Samtools

John McCallum
Marcus Davy
Samantha Baldwin

SAMtools

SOURCEFORGE.NET®

SAM Tools provides command line tools for manipulating alignments in the SAM format.

Binary BAM file support for efficient storage

Including sorting, merging, indexing and generating alignments in a per-position format.

Open source cross platform project on sourceforge

- *http://samtools.sourceforge.net/*

Li H., Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]*

De facto standard for alignment data

Aligners natively generating SAM

- BFAST, `Blat-like Fast Accurate Search Tool' for Illumina and SOLiD reads.
- Bowtie. Highly efficient short read aligner. Natively support SAM output in recent version. A convertor is also available in samtools-C.
- BWA, Burrows-Wheeler Aligner for short and long reads.
- GEM library. Short read aligner. Convertor provided by the developers.
- Karma, the K-tuple Alignment with Rapid Matching Algorithm.
- LASTZ, aligner for both short and long reads.
- Mosaik. The latest version support SAM output.
- Novoalign. An accurate aligner capable of gapped alignment for Illumina short reads. Academic free binary. Convertor is also available in samtools.
- SNP-o-matic, short read aligner and SNP caller.
- SOLiD BaseQV Tool. Developed by Applied Biosystems for converting SOLiD output files.
- SSAHA2 (since v2.4). Classical aligner for both short and long reads.
- Stampy, by Gerton Lunter. An accurate read aligner capable of gapped alignment for Illumina short reads. Used for indel discovery on the 1000 genomes data.
- TopHat for mapping short RNA-seq reads bridging exon junctions.

http://samtools.sourceforge.net/swlist.shtml

De facto standard for alignment data…

Programs processing SAM/BAM

- BAMTools, C++ APIs (not based on C APIs) for processing BAM files.
- BamView, BAM alignment viewer. It can be integrated to Artemis.
- BEDTools, a software package for manipulating BED files, with some utilities working with BAM. Built upon BAMTools.
- BreakDancer, structural variation caller for paired-end data.
- DNAA, DNA Analysis package including various post-alignment processing.
- Gambit, graphical BAM alignment viewer.
- GAP5, sequence assembly viewer, editor and analyzer. Capable of importing BAM files and outputing SAM.
- GATK, the Genome Analysis Toolkit. Rich funtionality including an accurate SNP caller. Built upon Picard.
- GBrowse, generic genome browser. Experimental SAM/BAM alignment viewing. Built upon Perl APIs.
- GenomeView, a Java based genome browser.
- IGB, the Integrated Genome Browser for various data formats.
- IGV, the Integrative Genomics Viewer, supporting multiple tracks and genome annotations. Built upon Picard.
- LookSeq, web-based alignment/annotation viewer.
- MagicViewer, graphical BAM alignment viewer.
- samToBed by Aaron Quinlan. Converting alignments in the SAM format to the BED format.
- Savant, a Java based genome browser.
- Tablet, alignment viewer. It also supports tons of other alignment/assembly formats.
- Vancouver Short Read Analysis Package (in particular FindPeaks), post alignment processing of new sequencing data.
- VarScan, variant caller for short sequence reads.

http://samtools.sourceforge.net/swlist.shtml

Header fields

Lines begin with character '**@**' followed by a two-letter record type code.

TAB delimited lookup table of the form ***TAG:TYPE:VALUE***

Tags with ***'*'*** are required when the ***TYPE*** is present

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
SQ - Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
RG - read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
	PL	Platform/technology used to produce the read.
PG - Program	ID*	Program name
	VN	Program version
	CL	Command line
CO - comment		One-line text comments

e.g.

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
```

Alignment fields

TAB delimited fields in the body of the format

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

http://samtools.sourceforge.net/SAM1.pdf

Flag field

Binary lookup table

Bitwise flag allows filtering using *samtools view*

e.g.
Unaligned reads = 4

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

Decimal to binary conversion;

SAMtool bitwise flag meaning explained

Picard flag conversion tool

The cigar string

The aligned sequence or reference may have additional/missing bases *(INDELS)*
The CIGAR string is a sequence of base lengths and the associated operation to indicate which;

- bases align (either a ***match/mismatch***) with the reference
- bases are ***deletions*** from the reference
- ***insertions*** that are not in the reference

e.g. 20M6I10M = 20 matches, 6 reference inserts, 10 matches

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

SAMtools documentaion

Usage: samtools <command> [options]

Type ***samtools*** in the ***shell terminal*** command prompt for help

```
$  samtools

Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.18 (r982:295)

Usage:   samtools <command> [options]

Command: view         SAM<->BAM conversion
         sort         sort alignment file
         mpileup      multi-way pileup
         depth        compute the depth
         faidx        index/extract FASTA
         tview        text alignment viewer
          ...                  ...
```

Manual pages: ***http://samtools.sourceforge.net/samtools.shtml***

Rsamtools

- Rsamtools is a Bioconductor package

http://www.bioconductor.org/packages/2.10/bioc/html/Rsamtools.html

- *The package provides an interface for R to efficiently access* and import BAM les into R

- Facility for file access such as record counting, index file creation, and filtering to create new files containing subsets of the original.

- Rsamtools is as a starting point for creating R objects suitable for a diversity of workfows, e.g. ***AlignedRead*** objects in the ***ShortRead*** package

Help documentation (vignette) in R;

```
library(Rsamtools)
vignette("Rsamtools-Overview")
```