

# Contents

<b>1</b>	<b>Guide for xerxes v0.2.0.0</b>	<b>1</b>
1.1	Fstats command	1
1.1.1	Allowed statistics	3
1.1.2	Defining statistics directly via <code>--stat</code>	3
1.1.3	Defining statistics in a simple text file	3
1.1.4	Input via a configuraton file	4
1.1.5	Output	5
1.1.6	Whitepaper	5
1.2	RAS (in development)	6

## 1 Guide for xerxes v0.2.0.0

### 1.1 Fstats command

Xerxes allows you to analyse genotype data across poseidon packages, including your own, as explained above by “hooking” in your own package via a `--baseDir` (or `-d`) parameter. This has the advantage that you can compute arbitrary F-Statistics across groups and individuals distributed in many packages, without the need to explicitly merge the data first. Xerxes also takes care of merging PLINK and EIGENSTRAT data on the fly. It also takes care of different genotype base sets, like Human-Origins vs. 1240K. It also flips alleles automatically across genotype files, and throws an error if the alleles in different packages are incongruent with each other. Xerxes is also smart enough to select only the packages relevant for the statistics that you need, and then streams through only those genotype data.

Here is an example command for computing several F-Statistics:

```
xerxes fstats -d ... -d ... \  
  --stat "F4(<Chimp.REF>, <Altai_published.DG>, Yoruba, French)" \  
  --stat "F3(<Chimp.REF>, <Altai_snpAD.DG>, Spanish)" \  
  --statFile fstats.txt  
  --statConfig fstats.yaml  
  -f outputfile.txt
```

First, the two options `-d ...` exemplify that you need to provide at least one base directory for poseidon packages, but can also give multiple. Second, F-Statistics can be entered in three different ways:

1. Directly via the command line using `--stat`.
2. Using a simple text file using `--statFile`
3. Using a powerful configuration file that allows more options.

These three input ways can be mixed and matched, and given multiple times. They are explained below.

Last, option `-f` can be used to write the output table into a tab-separated text file, beyond just printing a table into the standard out when the program finishes. Note that there are more options, which you can view using `xerxes fstats --help`:

```
Usage: xerxes fstats (-d|--baseDir DIR) [-j|--jackknife ARG]  
      [-e|--excludeChroms ARG]  
      (--stat ARG | --statConfig ARG | --statFile ARG)
```

```

40         [--noTransitions] [-f|--tableOutFile ARG]
41         [--blockTableFile ARG]
42
43 Compute f-statistics on groups and individuals within and across Poseidon
44 packages
45
46 Available options:
47 -h,--help                Show this help text
48 -d,--baseDir DIR         A base directory to search for Poseidon packages.
49 -j,--jackknife ARG       Jackknife setting. If given an integer number, this
50                           defines the block size in SNPs. Set to "CHR" if you
51                           want jackknife blocks defined as entire chromosomes.
52                           The default is at 5000 SNPs
53 -e,--excludeChroms ARG   List of chromosome names to exclude chromosomes,
54                           given as comma-separated list. Defaults to X, Y, MT,
55                           chrX, chrY, chrMT, 23,24,90
56 --stat ARG               Specify a summary statistic to be computed. Can be
57                           given multiple times. Possible options are: F4(a, b,
58                           c, d), F3(a, b, c), F3star(a, b, c), F2(a, b), PWM(a,
59                           b), FST(a, b), Het(a) and some more special options
60                           described at
61                           https://poseidon-framework.github.io/#/xerxes?id=fstats-command.
62                           Valid entities used in the statistics are group names
63                           as specified in the *.fam, *.ind or *.janno files,
64                           individual names using the syntax "<Ind_name>", so
65                           enclosing them in angular brackets, and entire
66                           packages like "*Package1*" using the Poseidon package
67                           title. You can mix entity types, like in
68                           "F4(<Ind1>,Group2,*Pac*,<Ind4>)". Group or individual
69                           names are separated by commas, and a comma can be
70                           followed by any number of spaces.
71 --statConfig ARG         Specify a yaml file for the Fstatistics and group
72                           configurations
73 --statFile ARG           Specify a file with F-Statistics specified similarly
74                           as specified for option --stat. One line per
75                           statistics, and no new-line at the end
76 --maxSnps ARG            Stop after a maximum nr of snps has been processed.
77                           Useful for short test runs
78 --noTransitions           Skip transition SNPs and use only transversions
79 -f,--tableOutFile ARG    a file to which results are written as tab-separated
80                           file
81 --blockTableFile ARG     a file to which the per-Block results are written as
82                           tab-separated file

```

### 83 1.1.1 Allowed statistics

84 The following statistics are allowed in the `--stat`, `--statFile` and `--statConfig` options. In all of the following,  
85 symbols `a`, `b`, `c` or `d` stand for arbitrary entities allowed in Poseidon, so groups (such as `French`), individuals  
86 (such as `<MA1.SG>`) or packages (such as `*2012_PattersonGenetics*`).

