Updates in New Versions

This file gives short information about the updates in new version.

* **Updates in Version 1.07 :** 
  + fcGENE can compute genetic similarity like F\_ST and G\_ST   
    <http://www.biomedcentral.com/1471-2156/16/90>

Fst and Gst can be computed by using the command

--fst --group-label population\_group\_label.txt

Population\_group\_label.txt file contacts two columns: individual ids in the first column and population group name in the second column as below.

Indiv1 pop1

Indiv2 pop1

Indiv3 pop2

* + fcGENE can convert shapeit outputs : haps and sample file, into many other formats Haps and sample files can be read using “--haps” and “--sample” options.

**Some examples:**   
./fcgene  --haps shapeit\_output.haps --sample shapeit\_output.sample --oformat snptest   
./fcgene  --haps shapeit\_output.haps --sample shapeit\_output.sample --oformat impute   
To convert it into plink   
./fcgene  --haps shapeit\_output.haps --sample shapeit\_output.sample --oformat plink  
./fcgene  --haps shapeit\_output.haps --sample shapeit\_output.sample --oformat plink-bed

* + fcGENE can convert genotype data into genABEL and probABEL programs with commands   
    --oformat genabel and –oformat probabel respectively

Since probABEL program accepts mach-formatted output data, fcGENE converts the outputs into mach-formatted mlinfo, mlprob files. Moreover, since mlinfo file contains imputation quality rsq and quality masses and they can’t be created by other tools, fcGENE writes 0.0 in the columns of these two masses.

* + fcGENE can read and convert input files into plink binary files( bed,bim and fam files):
    - For input --bfile filename or

--bed filename.bed --bim filename.bim --fam filename.fam

* + - For output --oformat plink-bed
  + fcGENE can perform the following additional commands in r-formatted data:
    - This version can read and write r-format files, this is the count of allele1 in the genotype data. This means homozygote of the reference allele is noted as 2 and homozygote of alternative allele is noted as 0. The heterozygote is noted as 1.

Command to read the genotype data

./fcgen --rgeno mydata --snpinfo allele\_info\_file.txt

* + - fcGENE can write both normal and transformed form of r-file
      * command --oformat r or --oformat R and
      * --transpose --oformat r --transpose --oformat R
  + fcGENE can read and write Gzipped files.
    - fcGENE can now read and write gezipped data. No commands are necessary (at the moment )
  + fcGENE can split genotype data :
    - fcGENE can all type of formatted data into any number of subsets given by the two commands --ssplitt and –isplitt
    - After splitting, one can change into another format by using

--oformat . If --oformat is not used, then splitting command writes the splitted data into the original format.

* + - for bp split --bpsplitt
  + fcGENE can write vcf format files
    - --oformat vcf
  + fcGENE can calculate pairwise fST among individuals.
    - Command - -fst
* **Updates in Version 1.06:**
  + This version of fcGENE can perform multiple command options at a time
    - 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 this command, we have performed two tasks at one time. The first task is to convert plink-formatted ped and data file into mach format and the send task is to convert plink-formatted dosage files into impute format. These two tasks has here independent of each other but given as one command.

* + This version can also merge two genotype data formatted in the same way or differently. 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”. An example of merging two data is given below.

* + - 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*

In this command first the best case genotypes of IMPUTE-formatted data are determined from their probability distribution and then they are merged into the primary data example.ped and example.map given in plink format. The merged data will be converted into eigensoft format and the name of final data will then be example\_merge

* + 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*

* + This version of fcGENE can convert data into phase/fastPHASE format also.
  + **Strand alignment:** Strand alignment between genotype dataset and reference data set is crucial for GWA analysis and imputation. Generally, reference panels such as HapMap are given as ‘+’ strand but they might be genotyped with respect to negative strand. fcGENE supports strand flipping by the following approach: Use fcGENE first to merge the genotype SNP data and the given reference panel and then to convert them into plink format While merging these two data, individuals of genotype SNP data and reference panel should be assigned with dummy cases and control status respectively. This can be done by applying command option “--force” in fcGENE. Finally, strand mismatches can be detected by applying “--flip-scan" or “-- flip-scan-verbose" commands of PLINK [1].
  + We realized that coding genotypes as the counts of minor alleles may be very ideal when we apply different statistical models like regression and ANOVAs. PLINK supports to produce this type of format by generating two forms of raw-formatted files by using command options “--recodeA” and “--recodeAD”, but these two options support only the hard calls of genotype data meaning use of only the numbers 0, 1 and 2 to represent genotypes of the form homozygote major, heterozygote and homozygote minor respectively. Alongside the support to PLINK-users by converting different types of imputation results into previously mentioned two forms of raw-formatted files, fcGENE provides the facility also to transform the data with genotype probability distribution into the form of PLINK’s recodeA-formatted files but filled with expected minor- allele-doses. This means the raw files transformed by fcGENE can contain not only 0,1 and 2 as minor allele counts but also the expected allele dose of minor allele, which can be any fractional number between 0 and 2. If “A” and “B” are two alleles of A SNP with “B” as minor allele, then the expected minor allele dose can be calculated as
* 

where  and  are the probabilities having genotypes ,  and  at an individual respectively. GWA analysis with this type of coding may be very useful especially when the accuracy of imputation results is low because it can account for the uncertainty in the imputed genotypes [18].

