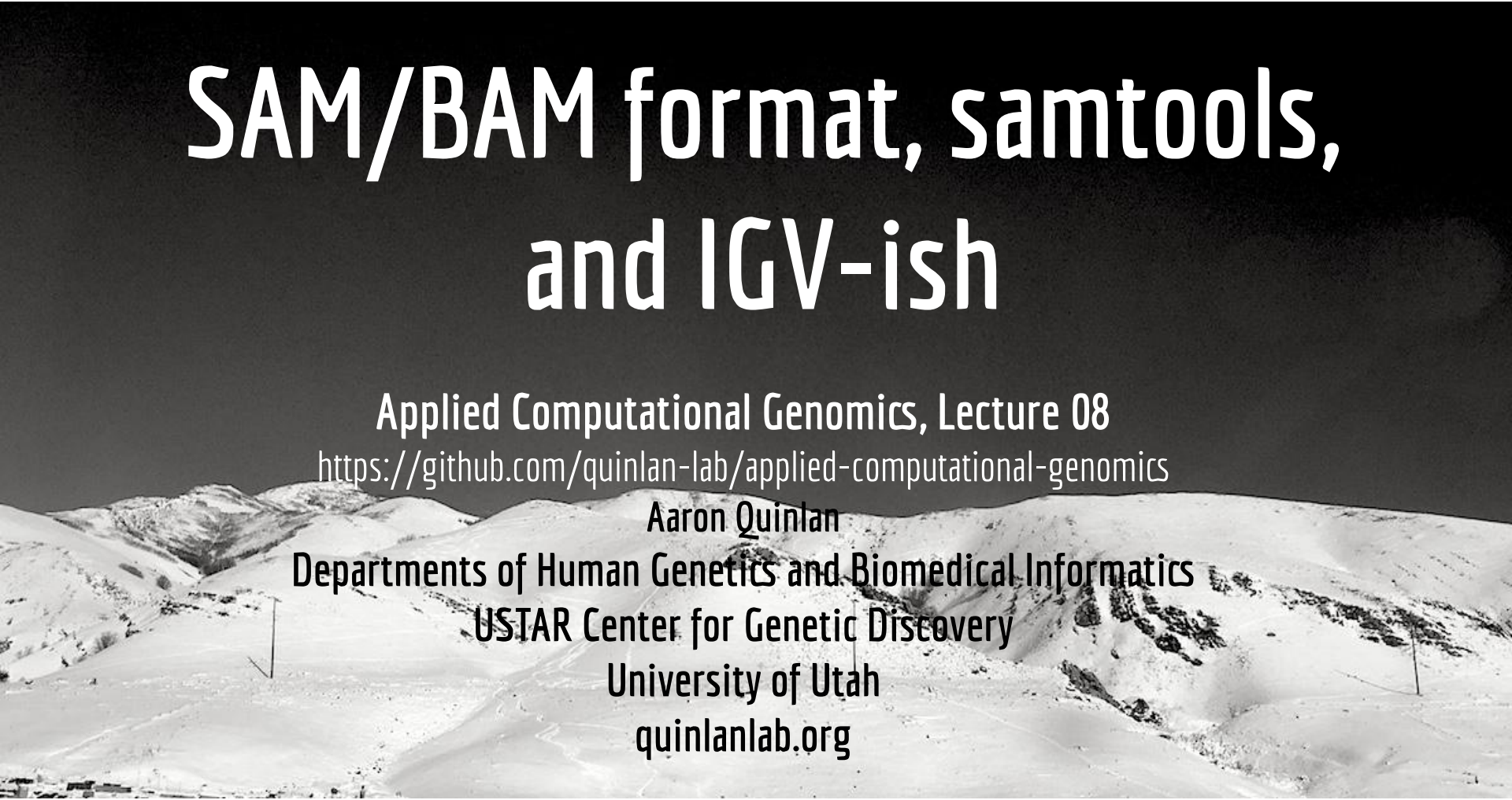
# SAM/BAM format, samtools, and IGV-ish



Applied Computational Genomics, Lecture 08

https://github.com/quinlan-lab/applied-computational-genomics

Aaron Quinlan

Departments of Human Genetics and Biomedical Informatics

USTAR Center for Genetic Discovery

University of Utah

quinlanlab.org

# SAM format: a **text**-based **standard(!)** for representing sequence alignments

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 25 no. 16 2009, pages 2078–2079*
*doi:10.1093/bioinformatics/btp352*

