User manual for *funbiased*

Mehreen R. Mughal, Michael DeGiorgio

November 16, 2020

Contents

1 Introduction

The `funbiased` script can be used to compute the unbiased estimators of F_2, F_3, normalized F_3, and normalized F_4 statistics introduced in Mughal and DeGiorgio (2020). These statistics are used to understand the relationships among two, three, or four populations using allele frequency data at biallelic polymorphic sites. Operation of this package requires a UNIX environment with Python 2.7.

Please cite this software as

MR Mughal and M DeGiorgio (2020) Properties and unbiased estimation of F- and D-statistics in samples containing related and inbred individuals. *bioRxiv* doi:XXXXX

If you experience any issues, then please contact Mehreen Mughal at mrm79@psu.edu for assistance.

2 Operation

2.1 Installation

To download `funbiased` to your current directory use the command

```
git clone https://github.com/MehreenRuhi/funbiased
```

This command will download a copy of the user manual, the software, and the example data.

2.2 Input file formats

Genotype data stored in VCF file format (Danecek et al., 2011) is used by this software to compute $\widetilde{F}$. In this format each line contains information regarding the position, alleles, and quality of the calls in addition to the genotype calls for each individual in the population. We require separate VCF files for each population, and require that all VCF files contain information for the same sites. The software will compute the estimator using all sites provided.

In addition, we require the user to input a kinship matrix for each population describing how each individual in the population contained in the corresponding VCF file is related to each other. The user must input as many kinship matrices are there are VCF files. This symmetric $(N \times N)$ matrix will contain the same number of rows and columns as there are individuals sampled in the population (N). The jth row and kth column of this matrix contains the kinship coefficients Φ_{jk} relating an individual in row j to an individual in column k, with the diagonal representing self kinship coefficient Φ_{kk}. For example, if the VCF file contains six unrelated individuals, then the kinship file will look like the following:

$$\begin{array}{cccccc} 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 \end{array}$$

Individuals in both VCF and kinship files must be arranged in the same order. For example, the individual whose relatedness is described in the first row and first column of the kinship matrix must be the first genotyped individual in the VCF file. If a genotype is missing, indicated by "./.", then the software will recompute the mean kinship coefficient for that site by removing individuals with missing genotypes.

2.3 Computing $\widetilde{F}_2(A, B)$ from genotypes in VCF files

To compute $F_2(A, B)$ using genotype information contained in VCF files for two populations, use the following command

```
python funbiased.py F2 <input_vcf_popA> <input_vcf_popB> <input_kinship_popA>
<input_kinship_popB>
```

where `input_vcf_popA` and `input_vcf_popB` are the VCF files for populations A and B (Mughal and DeGiorgio, 2020), and `input_kinship_popA` and `input_kinship_popB` are the corresponding kinship files. This command will print to the screen the unbiased estimator $\widetilde{F}_2(A, B)$, followed by the biased estimator $\widehat{F}_2(A, B)$.

2.4 Computing $\widetilde{F}_3(A; B, C)$ and normalized $\widetilde{F}_3(A; B, C \mid A)$ from genotypes in VCF files

To compute $\widetilde{F}_3(A; B, C)$ using genotype information contained in VCF files for three populations, use the following command

```
python funbiased.py F3 <input_vcf_popA> <input_vcf_popB> <input_vcf_popC>
<input_kinship_popA> <input_kinship_popB> <input_kinship_popC>
```

This command will print to the screen the unbiased estimator $\widetilde{F}_3(A; B, C)$, followed by the biased estimator $\widehat{F}_3(A; B, C)$. Here the `input_vcf_popX` are VCF files containing the genotype information for each population $X \in \{A, B, C\}$, while the `input_kinship_popX` files contain the relationship information for all individuals contained in the corresponding VCF files for each population X.