- 87 • `F2vanilla(a, b)`: F2-Statistics - Vanilla version. Computed using  $F2vanilla(a, b) = (a-b)^2$  across  
88 the genome.
- 89 • `F2(a, b)`: F2-Statistics (bias-corrected version). Computed as  $F2(a, b) = F2vanilla(a, b) - hA/sA$   
90  $- hB/sB$ , where `sA` is the number of non-missing alleles in entity `A`, and  $hA = nA * nA' / sA *$   
91  $(sA - 1)$  is an estimator of half the heterozygosity (see `Het(a)`), and likewise for `sB` and `nB` etc.
- 92 • `F3vanilla(a,b,c)`: F3-Statistics - Vanilla version, recommended if used as Outgroup-F3 statistics or with  
93 group `c` being pseudo-haploid: Are computed as  $F3(a, b, c) = (c-a)(c-b)$  across all SNPs.
- 94 • `F3(a,b,c)`: F3-statistics (bias-corrected version). Computed as  $F3(a, b, c) = F3vanilla(a, b) -$   
95  $hC/sC$ .
- 96 • `F3star(a,b,c)`: F3-Statistics as defined in Patterson et al. 2012 - normalised and bias-corrected version,  
97 recommended for Admixture-F3 tests. Are computed by i) first subtracting per SNP from the vanilla-F3  
98 statistic a bias-correction term  $hC/sC$ , as above for `F2`, and ii) then normalising the genome-wide estimate  
99 by a genome-wide estimate of the heterozygosity of entity `C` (`Het(c)`), in order to make results comparable  
100 between different groups `C` (see Patterson et al., Genetics, 2012)
- 101 • `F4(a,b,c,d)`: F4 statistics. Are computed by averaging the quantity  $(a-b)(c-d)$  across all SNPs. No bias  
102 correction is necessary for this statistic.
- 103 • `Het(a)`: An estimate of the heterozygosity across all SNPs, computed as  $2 * hA$ , with `hA` defined as above in  
104 `F2`
- 105 • `FST(a, b)`: An estimate of `FST` across the genome, following the formular from Appendix A in Patterson  
106 et al. 2012, which is a ratio of two terms, with numerator being `F2(a, b)` including bias correction, and  
107 the denominator being  $F2(a, b) + hA + hB$  including bias correction and `hA` and `hB` defined as above.
- 108 • `PWM(a, b)`: The pairwise mismatch rate between entities `a` and `b`, computed from allele frequencies as  $a$   
109  $(1 - b) + (1 - a) b$ .

110 All of these equations are from Patterson, Nick, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland,  
111 Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. 2012. "Ancient Admixture in Human History."  
112 Genetics 192 (3): 1065–93. See also Appendix A of this paper for the unbiased estimators used above.

113 For each of the "slots" `A`, `B`, `C` or `D`, you can enter: \* Individuals, using the syntax `<Individual_Name> *`  
114 Groups, using no special syntax `"Group_Name"` \* Packages, using syntax `*Package_Name*` (This can be useful  
115 if you happen to have a homogenous set of individuals from multiple groups in one package and want to consider  
116 all of these as one group.)

### 117 1.1.2 Defining statistics directly via --stat

118 This is the simplest option to instruct the program to compute a specified statistic. Each statistic requires a  
119 separate input using `--stat` using this input method. Example:

```
120 xerxes fstats -d ... -d ... --stat "F3(French, Spanish, <Chimp.REF>) --stat "FST(French,  
121 Spanish)"
```

### 122 1.1.3 Defining statistics in a simple text file

123 You can prepare a text file, into which you write the above statistics, one statistics per line. Example:

```

124 F4(<Chimp.REF>, <Altai_published.DG>, Yoruba, French)
125 F4(<Chimp.REF>, <Altai_snpAD.DG>, Spanish, French)
126 F4(Mbuti,Nganasan,Saami.DG,Finnish)
127 you can then load these statistics using the option --statFile fstats.txt.

