above)
  + fcGENE commands:
    - To read files:

*./fcgene –mach-mlprob example.geno -- mach-mlinfo example.info\*

*--rsq 0.3* (optional) *--maf-thresh 0.1* (optional)

*--pedinfo example\_pedinfo.txt --snpinfo example\_snpinfo.txt* (optional)

* + - To create files: not applicable

**MINIMAC format :**

* example.ped (see mach-formatted ped file )
* example.snps (see mach-formatted ref.snps file )
* fcGENE commands:
  + To read files : *./fcgene --ped example.ped -- snps example.snps*
  + To create files: *--oformat minimac*

**MINIMAC imputation outputs: see mach imputation outputs**

**IMPUTE (CHIAMO) format files:**

* example.gens

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| snp1 | Snp1 | 16212142.00 | T | G | 0.01 | 0.14 | 0.85 | 0.30 | 0.52 | 0.18 | 0.05 | 0.36 | 0.59 | 0.38 | 0.47 | 0.15 |
| snp2 | Snp2 | 20278224.00 | T | C | 0.08 | 0.41 | 0.51 | 0.39 | 0.54 | 0.07 | 0.07 | 0.40 | 0.54 | 0.89 | 0.11 | 0.00 |
| snp3 | Snp3 | 20304703.00 | G | A | 0.37 | 0.48 | 0.16 | 0.40 | 0.50 | 0.11 | 0.27 | 0.54 | 0.19 | 0.52 | 0.40 | 0.08 |

* example.strand

|  |  |
| --- | --- |
| rs1 | + |
| rs2 | + |
| rs3 | + |

* fcGENE commands:
  + To read files:

*--gens example.gens --strand example.strand\*

*--thresh maxProb --pedinfo example\_pedinfo.txt* (optional)

* + To create files: *--oformat impute*

**IMPUTE reference files:**

* Impute\_ref.hap

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |

This file contains 3 SNPs and four individuals. Each row represents a SNP. Each two columns represent an individual.

* Impute\_genetic\_map.txt

|  |  |  |
| --- | --- | --- |
| position | CEU\_rate(cM/Mb) | Genetic\_Map(cM) |
| 14431347 | 12.83039317 | 0 |
| 14432618 | 12.87908888 | 0.01630743 |
| 14433624 | 12.87908888 | 0.029263793 |
| 14433659 | 12.87568372 | 0.029714561 |
| 14433758 | 12.85861428 | 0.030989254 |
| 14434713 | 12.60322816 | 0.043269231 |
| 14435070 | 5.89938528 | 0.047768583 |
| 14439734 | 1.375554222 | 0.0752833160 |
|  |  |  |

* Impute\_ref.legend

|  |  |  |  |
| --- | --- | --- | --- |
| rsID | position | a0 | a1 |
| rs7288834 | 16212142 | G | T |
| rs16978746 | 20278224 | T | C |
| rs5754387 | 20304703 | G | A |

* fcGENE commands:
  + To read impute references with fcGENE:

*./fcgene --impute-hap impute\_ref.hap --impute-legend impute\_ref.legend\*

*--force pheno=unaff,sex=m* (optional)

* + To create files: not applicable

**IMPUTE outputs:**

* example\_impute2:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| snp1 | snp1 | 16212142.00 | T | G | 0.06 | 0.35 | 0.59 | 0.01 | 0.13 | 0.85 | 0.06 | 0.35 | 0.59 | 0.04 | 0.31 | 0.66 |
| snp2 | snp2 | 20278224.00 | T | C | 0.25 | 0.50 | 0.25 | 0.03 | 0.55 | 0.41 | 0.07 | 0.77 | 0.16 | 0.16 | 0.45 | 0.39 |
| snp3 | snp3 | 20304703.00 | G | A | 0.15 | 0.46 | 0.39 | 0.38 | 0.58 | 0.04 | 0.10 | 0.29 | 0.61 | 0.72 | 0.23 | 0.06 |

* example\_info

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| snp\_id | rs\_id | position | exp\_freq\_a1 | info | certainty | type | Concord\_type2 | r2\_type2 |
| snp1 | snp1 | 16212142 | 0.815 | -0.01 | 0.673 | 3 | -1 | -1 |
| snp2 | snp2 | 20278224 | 0.588 | 0.214 | 0.567 | 3 | -1 | -1 |
| snp3 | snp3 | 20304703 | 0.469 | 0.211 | 0.592 | 3 | -1 | -1 |

