ANGSD formats

tsk

May 4, 2015

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.

PREFIX.saf.gz bgzf compressed flat floats. With similar interpretation as version0.

PREFIX.saf.pos.gz bgzf compressed flat integer. Representing the position.

PREFIX.saf.idx uncompressed binary file containing blocks of data described in 1.2.

First 8 bytes in all three files is 8byte magic numer *char[8] "safv3"*. The next size_t value is the number of categories of the sample allelefrequency.

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.