*Sequence analysis*

**The Sequence Alignment/Map format and SAMtools**

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

Advance Access publication June 8, 2009

Associate Editor: Alfonso Valencia

**Table 1.** Mandatory fields in the SAM format

| No. | Name | Description |
|---|---|---|
| 1 | `QNAME` | Query NAME of the read or the read pair |
| 2 | `FLAG` | Bitwise FLAG (pairing, strand, mate strand, etc.) |
| 3 | `RNAME` | Reference sequence NAME |
| 4 | `POS` | 1-Based leftmost POSition of clipped alignment |
| 5 | `MAPQ` | MAPping Quality (Phred-scaled) |
| 6 | `CIGAR` | Extended CIGAR string (operations: `MIDNSHP`) |
| 7 | `MRNM` | Mate Reference NaMe ('=' if same as `RNAME`) |
| 8 | `MPOS` | 1-Based leftmost Mate POSition |
| 9 | `ISIZE` | Inferred Insert SIZE |
| 10 | `SEQ` | Query SEQuence on the same strand as the reference |
| 11 | `QUAL` | Query QUALity (ASCII-33=Phred base quality) |

# SAM format overview

- **In the dark ages,** sequence aligners used disparate output formats. **Pain.**
- 1000 Genomes Project sought to standardize. Standards are good.
- The result is imperfect, but it's a **huge** improvement.
- Strengths of the SAM and BAM formats
    - Compressed: less disk hungry
    - Indexed: fast viewing, slicing, etc.
    - Single-end and paired-end
    - Relatively simple to produce
    - *Good toolkits available*

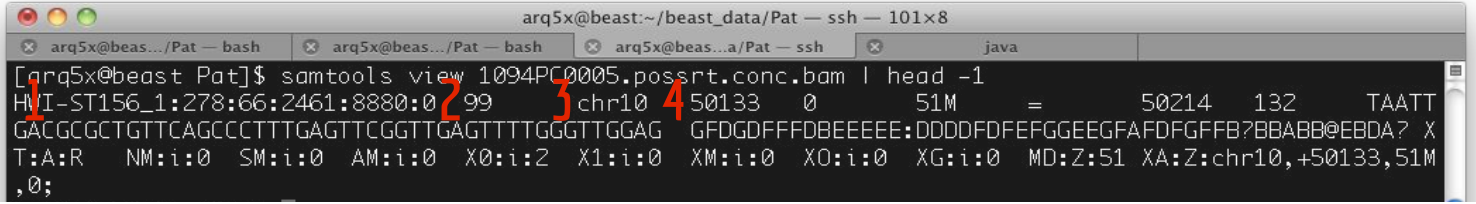
# What critical information do we need for sequence alignments?

# SAM format overview

| Col # | Name | Meaning | Example |
|---|---|---|---|
| 1 | QNAME | Read or Pair name | HWI:ST156_1:278:1:1058:4544:0 |
| 2 | FLAG | Bitwise FLAG | ***Much more soon!*** |
| 3 | RNAME | Reference sequence name | chr1 |
| 4 | POS | 1-based alignment start coordinate | 8,724,005 |
| 5 | MAPQ | Mapping quality | 60 |
| 6 | CIGAR | Extended CIGAR string | ***Much more soon!*** |
| 7 | MRNM | If paired, the mate's reference seq. | chr1 |
| 8 | MPOS | If paired, the mate's alignment start | 8,724,505 |
| 9 | ISIZE | If paired, the insert size | 562 |
| 10 | SEQ | The sequence of the query/mate | ACAAATTCAG... |
| 11 | QUAL | The quality string for the query/mate | HHH$^^%$$$... |
| 12 | OPT | Optional Tags | XA:i:2, MD:Z:0T34G15 |

