III. SAM, BAM AND CRAM FORMAT

SAM Format

- SAM = Sequence Alignment/Map format
- Tab delimited plain text
- Store large nucleotide sequence alignments
  - Alignment of every read
  - Including gaps, SNPs and structural variants
  - Pairing of reads
  - Can record more than one alignment location in the genome
  - Stores quality values
  - Stores information about duplication

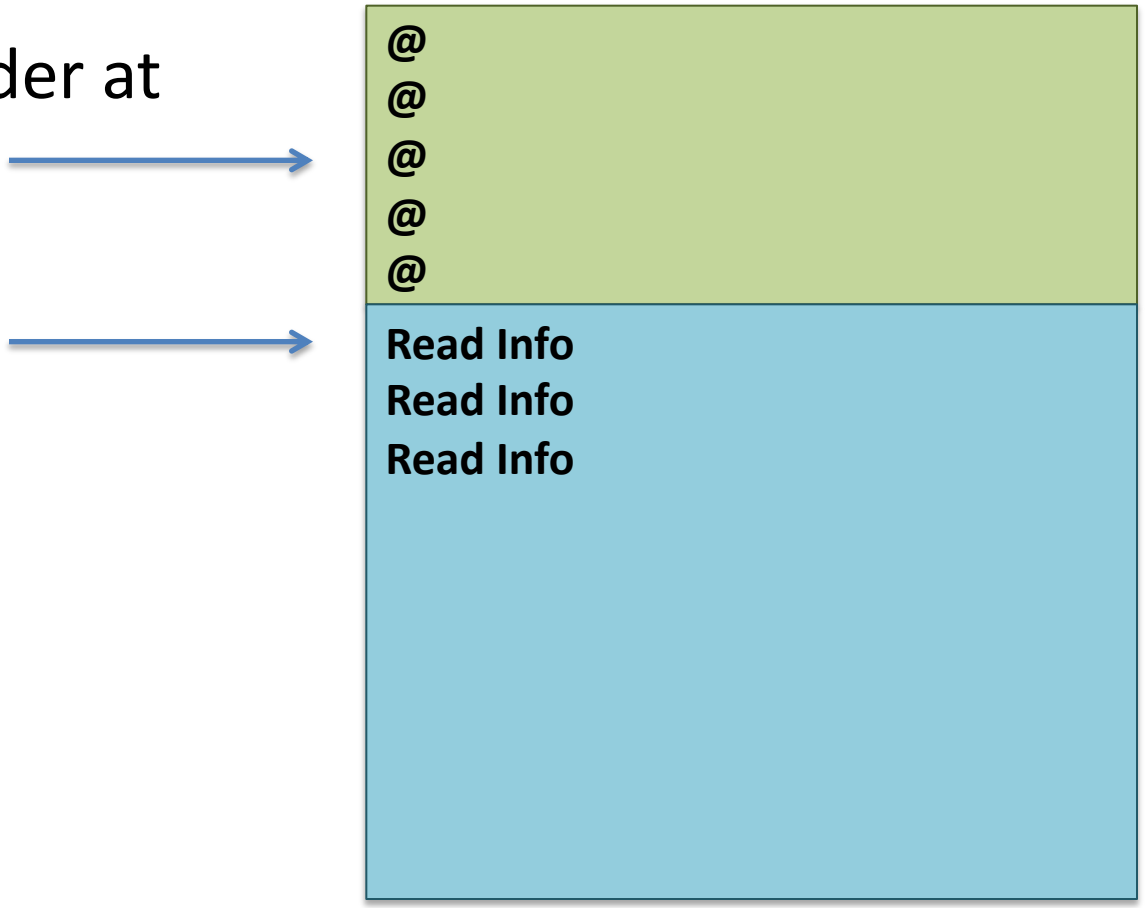
SAM Format

Strengths

- Flexible
- Useful for operations on very large sequences
- Extremely detailed documentation
  - https://github.com/samtools/hts-specs/blob/master/SAMv1.pdf
  - Manipulations can be done with the software samtools