* fcGENE commands:
  + To read files:

*./fcgene --gens example.impute2 --info example.impute2\_info\*

*--thresh 0.9* (optional) *--info-thresh 0.3\* (optional)

*--maf-thresh 0.1\* (optional)

*-- pedinfo example\_pedinfo.txt* (optional)

* + To create files: not applicable

**BEAGLE input files:**

* example.bgl

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| I | id | ind\_1 | ind\_1 | ind\_2 | ind\_2 | ind\_3 | ind\_3 | Ind\_4 | ind\_4 |
| A | phenotype | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| M | snp1 | G | G | T | G | G | G | T | G |
| M | snp2 | C | C | T | C | C | C | T | T |
| M | snp3 | G | A | G | A | G | A | G | G |

* fcGENE commands:
  + To read files: *./fcgene –bgl example.bgl*
  + To create files: *--oformat beagle*

**BEAGLE references file:**

* beagle\_ref.bgl : see example\_beagle.bgl file
* fcGENE commands:
  + To read data:

*./fcgene --bgl beagle\_ref.bgl\*

*--force pheno=unaff,sex=m* (optional)

* + To create files: not applicable

**BEAGLE-imputation output files**

* example.gprobs:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| marker | alleleA | alleleB | ind\_1 | ind\_1 | ind\_1 | ind\_2 | ind\_2 | Ind\_2 | ind\_3 | ind\_3 | Ind\_3 | ind\_4 | ind\_4 | ind\_4 |
| Snp1 | G | T | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| Snp2 | C | T | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| Snp3 | G | A | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |

* example.bgl.r2

|  |  |
| --- | --- |
| Snp1 | 1.000 |
| Snp2 | 1.000 |
| Snp3 | 1.000 |

* fcGENE commands:
  + To read files:

*./fcgene --bgl-gprobs example.bgl.gprobs\*

*--bgl-rsq example.bgl.r2 --rsq-thresh 0.3\* (optional)

*--pedinfo example\_pedinfo.txt --snpinfo example\_snpinfo.txt* (optional)

* + To create files:not applicable

**BIMBAM input files:**

* example.geno

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4 |  |  |  |  |
| 3 |  |  |  |  |
| IND, | ind\_1, | ind\_2, | ind\_3, | ind\_4 |
| snp1, | TG, | GG, | TG, | GG |
| snp2, | TT, | CT, | TT, | CC |
| snp3, | GG, | GA, | AG, | ?? |

* example.pheno

|  |
| --- |
| 1 |
| 1 |
| 0 |
| 0 |

|  |  |
| --- | --- |
| snp1, | 16212142 |
| snp2, | 20278224 |
| snp3, | 20304703 |

* example.pos.txt
* fcGENE commands:
  + To read files: *./fcgene -- example.geno.txt --pos example.pos.txt*
  + To create files: *--oformat bimbam*

**BIMBAM imputation output-files:**

* example. mean.genotype.txt

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| snp1 | T | G | 2 | 0 | 1 | 0 |
| snp2 | C | T | 0 | 1 | 0 | 2 |
| snp3 | A | G | 0 | 1 | 1 | 0.751 |

* example.genotype.distribution.txt

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| snp1 | T | G | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| snp2 | C | T | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| snp3 | A | G | 1 | 0 | 0 | 1 | 0 | 1 | 0.627 | 0.075 |

* Example.best.guess.genotype.txt

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Snp1 | GT | GG | GT | GG |
| Snp2 | TT | TC | TT | CC |
| Snp3 | GG | GA | GA | GG |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## af is the allele freq for A | | | | | |
| Rs | A | B | af | chr | Pos |
| snp1 | T | G | 0.25 | 22 | 16212142 |
| snp2 | C | T | 0 | 22 | 20278224 |
| snp3 | A | G | 0 | 22 | 20304703 |

* Example.snpinfo.txt

* fcGENE command:
  + To read files:

*./fcgene --wbg example.best.guess.genotype.txt --pos example.snpdata.txt*

*--maf-thresh 0.1* (optional)