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

| Col # | Name | Meaning | Example |
|---|---|---|---|
| 1 | QNAME | Read or Pair name | HWI:ST156_1:278:1:1058:4544:0 |
| 2 | FLAG | Bitwise FLAG | ***Much more soon!*** |
| 3 | RNAME | Reference sequence name | chr1 |
| 4 | POS | 1-based alignment start coordinate | 8,724,005 |
| 5 | MAPQ | Mapping quality | 60 |
| 6 | CIGAR | Extended CIGAR string | ***Much more soon!*** |
| 7 | MRNM | If paired, the mate's reference seq. | chr1 |
| 8 | MPOS | If paired, the mate's alignment start | 8,724,505 |
| 9 | ISIZE | If paired, the insert size | 562 |
| 10 | SEQ | The sequence of the query/mate | ACAAATTCAG... |
| 11 | QUAL | The quality string for the query/mate | HHH$^^%$$$... |
| 12 | OPT | Optional Tags | XA:i:2, MD:Z:0T34G15 |



```
ST-E00223:32:H5J57CCXX:6:2123:15189:52872       97      1       10001   0       4S15M1I54M2I50M25S      4       699063  0
ACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCC
CCCCCACCCAACCCCACCCCCCAC
AAFFFKKKKKKKKKKKKKKKKKKKFKKFKKKKKKFKKAKKKKFFKKKKKFKF<FF7FFFFK7FK,AA,FKFFKF,FKKK7<KKK,,7FFF7F,AAFFK,AKA,,FFF<,A7FKK,,,7A,,,7,,,
,,((,(,,(,,,,,,,(,,A(,,( MC:Z:80S11M1D58M        MD:Z:119        RG:Z:15-0017315_1       NM:i:3  MQ:i:47 AS:i:104
XS:i:103
ST-E00223:46:HG7V5CCXX:2:1116:12601:22862       1123    1       10006   0       81M70S  =       10106   143
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACACTCACCCTAACCCTACCCCTAACCCTAACCCTATCATTCACTCGAACCCTAACACTACCGCTAGCGCTAACTCCAGCCC
GCACACTATCGCTAACCCTCACGC
AAFFFKKKKKKKKKKKKKKKFKKA,,A,A,,77<,,A,,7FFK7,,,,AF,,,A7,,,77,,,7AA,,7,FFK,<A,,,,7<<,,,,AA,7,AA,,7,,,7,,,,,,,,,,A7,,,,,,7,7,7(,
(,((,,,7,A,F,7,,,<,,AA,, MC:Z:108S43M    MD:Z:47C2A12A17 RG:Z:15-0017315_1       NM:i:3  MQ:i:2  AS:i:66 XS:i:71
ST-E00223:32:H5J57CCXX:5:2208:10074:43308       99      1       10008   36      101M1I41M7S     =       10107   137
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTACCCCTAACCCTAACCCTAACCCTAACCCTCACCCTCACCCTCACCCTAGCCCAAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTACCCCG
AAFFFKFKKKKKKKKKKKKKKKKFKKKKKKKFKKK7<KA7<F,AF7F,7,A,FA7<KKKKKFKKKK<,,FKAKKAFFA,A7,7,,,7,7,,,,,7F<,,7,,,,7,,A,,,<F7,,,7FA,7,,<F,
,,<A,7<AFKK<,<,,7,,,,,(  MC:Z:112S38M    MD:Z:49A28A5A5A6A44     RG:Z:15-0017315_1       NM:i:6  MQ:i:36 AS:i:110
XS:i:113
ST-E00223:46:HG7V5CCXX:5:2119:12936:64896       99      1       10013   0       90M61S  =       10176   211
TAACCCTAAGCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCATAACCCAAACCCTAACCCTAACCCTAACCCGAACCGTAAGCCAAAACATAACCACAACCATAACAAT
AACCAAAACCTTAACGTTAAACAT
A<AFFKKKK,AAAFKK<FKKKKAK,<A,AFKKKKAFFFKKKKAFKKFAFFFKKA7,FFA,F,F,7F,AFF77AFFAFFKAFKKA,,FFKF,AA,F,,7F,A,,7A,,,7A,,,,,AFK,,,AA,,,
7FKA,A,AFF,<,,,,,,,77<AA MC:Z:99S48M4S   MD:Z:9C49C6T23  RG:Z:15-0017315_1       NM:i:3  MQ:i:0  AS:i:75 XS:i:75
ST-E00223:32:H5J57CCXX:1:1205:17290:54577       99      1       10019   1       92M59S  =       10354   414
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAAACCCCCCACCCCAAACCCAAGCCCAACCCTACCCCC
TAACCCCTAACCCCAACCCTGACC
AAFFFKKKKFKKKKAKKKKKKKKKKKKKKKF7KKKKKKKKKKKKKKFKKKKFKKKKKKKAKKKFFKKFFKKKK,<,7<FA,FFFK,,7FFF,,,,,A,((,(((,7,7A,,,,,A,,,777A,AAKK<
<,7FFA7,A,,AA,<,,AFA,7AF MC:Z:72S79M     MD:Z:92 RG:Z:15-0017315_1       NM:i:0  MQ:i:20 AS:i:92 XS:i:97
```

# Recall: Edit distance

How many edits (changes) must be made to a word or kmer to make it match (align) to another word or kmer?