```

#### 1.1.4 Input via a configuraton file

This is the most powerful way to input F-Statistics. Here is an example:

```

130 groupDefs:
131   CEU2: ["CEU.SG", "-<NA12889.SG>", "-<NA12890.SG>"]
132   FIN2: ["FIN.SG", "-<HG00383.SG>", "-<HG00384.SG>"]
133   GBR2: ["GBR.SG", "-<HG01791.SG>", "-<HG02215.SG>"]
134   IBS2: ["IBS.SG", "-<HG02238.SG>", "-<HG02239.SG>"]
135 fstats:
136 - type: F2
137   a: ["French", "Spanish"]
138   b: ["Han", "CEU2"]
139   # Ascertainment is optional
140 - type: F3 # This will create 3x2x1 = 6 Statistics
141   a: ["French", "Spanish", "Mbuti"]
142   b: ["Han", "CEU2"]
143   c: ["<Chimp.REF>"]
144   ascertainment:
145     outgroup: "<Chimp.REF>" # ascertaining on outgroup-polarised derived allele frequency
146     reference: "CEU2"
147     lower: 0.05
148     upper: 0.95
149 - type: F4 # This will create 5x5x4x1 = 100 Statistics
150   a: ["<I0156.SG>", "<I0157.SG>", "<I0159.SG>", "<I0160.SG>", "<I0161.SG>"]
151   b: ["<I0156.SG>", "<I0157.SG>", "<I0159.SG>", "<I0160.SG>", "<I0161.SG>"]
152   c: ["CEU2", "FIN2", "GBR2", "IBS2"]
153   d: ["<Chimp.REF>"]
154   ascertainment:
155     # A missing outgroup means: ascertain on minor allele frequency
156     reference: "CEU.SG"
157     lower: 0.00
158     upper: 0.10

```

The top level structure of this [YAML](#) file is an object with two fields: **groupDefs** (which is optional) and **fstats** (which is mandatory).

**1.1.4.1 Group Definitions** You can specify adhoc group definitions using the syntax above. Every group consists of a name (used as object key) and then a JSON- or YAML-list of signed entities, following the same syntax of **trident** forge (see [trident](#)). Briefly: Individuals, Groups and Packages can be added or excluded (prefixed by a -) in order. In the example above, two individuals are removed from each group.

Note that currently, groups can be defined only independently, so not incremental to each other. That means, you cannot currently use an already defined new group name in the entity list of a following group name.

**1.1.4.2 Statistic input using YAML** Each statistic defined in the `fstats` section of the YAML file, actually defines a loop over multiple populations in each statistic. In the example above, there are 6 F3-Statistics, each using a different combination of the input groups defined in each of the `a:`, `b:` and `c:` slots. There are also 100 (!) F4 statistics, following all combinations of 5x5x4x1 slots defined in `a:`, `b:`, `c:` and `d:`. This makes it very convenient to loop over statistics.

**1.1.4.3 Ascertainment (experimental feature)** In addition, every statistic section allows for a definition of an ascertainment specification, using a special key `ascertainment:`, which is optional. If given, you can specify an optional `outgroup`, a `reference` group in which to ascertain SNPs, and `lower` and `upper` allele frequency bounds. If specified, only SNPs for which the `reference` group has an allele frequency within the given bounds are used to compute the statistic (note that normalisation is still using all non-missing SNPs for that given statistic). If an `outgroup` is defined, then the outgroup-polarised derived allele frequency is used. If no `outgroup` is defined, then the minor allele frequency is used instead. If an outgroup is defined, any sites where the outgroup is polymorphic are treated as missing.

You can save this into a text file, for example named `fstats_config.yaml`, and load it via `--statConfig fstats_config.yaml`.