SAM - Header

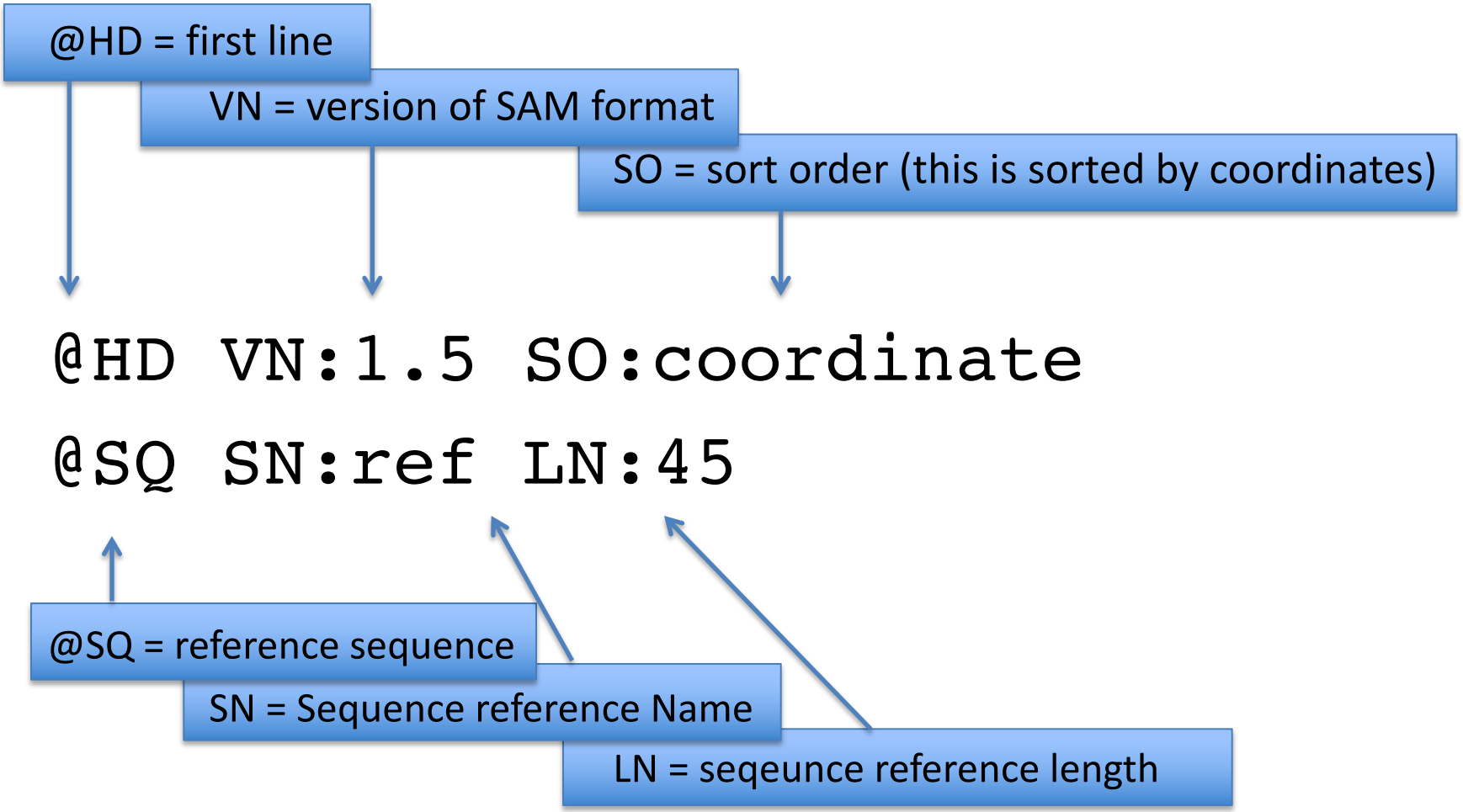
- Structure
  - Optional Header at top of file
  - Alignment information



SAM - Header

- Header lines start with @ symbol
- Always at top of file
- Contain lots of information about what was mapped, what it was mapped to, and how (metadata)
  - the version information for the SAM/BAM file
  - whether or not and how the file is sorted
  - information about the reference sequences
  - any processing that was used to generate the various reads in the file
  - software version

Simple Header



Decipher header information: https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment Line

