ANGSD formats

tsk

July 11, 2020

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. 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.

1.3 Version 2

This section describes the banded representation of the sample allele frequency likelihoods. First 8 bytes in all three files is 8byte magic numer *char[8] "safv4"*.

PREFIX.saf.gz bgzf compressed. Full description in Table 3.

PREFIX.saf.pos.gz bgzf compressed flat integer. Representing the position. Each element is a cdatatype 'int' which is 4bytes. Similar to version 1.

PREFIX.saf.idx uncompressed binary file containing blocks of data described in 1.3.

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
3	SUMBAND	size_t	Sum of bins 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 2: Content of entry for a single reference name in the PREFIX.saf.idx file.

Col	Field	Type	Brief description
1	FIRST	int32_t	First category with data)
2	NCATS	int32_t	Number of categories with sample allele frequencies
3	SAFLH	float4_t[NCATS]	The actual sample allele frequencies

Table 3: Record for the sample allele frequencies for a single site. Notice that the SAFLH are loglikelihood ratios to the most likely. Scaling is natural log.

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

Table 4: Content of the PREFIX.fst.idx

2.1.2 PREFIX.fst.gz

Col	Field	Type	Brief description
1	POSI	int	Length of CHR (not including terminating null)
2	acoef1	double*	α coefficients from either reynolds estimator or Bhatia
3	bcoef2	double*	β coeffficients from eithre Reynolds estimator or Bhatia

Table 5: Contents of the PREFIX.fst.gz file

3 theta formats

From 0.917 onwards, the -doThetas in angsd wont generate the old ASCII files but rather the indexed file as described below. The original format will not be described in this document.

3.1 thetav2

Second iteration of the theta files now contains one raw bgzf compressed data file and an uncompressed index file. First 8 bytes in the (uncompressed) files are 8byte magic numer *char[8] "thetav2"*. These are generated if the options -dosaf 1 and -doThetas 1 has been selected. This will output the following two files:

prefix.thetas.idx Small uncompressed binary file that contains chr,number of sites, number of chromosomes and the offset into the main data file contain the theta estimates. See below

prefix.thetas.gz Main file. Does also contain chr, number of sites number of chromsomes.

3.1.1 Theta definitions

Let η be the site frequency spectra. Then η_i is the posterior probablity of being in frequency i.

Watterson $\sum_{i=1}^{n-1} \eta_i/a^{-1}, a = \sum_{i=1}^{n-1} i$

π $\binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-1)\eta_i$

FuLi η_1

FayH $\binom{n}{2}^{-1} \sum_{i=1}^{n-1} i^2\eta_i$

L $n-1^{-1} \sum_{i=1}^{n-1} i\eta_i$

3.1.2 Description of binary files

Col	Field	Type	Brief description
1	CLEN	size_t	Length of CHR (inferred by strlen)
2	CHR	char*	Reference sequence name. Length is CLEN
3	NSITES	size_t	Number of sites with coverage from reference CHR
4	NCHR	size_t	number of possible derived/minor allels. (2*nInd for the unfolded, nInd for the folded)
5	OFF	long int	CHR offset into the thetas.gz

Table 6: Content of entry for a single reference name in the PREFIX.thetas.idx file. Note that there exists a 8byte magicnumber in the beginning of the file.

Col	Field	Type	Brief description
1	CLEN	size_t	Length of CHR (inferred by strlen)
2	CHR	char*	Reference sequence name. Length is CLEN
3	NSITES	size_t	Number of sites with coverage from reference CHR
4	NCHR	size_t	number of possible derived/minor allels. (2*nInd for the unfolded, nInd for the fold
5	POSI	int[NSITES]	zero indexed positions for CHR
5	Watterson	float[NSITES]	logscaled persite estimates of Watterson theta estimator (number of segregating site
5	π	float[NSITES]	logscaled persite estimates of the Tajima theta estimator (pairwise differences)
5	FuLi	float[NSITES]	logscaled persite estimates of the fuli theta estmator (singleton category)
5	FayH	float[NSITES]	logscaled persite estimates of the FayH theta estimator
L	L	float[NSITES]	logscaled persite estimates of the L theta estimator

Table 7: Content of the PREFIX.thetas.gz file. Note that there exists a 8byte magicnumber in the beginning of the file.