ANGSD formats

tsk

April 13, 2016

1   SAF formats

SAF files are files that contain sample allele frequency. These are generated with -doSaf in main ANGSD. These contains either the loglikelihood ratio to the most likely category or the pp. This is determined if the -prior has been supplied. The first 8 bytes magic number determines which SAF version. If no magic number is present then version0 is assumed.

1.1   version 0

First version of the SAF files were simply flat binary double files `PREFIX.saf` along with an associated `PREFIX.saf.pos.gz` which contains the gzip compressed 'chromosome position'. Assuming $nChr$ number of chromosomes, then we have $nChr+1$ categories for each site. The number of sites can therefore be deduced either directly from the number of lines in the uncompressed output of the `PREFIX.saf.pos.gz`, or by using the filesize ($fsize$) of the `PREFIX.saf`

$$\frac{fsize}{sizeof(double) * (nChr + 1)}.$$

1.2   version 1

Second iteration of the saf files now contains two raw files and an index file. Still under development. First 8 bytes in all three files is 8byte magic numer $char[8]$ "safv3".

PREFIX.saf.gz   bgzf compressed flat floats. With similar interpretation as version0. Each element is a cdatatype 'float' which is 4 bytes.

PREFIX.saf.pos.gz   bgzf compressed flat integer. Representing the position. Each element is a cdatatype 'int' which is 4bytes

PREFIX.saf.idx   uncompressed binary file containing blocks of data described in 1.2. This is preceeded by a size_t value which indicates the number of categories in the spectrum.

Note that it is not possible to deduce the number of sites directly from the compressed files.

Col	Field	Type	Brief description
1	CLEN	size_t	Length of CHR (not including terminating null)
2	CHR	char*	Reference sequence name. Length is CLEN
3	NSITES	size_t	Number of sites with coverage from reference CHR
4	OFF1	long int	CHR offset into the PREFIX.saf.pos.gz
5	OFF2	long int	CHR offset into the PREFIX.saf.gz

Table 1: Content of entry for a single reference name in the PREFIX.saf.idx file.

2 fst formats

This section describes the binary output generated by a **realSFS fst index pop1.saf.idx pop2.saf.idx -sfs prior**

2.1 fstv1

First iteration of the fst files contains two files. 1) PREFIX.fst.idx 2) PREFIX.fst.gz. First 8bytes is a magic number determining which binary version.

PREFIX.fst.idx flat file, described in table 2.1.1

PREFIX.fst.gz bgzf compressed binary file.

2.1.1 PREFIX.fst.idx

The fst.idx has a very simple header 8bytes magicheader followed by a `size_t` containing the number of samples for which we have generated fst results.

Col	Field	Type	Brief description
1	CLEN	size_t	Length of CHR (not including terminating null)
2	CHR	char*	Reference sequence name. Length is CLEN
3	NSITES	size_t	Number of sites with coverage from reference CHR
4	OFF1	long int	CHR offset into the PREFIX.saf.pos.gz

2.1.2 PREFIX.fst.gz

Col	Field	Type	Brief description
1	POSI	int	Length of CHR (not including terminating null)
2	acoef1	char*	Reference sequence name. Length is CLEN
3	bcoef2	size_t	Number of sites with coverage from reference CHR

2.2 fstv2

PREFIX.fst.idx flat file, described in table 2.1.1

PREFIX.fst.gz bgzf compressed binary file.

2.2.1 PREFIX.fst.idx

2.2.2 PREFIX.fst.gz

FiXme Fatal:
This is still under
development

Col	Field	Type	Brief description
1	MAGIC	char[8]	magic nr = fstv2
2	NPOP	size_t	Number of populations
3	NPOP_LEN_i	size_t	length of population name_i
4	NPOP_i	char[NPOP_LEN]	name of population_i.
5	NDIM_j	size_t	number of categories for pair j
6	PRIOR_j	double[NDIM_j]	prior for pair j
6	AS_j	double[NDIM_j]	alfa for pair j
6	BS_j	double[NDIM_j]	beta for pair j
7	CLEN	size_t	Length of CHR (not including terminating null)
8	CHR	char*	Reference sequence name. Length is CLEN
9	NSITES	size_t	Number of sites with coverage from reference CHR
10	OFF1	long int	CHR offset into the PREFIX.saf.pos.gz

Table 2: Format of saf.idx file. i loops over number of populations. j loops over number of combinations, in this order pop1/pop2,pop1/pop3,pop1/pop4,pop2/pop3,….

Col	Field	Type	Brief description
1	POSI	int	Length of CHR (not including terminating null)
2	acoef1	char*	Reference sequence name. Length is CLEN
3	bcoef2	size_t	Number of sites with coverage from reference CHR