Using command options “--oformat plink-recodeA” and “--oformat plink-recodeAD , we can obtain PLINK’s raw files.

To crate PLINK-formatted raw dose file (see below example\_dose.raw), command option “--format recodeA-dose” is used.

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\_dose.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 |

* **Updates in version: 1.05:**
  + This version of fcGENE can calculate snp-wise quality measure call rate, HWE and MAF and individual-wise call rate. For this purpose use “- -crate”, “- -hardy”

and “- - freq” options.

* + Filtering SNPs and individual according as their quality. No need to calculate quality measures separately. Just use “- -filter-snp” and “- -filter-indiv option” .
  + Format conversion into plink’s raw format (- -recodeA option), use

“- -oformat plink-recodeA” option.

Format conversion of plink’s raw format (- - recode AD option)

use “- -oformat plink-recodeAD” option.

* **Updates in version: 1.04:**

In this version, we have updated the followings:

* + Individuals in plink or mach (merlin) formatted family genotype data may be observed as unique only if we consider family ids , individual ids, paternal and /or maternal ids collectively. This type of uniqueness may be lost if we observe individual or family ids only. However some imputation software like SHAPEIT or Eigensoft, which observe only individuals ids, may not work properly unless we create new unique id for each individual. At such condition, We may have to combine family-ids, and /or individual-ids, and/or paternal ids and/or maternal ids. We can specify a rule, with which fcGENE can create unique individual ids while converting genotype data into the required format. Command option “ - -iid ” specify this type of rule. For example if we want to create new individual ids by combining family ids and individual ids with a character “\_” as separator, then we can mention this rule in fcGENE as “- - iid fid,iid,sep=\_ ”, which creates new individual ids in the form: “familyid\_individualid”. Any other character like “-”, “.”, “->” or even a string or nothing can be used as separator between family ids and individual ids and/or others. One can also combine patids and matids together with famids and individual ids by using the command option for example

“- -iid fid,iid,patid,matid,sep=->”.

* + Converting data into haploview format
  + Converting data into smartpca format
  + Converting minimac dose file into plink.dose file
  + Converting genotype data into simple standard r-readable format . By standard we mean rows for observations (individuals) and columns for variables (SNPs), with simple headers for gene names(rsids) and first column to identify subjects (i.e. pedigree ids). An sized genotype data matrix with dosage values for genotype (number of minor alleles) would look like:

  SMAPLE\_ID        rsid\_1     rsid\_2      …             rsid\_M

ID\_1                       0             2                                 2

ID\_2           1             1         …                  2

… … … … …

ID\_N                      1              0   …                0

Similarly an affection status file having 0 for cases and 1 for control would look like:

SMAPLE\_ID        AFFSTAT

ID\_1                       1

ID\_2                       0

…                           ..

ID\_N                      1

Previously mentioned two text files can be obtained as output of fcGENE by using the command option “- -oformat R ” or “- -oformat r”. These two files may be useful when we analyse the data in R. There are R packages for the analysis of genetic data however it is often useful to get data into the previously mentioned format when using self-planned analysis in R.

* **Other commands are same as in the older version except “ *- -oformat” . i.e.***
  + To convert data into haploview use “ - -oformat haploview “
  + To convert data into smartpca use “- - oformat smartpca ”
  + To convert data into plink dose file use “ - - oformat plink-dosage”
  + For example to convert genotype data given in any format (plink , mach, impute , beagle and BIMBAM) according as it is used in the following two example commands.
  + To convert RESULT of impute 2 into plink dosage file

./fcgene - - gens mydata.gens\

- - pedinfo mydata\_pedinfo.txt\

- -oformat plink-dosage - -out mydata

* + To convert RESULT of MACH into plink dosage file

./fcgene -- -mach-mlprob mydata.mlprob\

- -mach-mlinfo mydata.mlinfo\

- - pedinfo mydata\_pedinfo.txt \

- -oformat plink-dosage - -out mydata

* + Note that if there are missings in original genotype file, then plink-dose file cannot be created since plink assumes that all genotypes are already genotyped or imputed.
  + The plink command to read the plink-dose formatted data is :

***./plink --dosage myfile.dose --fam mydata.fam --map mymap.map***

This command will create a plink.assoc.dosage

* + For more information on plink dosage files, please visit the website

[*http://pngu.mgh.harvard.edu/~purcell/plink/dosage.shtml*](http://pngu.mgh.harvard.edu/~purcell/plink/dosage.shtml)