*./fcgene --wgd example.genotype.distribution.txt --pos example.snpdata.txt\*

*--maf-thresh 0.1* (optional)

* + To create files: not applicable

**Haploview linkage format**

* Example.ped

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fam\_1 | ind\_1 | 0 | 0 | 1 | 1 | T | T | T | T | G | G |
| fam\_2 | ind\_2 | 0 | 0 | 1 | 1 | G | G | C | T | G | A |
| fam\_3 | ind\_3 | 0 | 0 | 1 | 2 | T | G | T | T | A | G |
| fam\_4 | ind\_4 | 0 | 0 | 2 | 2 | G | G | C | C | 0 | 0 |

* Example.info

|  |  |
| --- | --- |
| Snp1 | 16212142 |
| Snp2 | 20278224 |
| Snp3 | 20304703 |

* fcGENE commands:
  + To read files: not applicable
  + To create files: *--oformat haploview*

**Eigensoft input format**

* example.ped

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fam\_1 | ind\_1 | 0 | 0 | 1 | 1 | T | T | T | T | G | G |
| fam\_2 | ind\_2 | 0 | 0 | 1 | 1 | G | G | C | T | G | A |
| fam\_3 | ind\_3 | 0 | 0 | 1 | 2 | T | G | T | T | A | G |
| fam\_4 | ind\_4 | 0 | 0 | 2 | 2 | G | G | C | C | 0 | 0 |

* Example.pedind

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fam\_1 | ind\_1 | 0 | 0 | 1 | 1 |
| Fam\_2 | ind\_2 | 0 | 0 | 1 | 1 |
| Fam\_3 | ind\_3 | 0 | 0 | 1 | 2 |
| Fam\_4 | ind\_4 | 0 | 0 | 2 | 2 |

* Example.pedsnp

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 22 | snp1 | 0.02 | 16212142 | T | G |
| 22 | snp2 | 0.13 | 20278224 | C | T |
| 22 | snp3 | 0.13 | 20304703 | A | G |
|  |  |  |  |  |  |

* fcGENE commands:
  + To read files: not applicable
  + To create files with fcGENE: *--oformat eigensoft*

**SNPTEST format**

* example.gen : see impute-formatted example.gens
* example.strand : see impute-formatted example.strand.txt
* example.sample :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID\_1 | ID\_2 | missing | sex | status |
| 0 | 0 | 0 | B | B |
| fam\_1 | ind\_1 | 0 | 1 | 0 |
| fam\_2 | ind\_2 | 0 | 1 | 0 |
| fam\_3 | ind\_3 | 0 | 1 | 1 |
| fam\_4 | ind\_4 | 0 | 0 | 1 |

* fcGENE commands:
  + To read files: not applicable
  + To create files with fcGENE:

*--oformat snptest\*

*--covar example/plink\_cov.txtn\*

*--covar-name pheno1,pheno2,covar\_A,covar\_B\*

*--covar-type P,B,D,C*

**Standard text files saved as minor allele coding (012):**

* Example\_genotype.txt

|  |  |  |  |
| --- | --- | --- | --- |
| SMAPLE\_ID | snp1 | Snp2 | snp3 |
| ind\_1 | 2 | 0 | 0 |
| ind\_2 | 0 | 1 | 1 |
| ind\_3 | 1 | 0 | 1 |
| ind\_4 | 0 | 2 | NA |

* example\_affection.txt

|  |  |
| --- | --- |
| SMAPLE\_ID | AFFSTAT |
| ind\_1 | 0 |
| ind\_2 | 0 |
| ind\_3 | 1 |
| ind\_4 | 1 |

* example.freq: (A1 is minor allele )

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CHR | SNP | A1 | A2 | MAF | NCHROBS |
| 22 | snp1 | G | T | 0.375 | 8 |
| 22 | snp2 | C | T | 0.375 | 8 |
| 22 | snp3 | A | G | 0.25 | 6 |

* fcGENE commands: *--oformat r* (or *--oformat R*)

**Standard text files with genotype coding as expected dose of minor alleles:**

* Example\_genotype.txt

|  |  |  |  |
| --- | --- | --- | --- |
| SMAPLE\_ID | snp1 | Snp2 | snp3 |
| ind\_1 | 1.89 | 0.19 | 0.18 |
| ind\_2 | 0.15 | 1.3 | 0.96 |
| ind\_3 | 1.23 | 0.04 | 1.11 |
| ind\_4 | 0.35 | 1.95 | NA |