CURLED
HURLED → Edit distance = 1. Substitute C for H

SHORT
SHO-T → Edit distance = 1. Delete R

TGTTACGG
GGTTGACTA
?

TG-TT-ACGG
-GGTTGACTA

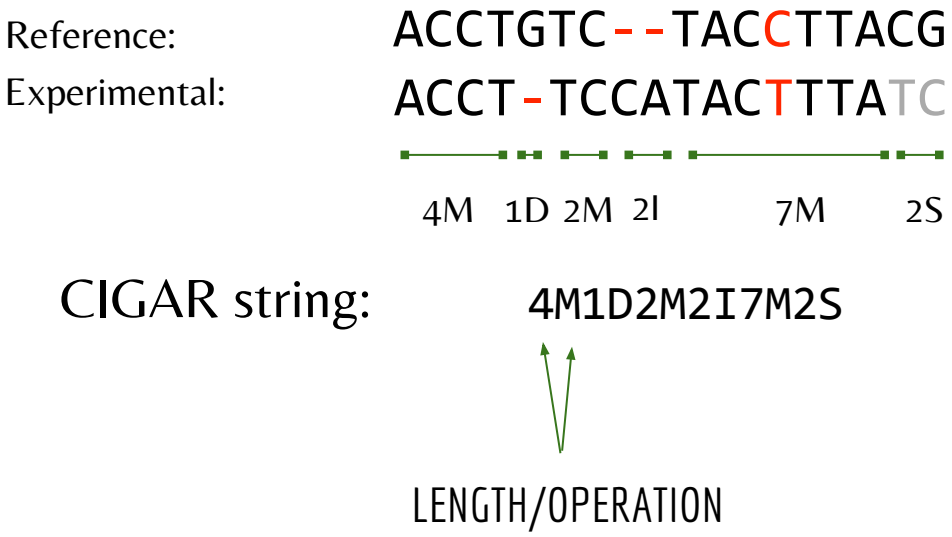
Edit distance = 5

TGTT-ACGG
GGTTGACTA

Edit distance = 4

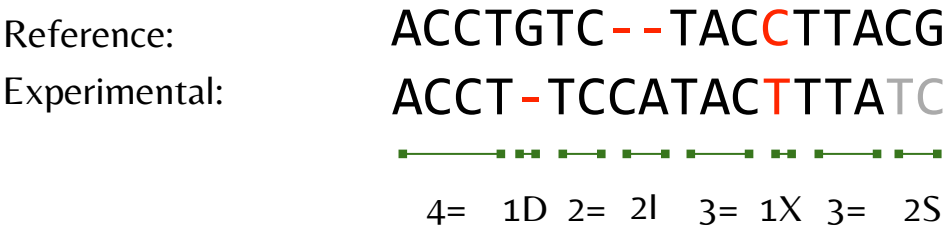
# The CIGAR string: encode the details of the alignment

| Operation | Meaning |
|---|---|
| M | Match* |
| D | Deletion w.r.t. reference |
| I | Insertion w.r.t. reference |
| N | Split or spliced alignment |
| S | Soft-clipping |
| H | Hard-clipping |
| P | Padding |



# The extended CIGAR string: M become = and X

| Operation | Meaning |
|---|---|
| = | Exact match |
| X | Mismatch |
| D | Deletion w.r.t. reference |
| I | Insertion w.r.t. reference |
| N | Split or spliced alignment |
| S | Soft-clipping |
| H | Hard-clipping |
| P | Padding |