Similarly, to compute the normalized $F_3(A; B, C \mid A)$ estimator instead use the command

```
python funbiased.py F3norm <input_vcf_popA> <input_vcf_popB> <input_vcf_popC>
<input_kinship_popA> <input_kinship_popB> <input_kinship_popC>
```

This command will print to the screen the approximately unbiased estimator $\widetilde{F}_3(A; B, C \mid A)$, followed by the biased estimator $\widehat{F}_3(A; B, C \mid A)$.

2.5 Computing normalized $\widetilde{F}_4(A, B; C, D \mid P)$ from genotypes in VCF files

To compute the normalized $F_4(A, B; C, D \mid P)$ estimator use the command

```
python funbiased.py F4norm <input_vcf_popA> <input_vcf_popB> <input_vcf_popC>
<input_vcf_popD> <input_kinship_popA> <input_kinship_popB> <input_kinship_popC>
<input_kinship_popD> <population_P>
```

Here the `input_vcf_popX` are VCF files containing the genotype information for each population $X \in \{A, B, C, D\}$, while the `input_kinship_popX` files contain the relationship information for all individuals contained in the corresponding VCF files for each population X. The parameter `population_P` is used to specify which population $P \in \{A, B, C, D\}$ is used in the denominator to normalize the F_4 statistic. This command will print to the screen the approximately unbiased estimator $\widetilde{F}_4(A, B; C, D \mid P)$, followed by the biased estimator $\widehat{F}_4(A, B; C, D \mid P)$.

3 Examples

We provide some example VCF files, along with example kinship matrices for users to test our software and to compare their own input files to the files we have provided. In the directory titled `Example_data` we have the four VCF files `popA.vcf`, `popB.vcf`, `popC.vcf`, and `popD.vcf`, along with the four kinship matrices describing the relatedness between individuals sampled within these populations, `popA.kin`, `popB.kin`, `popC.kin`, and `popD.kin`. Each population contains $N = 20$ sampled individuals, and for this reason the kinship matrix has 20 rows and 20 columns. However, our software does not require that all input populations contain an identical number of individuals.

To use these example files to compute $F_2(A, B)$ use the command

```
python funbiased.py F2 Example_data/popA.vcf Example_data/popB.vcf
Example_data/popA.kin Example_data/popB.kin
```

This command will print out the unbiased $\widetilde{F}_2(A, B)$ and biased $\widehat{F}_2(A, B)$ estimators of $F_2(A, B)$ using the genotype data found in `popA.vcf` and `popB.vcf`.

To use the example files to compute $F_3(A; B, C)$ use the command

```
python funbiased.py F3 Example_data/popA.vcf Example_data/popB.vcf
Example_data/popC.vcf Example_data/popA.kin Example_data/popB.kin
Example_data/popC.kin
```

This command will print out the unbiased $F_3(A; B, C)$ and biased $\widehat{F}_3(A; B, C)$ estimators of $F_3(A; B, C)$ using the genotype data found in `popA.vcf`, `popB.vcf`, and `popC.vcf`. To compute the normalized $F_3(A; B, C \mid A)$ estimators use `F3norm` intsead of `F3` in the above command.

Finally, to use the example files to compute normalized $\widetilde{F}_4(A, B; C, D \mid A)$ use the command

```
python funbiased.py F4norm Example_data/popA.vcf Example_data/popB.vcf
Example_data/popC.vcf Example_data/popD.vcf Example_data/popA.kin
Example_data/popB.kin Example_data/popC.kin Example_data/popD.kin A
```

This command will print out the approximately unbiased $\widetilde{F}_4(A, B; C, D \mid A)$ and biased $\widehat{F}_4(A, B; C, D \mid A)$ estimators of normalized $F_4(A, B; C, D \mid A)$ using the genotype data found in `popA.vcf`, `popB.vcf`, `popC.vcf`, and `popD.vcf`.

References

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and . G. P. A. Group. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

M. R. Mughal and M. DeGiorgio. Properties and unbiased estimation of F- and D-statistics in samples containing related and inbred individuals. *bioRxiv*, 2020.