* fcGENE commands: *--oformat r-dose* (or *--oformat R-dose*)
* example\_affection.txt : see previous type

**fcGENE’s pedInfo and SNPInfo files**:

* example\_pedinfo.txt

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Famid | indid | matid | patid | Sex | phenotype |
| Fam\_1 | ind\_1 | 0 | 0 | 1 | 1 |
| Fam\_2 | ind\_2 | 0 | 0 | 1 | 1 |
| Fam\_3 | ind\_3 | 0 | 0 | 1 | 2 |
| Fam\_4 | ind\_4 | 0 | 0 | 2 | 2 |

* fcGENE’s commands:
  + To read files: *--pedinfo example\_pedinfo.txt*
* example\_snpinfo.txt

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Nchr | snpid | rsid | bp | cm\_pos | allele1 | Allele2 |
| 22 | snp1 | snp1 | 16212142 | 0.02 | T | G |
| 22 | snp2 | snp2 | 20278224 | 0.13 | T | C |
| 22 | snp3 | snp3 | 20304703 | 0.13 | G | A |

* fcGENE commands:
  + To read file: *--snpinfo example\_pedinfo.txt*
  + *--snpinfo* and *--pedinfo* commands can be used with any type of data format readable by fcGENE. An example is given below:

*./fcgene --wgd example.genotype.distribution.txt --pos example.snpdata.txt\*

*--maf-thresh 0.1 --snpinfo example\_snpinfo.txt –pedinfo example\_pedinfo.txt*

**Commands for quality control:**

In this section, we present fcGENE commands required for quality control prior to or after genotype imputation.

**Calculation of SNP-wise call rate and individual wise call rate:** In order to calculate SNP-wise and individual-wise call rate, one can use command option *--crate*. This command will produce two files, namely “snp\_crate.txt” and “indiv\_crate.txt”.

* example\_snp\_crate.txt:

|  |  |  |
| --- | --- | --- |
| CHR | SNP | CRATE |
| 22 | snp1 | 1 |
| 22 | snp2 | 1 |
| 22 | snp3 | 0.75 |
|  |  |  |

* example\_indiv\_crate.txt:

|  |  |  |
| --- | --- | --- |
| FID | IID | CRATE |
| fam\_1 | ind\_1 | 1 |
| fam\_2 | ind\_2 | 1 |
| fam\_3 | ind\_3 | 1 |
| fam\_4 | ind\_4 | 0.75 |

* Example of fcGENE command: *./fcgene --bgl-gprobs example.bgl.gprobs --crate*

**Calculation of p-Values of HWE:** This option performs asymptotic Chi-square testing of Hardy-Weinberg equilibrium. For more details see Wigginton JE, Cutler DJ and Abecasis GR.

“A Note on Exact Tests of Hardy-Weinberg Equilibrium”, Am J Hum Genet (2005) 76: 887-93.

* Resulting file: snp\_hwe.txt:

|  |  |  |
| --- | --- | --- |
| CHR | SNP | PVALUE\_EXACT |
| 22 | snp1 | 0.428571 |
| 22 | snp2 | 0.428571 |
| 22 | snp3 | 1 |

* fcGENE commands:
  + To calculate p-values for testing Hardy-Weinberg Disequilibrium for each SNP, use command option *--hardy*.
  + Example of fcGENE command:

*./fcgene --ped example.ped --map example.map --hardy*

**Calculation of allele frequencies**

* example\_snp.frq:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CHR | SNP | A1 | A2 | MAF | NCHROBS |
| 22 | snp1 | G | T | 0.375 | 8 |
| 22 | snp2 | C | T | 0.375 | 8 |
| 22 | snp3 | A | G | 0.25 | 6 |

* fcGENE commands: Calculation of minor allele frequencies can be performed with command *--freq*, for example *./fcgene --ped example.ped --map example.map --freq*

**Remark:** Calculation of call rate, p-values of HWE, and MAF can be performed within a single command:

*./fcgene --ped example.ped --map example.map\*

*--freq --crate –hardy --out example*

**Discarding SNPs from analysis:**The following command first excludes SNPs specified in the file *snplist.txt* and then performs other tasks such as converting format into BIMBAM or calculating HWE and MAF.