- Below the headers are the alignment records
- Tab-delimited fields

1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe (`=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Lets unpack this alignment line, taken from a SAM file:

SRR030257.2000020 83 gi|254160123|ref|NC_012967.1| 32957526036M = 3295706-82 TGCTGGCGGCGATATCGTCCGTGGTTCCGATCTGGT ?%<91<?>>??AAAAAAAAAAAAAAAAAAAAAAAAA XT:A: NM:i:0 SM:i:37 AM:i:37 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:36

SAM Field 1

Query name

SRR030257.2000020

Field 2: Flag

83

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

64 + 16 + 2 + 1

1 = Read is paired
2 = Read mapped in proper pair
16 = Read mapped to reverse strand
64 = First in pair

Look up a SAM flag: https://broadinstitute.github.io/picard/explain-flags.html

SAM Field 3

Reference sequence name (useful especially if you have multiple chromosomes)

```
gi|254160123|ref|NC_012967.1|
```

SAM Field 4

Position- 1-based leftmost mapping POSition of the first matching base

```
3295752
```

SAM Field 5

Mapping Quality

- equals $-10 \log_{10}$ Pr{mapping position is wrong}, rounded to the nearest integer
- Same as phred!
- Probability of 99.9% correct = map quality of 30
- Probability of 0% correct = map quality of 0
- value 255 indicates that the mapping quality is not available.

60

.000001% probability wrong

SAM Field 6

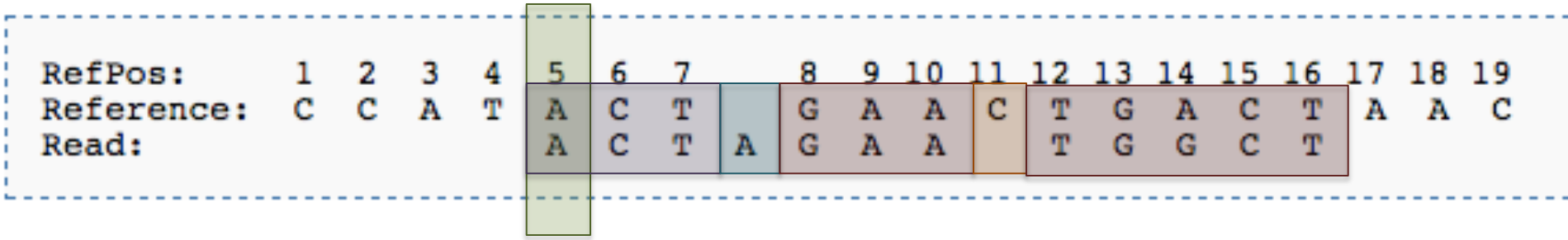
CIGAR String

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

```
36M   36 nucleotides match (perfect match)
8S28M 8 nucleotides clipped, 28 match
```

More CIGAR

Aligning these two:



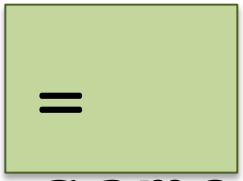
Position: 5

CIGAR: 3M1I3M1D5M

SAM Field 7

Reference sequence for the next read in the template

- For a forward read, this is the reference where the reverse read maps
- For a reverse read, this is the reference where the forward read maps



`=  reverse read maps on the same reference`

SAM Field 8

Position where the next read maps

```
3295706
```

```
(Forward read mapped at 3295752.
Remember the forward read mapped
to the reverse strand)
```