## 1.1.5 Output

The final output of the `fstats` command looks like this:

```

-----
| Statistic |      a      |      b      |      c      |      d      | NrSites | Estimate_Total | Estimate_Jackknife |
:=====:
| F3        | French      | Italian_North | Mbuti      |              | 593124 | 5.9698e-2      | 5.9698e-2          |
| F3        | French      | Han           | Mbuti      |              | 593124 | 5.0233e-2      | 5.0233e-2          |
| F3        | Sardinian   | Pima          | French     |              | 593124 | -1.2483e-3     | -1.2483e-3         |
| F4        | French      | Russian       | Han        | Mbuti      | 593124 | -1.6778e-3     | -1.6778e-3         |
| F4        | Sardinian   | French        | Pima       | Mbuti      | 593124 | -1.4384e-3     | -1.4384e-3         |
-----

```

which lists each statistic, the slots a, b, c and d, the number of sites with non-missing data for that statistic, Ascertainment information (outgroup, reference, lower and upper bound, if given), the genome-wide estimate, its standard error and its Z-score. If you specify an output file using option `--tableOutFile` or `-f`, these results are also written as tab-separated file.

Additionally, an option `--blockOutFile` can be specified, to which then a table with estimates per Jackknife block is written.

## 1.1.6 Whitepaper

The repository comes with a [detailed whitepaper](#) that describes some more mathematica details of the methods implemented here.

## 202 1.2 RAS (in development)

203 The RAS command computes pairwise RAS statistics between a collection of “left” entities, and a collection of  
204 “right” entities. Every Entity is either a group name or an individual, with the similar syntax as in F-statistics  
205 above, so **French** is a group, and **<IND001>** is an individual.

206 The input of left-pops and right-pops uses a YAML file via `--popConfigFile`. Here is an example:

```
207 groupDefs:
208   group1: a,b,-c,-<d>
209   group2: e,f,-<g>
210 popLefts:
211   - <I13721>
212   - <I14000>
213   - <I13722>
214   - <Iceman.SG>
215 popRights:
216   - Mbuti
217   - Mixe
218   - Spanish
219 outgroup: <Chimp.REF>
```

220 In this case, two groups are defined on the fly: **group1** comprises groups **a** and **b**, but excludes group **c** and  
221 individual **d**. Note that inclusions and exclusions are executed in order. **group2** comprises of group **e** and group  
222 **f**, but excludes individual **<g>**.

223 As in [RAScalculator](#), the allele frequency ascertainment is done across right populations only.

224 There are a couple of options, as specified in the CLI help (`xerxes ras --help`):

```
225 Usage: xerxes ras (-d|--baseDir DIR) [-j|--jackknife ARG]
226               [-e|--excludeChroms ARG] --popConfigFile ARG
227               [-k|--maxAlleleCount ARG] [-m|--maxMissingness ARG]
228               (-f|--tableOutFile ARG)
229 Compute RAS statistics on groups and individuals within and across Poseidon
230 packages
```

232 Available options:

233	<code>-h,--help</code>	Show this help text
234	<code>-d,--baseDir DIR</code>	a base directory to search for Poseidon Packages (could be a Poseidon repository)
235		
236	<code>-j,--jackknife ARG</code>	Jackknife setting. If given an integer number, this 237 defines the block size in SNPs. Set to "CHR" if you 238 want jackknife blocks defined as entire chromosomes. 239 The default is at 5000 SNPs
240	<code>-e,--excludeChroms ARG</code>	List of chromosome names to exclude chromosomes, 241 given as comma-separated list. Defaults to X, Y, MT, 242 chrX, chrY, chrMT, 23,24,90
243	<code>--popConfigFile ARG</code>	a file containing the population configuration

```

244  -k,--maxAlleleCount ARG  define a maximal allele-count cutoff for the RAS
245                          statistics. (default: 10)
246  -m,--maxMissingness ARG  define a maximal missingness for the right
247                          populations in the RAS statistics. (default: 0.1)
248  -f,--tableOutFile ARG    the file to which results are written as
249                          tab-separated file

```

250 The output gives both cumulative (up to allele-count k) and and per-allele-frequency RAS (for allele count k) for  
251 every pair of left and rights. The standard out contains a pretty-printed table, and in addition, a tab-separated  
252 file is written to the file specified using option -f.

253 **xerxes ras** makes a few important assumptions: 1) It assumes that the Right Populations are “nearly” completely  
254 non-missing. Any allele that is actually missing from the rights is in fact treated as homozygous-reference! A  
255 different approach would be to compute the actual frequencies on the non-missing right alleles, but then we  
256 cannot anymore nicely accumulate over different ascertainment allele counts. 2) If no outgroup is specified, the  
257 ascertainment operates on minor-allele frequency (as in fstats) 3) If an outgroup is specified and missing from a  
258 SNP, or if the SNP is polymorphic, the SNP is skipped as missing