

File Extensions (version 0.1)

farjoun@broadinstitute.org

5 Sep 2016

The master version of this document can be found at <https://github.com/samtools/hts-specs>.
This printing is version a5b2693 from that repository, last modified on the date shown above.

license: Apache 2.0

1 Overview

This document's purpose is to collect all the file extensions that are supported/recommended for SAM-related files.

The currently supported index formats are listed in table 1. The table is organized in blocks, where the first line in each block describes the file that might be indexed, and the remaining lines describe the possible index formats & file extensions for said file. The order of the index suffixes is the recommended order by which an implementation will search for an index file.

2 Notes

This list exposes a few weaknesses:

1. We do not have a definitive format specified for tribble (e.g. VCF), BGZIP, FASTA, FAI, bgzipped fasta index (does that even exist?), or FASTQ
2. We have several index formats for the same kind of data: tribble and fai seem to be very similar. bcf.idx and bai are the same (really?) but do not use the same extension
3. Index formats and file formats are rarely written in the same document.

File Type	Algorithm	File extension	Specified in
SAM file	plain	.sam	SAM
CRAM file	cram	.cram	CRAM
CRAM index	cram index	.cram.crai	CRAM+SAM
BAM index	bam index	.cram.bai	CRAM+BAM
CRAM index	cram index	.crai	CRAM
BAM file	bgzip	.bam	SAM
BAM index	bam index	.bai	SAM
BAM index	csi	.csi	CSI
VCF file	plain	.vcf	VCF
VCF index	tribble	.vcf.idx	CODE ONLY
gzipped VCF file	bgzip	.vcf.gz	CODE ONLY
gzipped VCF tabix index	tabix	.vcf.gz.tbi	tabix
BCF file (compressed or not)	bcf or bcf+bgzip	.bcf	BCF
BCF tabix index	bam index	.bcf.idx	SAM
FASTA file	plain	.fasta	https://blast.ncbi.nlm.nih.gov/Blast.cgi https://blast.ncbi.nlm.nih.gov/Blast.cgi CODE ONLY
FASTA file	plain	.fa	
FASTA index	fai	.fai	
gzipped FASTA file	bgzip	.fasta.gz	CODE ONLY
gzipped FASTA index	???	.fasta.gz.gzi	???
gzipped FASTA file	bgzip	.fa.gz	CODE ONLY
gzipped FASTA index	???	.fa.gz.gzi	???
FASTQ	???	.fastq	"crappy paper"
FASTQ	???	.fq	"crappy paper"

Table 1: Supported index file names in the recommended order they be searched