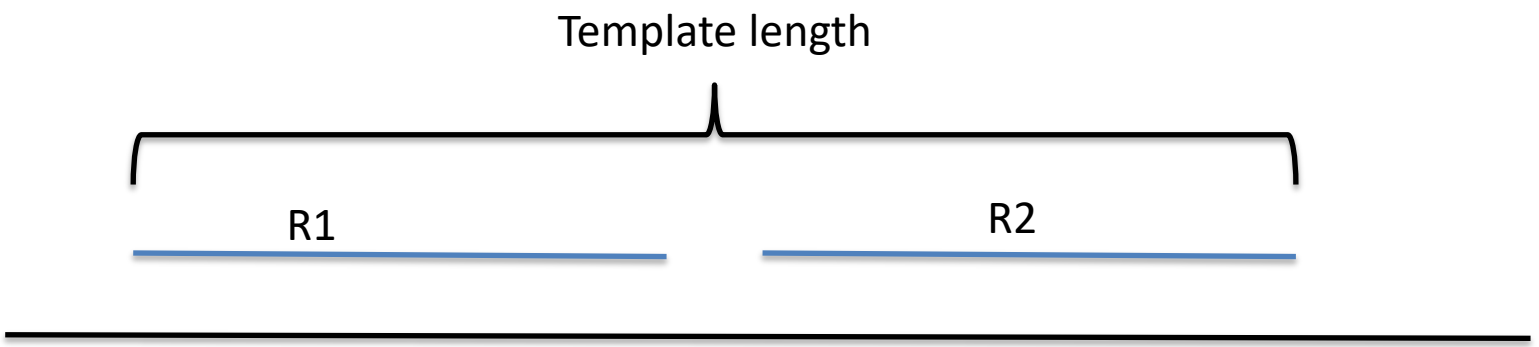
SAM Field 9

Observed template length

From leftmost base to rightmost base

Negative if this read is the rightmost read

-82



SAM Field 10

Sequence of the read

```
TGCTGGCGGCGATATCGTCCGTGGTTCCGATCTGGT
```

SAM Field 11

Quality of the read

```
?%<91<?>>??AAAAAAAAAAAAAAAAAAAAAAAAAA
```

SAM Field 12

Optional MORE information

TAG:TYPE:VALUE format

```
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1
X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:36
```

Anything with an X is specified by the user or by the mapping software, and is not part of the SAM spec.

Decipher the last fields

```
XT:A:U   One of Unique/Repeat/N/Mate-sw
NM:i:0   Edit distance to the reference
SM:i:37  Template-independent mapping quality
AM:i:37  Smallest template-independent mapping
       quality of other segments
X0:i:1   Number of best hits
X1:i:0   Number of suboptimal hits found by BWA
XM:i:0   Number of mismatches in the alignment
XO:i:0   Number of gap opens
XG:i:0   Number of gap extentions
MD:Z:36  String for mismatching positions
```

SAM

Example sam with one read:

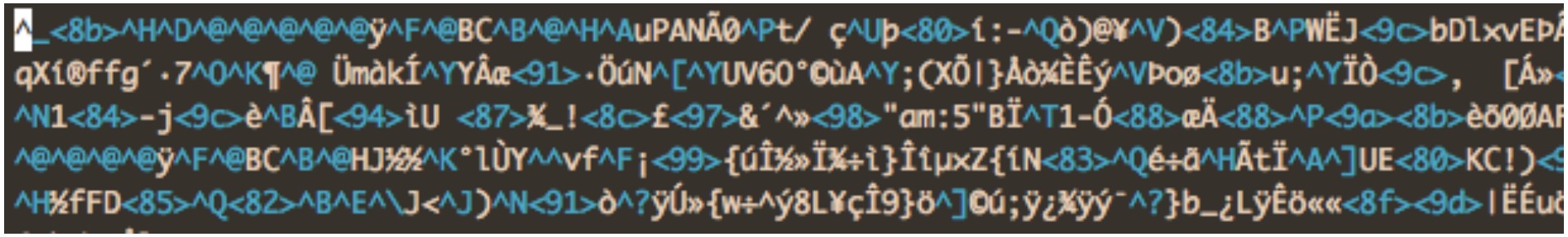
@SQ SN:gi|254160123|ref|NC_012967.1| LN:4629812

@PG ID:bwaPN:bwa VN:0.7.12-r1039
CL:/lustre/projects/rnaseq_ws/apps/bwa-0.7.12/bwa sampe ../raw_data/NC_012967.1.fasta aln_SRR030257_1.sai aln_SRR030257_2.sai ../raw_data/SRR030257_1.fastq ../raw_data/SRR030257_2.fastq

SRR030257.1 99 gi|254160123|ref|NC_012967.1| 950180
60 36M = 950295 151
TTACACTCCTGTTAATCCATACAGCAACAGTATTGG
AAA;A;AA?A?AAAAA?;?A?1A;;????566)=*1 XT:A:U
NM:i:1SM:i:37 AM:i:25 X0:i:1 X1:i:0 XM:i:1 XO:i:0
XG:i:0 MD:Z:32C3