CIGAR string: 4=1D2=2I3=1X3=2S

# The FLAG column

| Sequence ID | FLAG | CHROM | POS |
|---|---|---|---|
| ST-E00223:32:H5J57CCXX:6:2123:15189:52872 | 97 | 1 | 10001 |
| ST-E00223:46:HG7V5CCXX:2:1116:12601:22862 | 1123 | 1 | 10006 |
| ST-E00223:32:H5J57CCXX:5:2208:10074:43308 | 99 | 1 | 10008 |
| ST-E00223:46:HG7V5CCXX:5:2119:12936:64896 | 99 | 1 | 10013 |
| ST-E00223:32:H5J57CCXX:1:1205:17290:54577 | 99 | 1 | 10019 |
| ST-E00223:32:H5J57CCXX:6:1115:16844:11013 | 81 | 1 | 10026 |
| ST-E00223:32:H5J57CCXX:7:2113:18935:32356 | 99 | 1 | 10032 |
| ST-E00223:46:HG7V5CCXX:6:2117:3082:44239 | 99 | 1 | 10040 |
| ST-E00223:46:HG7V5CCXX:5:2213:10744:58813 | 163 | 1 | 10074 |
| ST-E00223:32:H5J57CCXX:4:1220:14651:8868 | 99 | 1 | 10086 |

# The FLAG score

| base2 | base10 | base16 | Meaning | Applies to: |
|---|---|---|---|---|
| 00000000001 | 1 | 0x0001 | The read originated from a paired sequencing molecule | Both |
| 00000000010 | 2 | 0x0002 | The read is mapped in a **proper** pair | Pairs only |
| 00000000100 | 4 | 0x0004 | The query sequence itself is unmapped | Both |
| 00000001000 | 8 | 0x0008 | The query's mate is unmapped | Pairs only |
| 00000010000 | 16 | 0x0010 | Strand of the query (0 for forward; 1 for reverse strand) | Both |
| 00000100000 | 32 | 0x0020 | Strand of the query's mate | Pairs only |
| 00001000000 | 64 | 0x0040 | The query is the first read in the pair | Pairs only |
| 00010000000 | 128 | 0x0080 | The read is the second read in the pair | Pairs only |
| 00100000000 | 256 | 0x0100 | The alignment is not primary | Both |
| 01000000000 | 512 | 0x0200 | The read fails platform/vendor quality checks | Both |
| 10000000000 | 1024 | 0x0400 | The read is either a PCR duplicate or an optical duplicate | Both |

ST-E00223:32:H5J57CCXX:4:1220:14651:8868 **99** 1 10086

| base2 | base10 | base16 | Meaning | Applies to: |
|---|---|---|---|---|
| 00000000001 | 1 | 0x0001 | The read originated from a paired sequencing molecule | Both |
| 00000000010 | 2 | 0x0002 | The read is mapped in a **proper** pair | Pairs only |
| 00000000100 | 4 | 0x0004 | The query sequence itself is unmapped | Both |
| 00000001000 | 8 | 0x0008 | The query's mate is unmapped | Pairs only |
| 00000010000 | 16 | 0x0010 | Strand of the query (0 for forward; 1 for reverse strand) | Both |
| 00000100000 | 32 | 0x0020 | Strand of the query's mate | Pairs only |
| 00001000000 | 64 | 0x0040 | The query is the first read in the pair | Pairs only |
| 00010000000 | 128 | 0x0080 | The read is the second read in the pair | Pairs only |
| 00100000000 | 256 | 0x0100 | The alignment is not primary | Both |
| 01000000000 | 512 | 0x0200 | The read fails platform/vendor quality checks | Both |
| 10000000000 | 1024 | 0x0400 | The read is either a PCR duplicate or an optical duplicate | Both |

00001100011

$2^6+2^5+2^1+2^0 = 64+32+2+1 = 99$

*What is the best way to use tons of disk space and have very inefficient analyses?*

**Text files of billions of alignments**

# Use samtools to convert SAM to BAM.

*This takes a long time, but you do it once*

Create BWT of reference genome.

```
$ bwa index grch38.fa
```

***Output is in SAM format.***
*Use multiple threads if you have a computer with multiple CPUs.*

Align paired-end FASTQ to BWT index.

```
$ bwa mem -t 16 grch38.fa 1.fq 2.fq > sample.sam
```

***Output is in BAM format.***
*However, it is unsorted - that is, random genomic order as reads are randomly placed in FASTQ by sequencer.*