*./fcgene --ped example.ped --map example.map\*

*--exclude snplist.txt --hardy --maf --oformat bimbam\*

*--out bimbam/plink\_bimbam*

**Discarding samples from analysis:** Using command option *--remove*, one can discard individuals from subsequent analyses. The following command first excludes individuals specified in the file *indivlist.txt* and then performs other tasks such as converting format into IMPUTE format and calculating call rates.

*./fcgene --ped example.ped --map example.map\*

*--remove indivlist.txt --crate --oformat impute\*

*--out impute/plink\_impute*

**Quality control of SNPs and Individuals:** In order to filter SNPs and individuals according to specified quality criteria (SNP-wise call rate, HWE and MAF, sample-wise call rate), we can use command options *--filter-snp* and *--filter-indiv*. The following command filters SNPs and individuals first, and then converts the genotype data into IMPUTE format

*./fcgene --ped example.ped --map example.map\*

*--filter-snp hwe=1e-2,maf=0.1,crate=0.95 --filter-indiv crate=0.95*

*--oformat impute\*

*--out impute/plink\_impute*

**Creation of new sample ids for family data**: This option is used to construct hybrid IDs on the basis of family data. We can specify an adequate rule using command option “--iid” in the following way:

|  |  |
| --- | --- |
| Command option | New IDs |
| - -iid famid,iid,sep=\_ | famid\_iid |
| - -iid famid,iid,patid,matid,sep=\_ | famid\_iid\_patid\_matid |
| - -iid famid,iid,patid,matid,sep=- | famid-iid-patid-matid |
| - -iid famid,iid,patid,matid | famidiidpatidmatid |

Table 4: Command options to create new IDs on the basis of pedigree information.

**Example:** *./fcgene --example.ped --map example.map --iid fid,iid,sep=\_ --oformat haploview*

Here, command option *“-- iid fid,iid,sep= \_”* creates new dummy sample IDs of the form “FamilyID\_IndividualID”.

**Multiple commands at a time:**

For the efficient use of fcGENE, one can apply multiple commands at a time. If we want to perform two or more tasks in fcGENE, these tasks can be separated by command options *--new-start* and *--new-end*. The following examples show how to perform multiple tasks in fcGENE simultaneously.

* When a fcGENE’s command contains multiple format converting tasks, each new task, except for the first, is separated by command identifiers *--new-start* and *--new-end*. The following command reads two PLINK-formatted files, and convert the first into MaCH and second into IMPUTE format.

*./fcgene --ped example1.ped --map example1.map\*

*--oformat mach --out mach/example\*

*--new-start\*

*--dosage example2.dose --fam example2.fam\*

*--map example2.map --oformat impute --out impute/example*

*--new-end*

* In the next example, two datasets were merged and then converted into EIGENSOFT format

*./fcgene --ped example1.ped --map example1.map\*

*--new-start\*

*--gens impute/example2.gens --pedinfo impute/example2\_pedinfo.txt \*

*--merge\*

*--new-end \*

*--out eigensoft/example\_merge --oformat eigensoft*

While merging two data, the first data given outside of   
--new-start and --new-end is considered as the basics of the final merged data. That means only SNPs and individuals given in other data but not in the first basic data are added to it . We can use “--merge” command if we want to merge a new data given within “--new-start” and “--new-end”, we mention the command “--merge” with in “--new-start” and “--new-end”.

* The following command reads three datasets, merge the first two of them before the merged datasets are converted into PLINK dosage format. Call rate, HWE and MAFs are calculated for the third dataset. Then it is converted into BEAGLE format.
  + *./fcgene --ped example1.ped --map example1.map\*

*--new-start\*

*--ped mach/example2.ped --dat mach/example2.dat\*

*--snpinfo mach/example2\_snpinfo --filter-snp hwe=1e-2\*

*--merge\*

*--new-end\*

*--new-start \*

*--gens impute/example3.gens --pedinfo impute/example3\_pedinfo.txt \*

*--hardy --crate --freq --oformat beagle \*

*--out beagle/impute\_beagle\*

*--new-end \*

*--out plink/example\_dosage --oformat plink-dosage*