BAM Format

- Sister format to SAM
- BAM – Binary version of SAM
- compressed BGZF (Blocked GNU Zip Format) - a variant of GZIP (GNU ZIP),
- files are bigger than GZIP files, but they are much faster for random access
- Can index and then look up information embedded in the file with decompressing the whole file
- up to 75% smaller in size
- Not readable by people

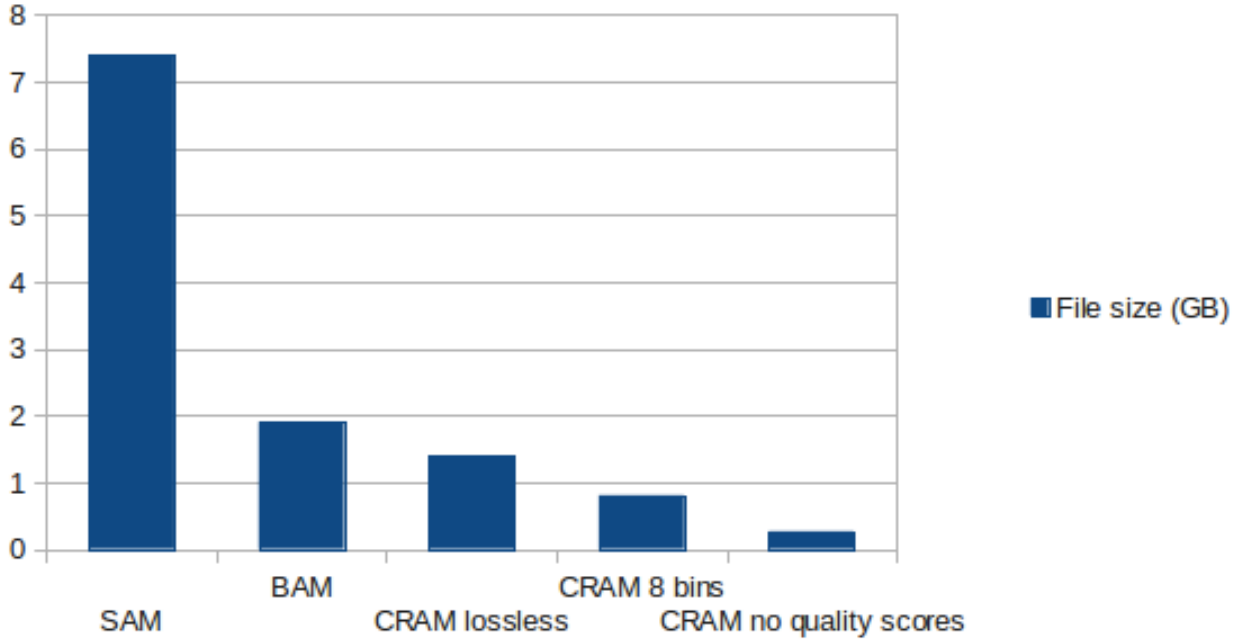


http://blastedbio.blogspot.com/2011/11/bgzf-blocked-bigger-better-gzip.html

CRAM

- Introduced in 2011 by EMBL/EBI
- Even smaller and more efficient than BAM files
- Rare

EBI has a cram toolkit
https://www.ebi.ac.uk/ena/software/cram-toolkit



Fritz, Markus Hsi-Yang, et al. "Efficient storage of high throughput DNA sequencing data using reference-based compression." Genome research 21.5 (2011): 734-740.

http://www.uppmax.uu.se/using-cram-to-compress-bam-files-on-uppnex

Software – htslib.org

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

Samtools	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
BCFtools	Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
HTSlib	A C library for reading/writing high-throughput sequencing data

Samtools and BCFtools both use HTSlib internally, but these source packages contain their own copies of htslib so they can be built independently.

Download

Source code releases can be downloaded from GitHub or Sourceforge:

Source release details

Workflows

We have described some standard workflows using Samtools:

- WGS/WES Mapping to Variant Calls
- Using CRAM within Samtools