Convert SAM to BAM

```
$ samtools view -Sb sample.sam > sample.bam
```

# SAMTOOLS: Converting and manipulating SAM/BAM

```
Commands:
  -- Indexing
     dict           create a sequence dictionary file
     faidx          index/extract FASTA
     index          index alignment

  -- Editing
     calmd          recalculate MD/NM tags and '=' bases
     fixmate        fix mate information
     reheader       replace BAM header
     rmdup          remove PCR duplicates
     targetcut      cut fosmid regions (for fosmid pool only)
     addreplacerg   adds or replaces RG tags

  -- Viewing
     flags          explain BAM flags
     tview          text alignment viewer
     view           SAM<->BAM<->CRAM conversion
     depad          convert padded BAM to unpadded BAM
```

http://www.htslib.org/doc/samtools.html

# SAMTOOLS: Converting and manipulating SAM/BAM

```
Commands:

  -- File operations
     collate         shuffle and group alignments by name
     cat             concatenate BAMs
     merge           merge sorted alignments
     mpileup         multi-way pileup
     sort            sort alignment file
     split           splits a file by read group
     quickcheck      quickly check if SAM/BAM/CRAM file appears intact
     fastq           converts a BAM to a FASTQ
     fasta           converts a BAM to a FASTA

  -- Statistics
     bedcov          read depth per BED region
     depth           compute the depth
     flagstat        simple stats
     idxstats        BAM index stats
     phase           phase heterozygotes
     stats           generate stats (former bamcheck)
```

https://gist.github.com/arq5x/4716b710f967998e9feaeb134e0ebe2b#file-bam-md

# Tutorial for working with samtools

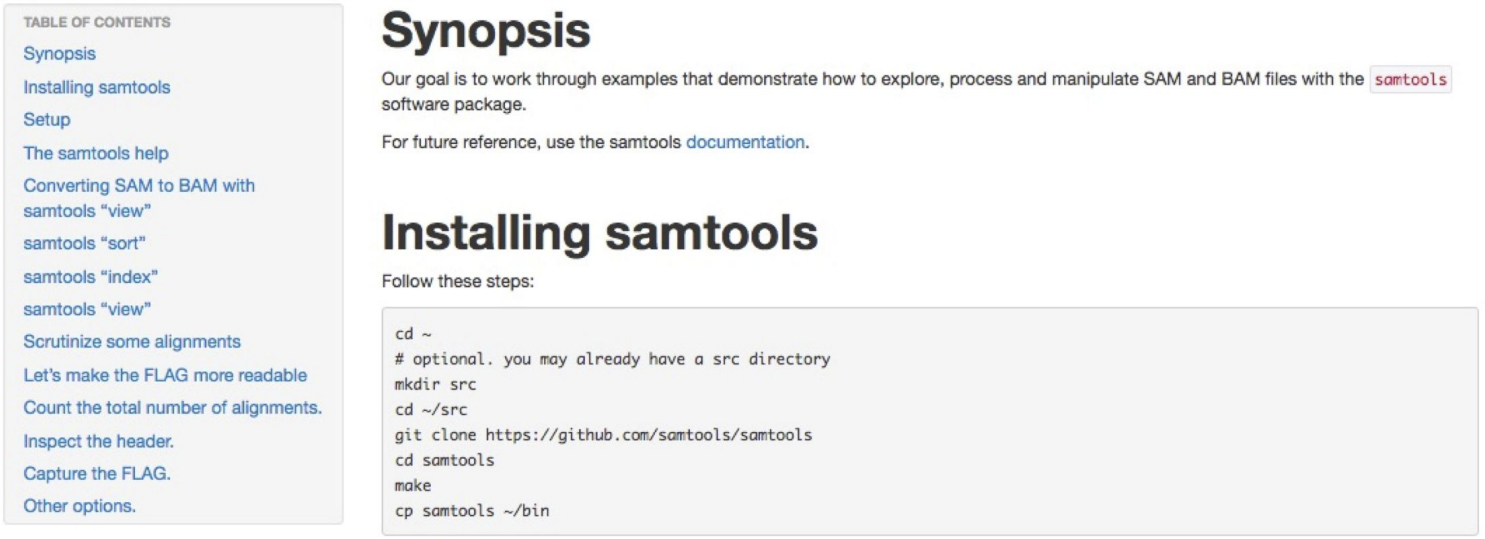


TABLE OF CONTENTS



## Synopsis

Our goal is to work through examples that demonstrate how to explore, process and manipulate SAM and BAM files with the `samtools` software package.

For future reference, use the samtools documentation.

## Installing samtools

Follow these steps:

```
cd ~
# optional. you may already have a src directory
mkdir src
cd ~/src
git clone https://github.com/samtools/samtools
cd samtools
make
cp samtools ~/bin
```

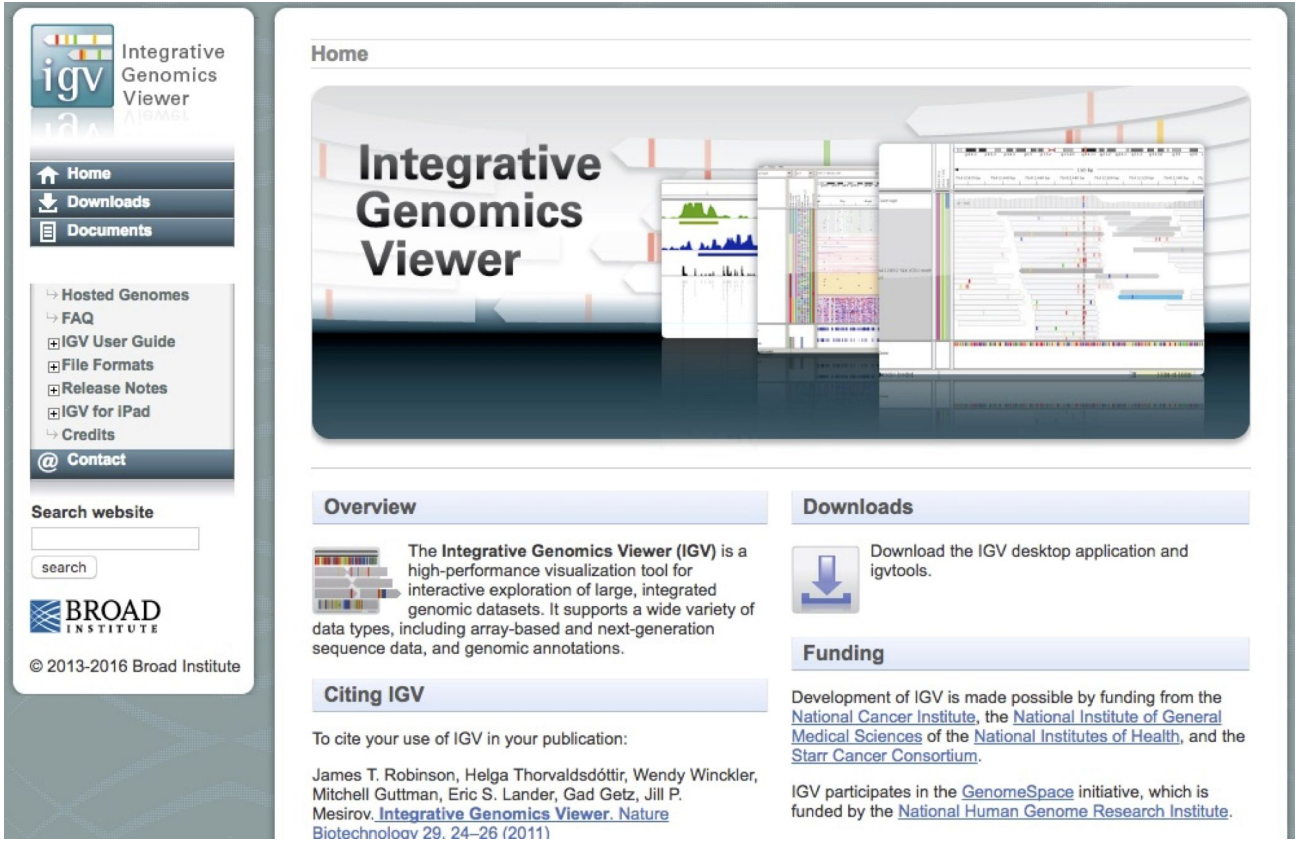
http://quinlanlab.org/tutorials/samtools/samtools.html

# Let’s play around with a real BAM file using samtools

# Use IGV to look at your BAM file

https://gist.github.com/arq5x/4716b710f967998e9feaeb134e0ebe2b#file-igv-md

# IGV tutorial




https://github.com/griffithlab/rnaseq_tutorial/wiki/IGV-Tutorial