

Cold Spring Harbor Laboratory

RNA-Seq Module 2 SAM/BAM/BED file formats

Kelsy Cotto, Felicia Gomez, Obi Griffith, Malachi Griffith, Huiming Xia Advanced Sequencing Technologies & Applications November 5- 16, 2019





Module 2 rnabio.org

Example of SAM/BAM file format

Example SAM/BAM/CRAM header section (abbreviated)

```
mgriffit@linus270 -> samtools view -H /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam | grep -P "SN\:22|HD|RG|PG"
              VN:1.4 SO:coordinate
              SN:22 LN:51304566
                                                           UR:ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa.gz AS:GRCh37-lite M5:a718acaa6135fdca8357d5bfe9
4211dd SP:Homo sapiens
              ID:2888721359 PL:illumina
                                                                             PU:D1BA4ACXX.3 LB:H_KA-452198-0817007-cDNA-3-lib1
                                                                                                                                                                                          PI:365 DS:paired end DT:2012-10-03T19:00:00-0500
              ID:2888721359 VN:2.0.8
                                                                             CL:tophat --library-type fr-secondstrand --bowtie-version=2.1.0
                                                             PN:MarkDuplicates
                                                                                                            PP:2888721359 VN:1.85(exported)
                                                                                                                                                                                          CL:net.sf.picard.sam.MarkDuplicates INPUT=[/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blad
e10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300.bam] OUTPUT=/gscmnt/gc13001/info/build_merged_alignments/merged_alignment-blade10-2-5.gsc.wustl.edu-jw
alker-15434-136080019/scratch-ILq6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300-post_dup.bam METRICS_FILE=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-1543
4-136080019/staging-liuJS/H_KA-452198-0817007-cDNA-3-lib1-2888360300.metrics REMOVE_DUPLICATES=false ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=9500 TMP_DIR=[/gscmnt/gci3001/info/build_merged_al
ignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILq6Y] VALIDATION_STRINGENCY=SILENT_MAX_RECORDS_IN_RAM=500000 PROGRAM_RECORD_ID=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicates_PROGRAM_GROUP_NAME=MarkDuplicat
DUPLICATE MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=50000 SORTING_COLLECTION_SIZE_RATIO=0.25 READ_NAME_REGEX=[a-za-Z0-9]+:[0-9]+:[0-9]+):([0-9]+).* OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 VERBOSITY=INFO
QUIET=false COMPRESSION_LEVEL=5 CREATE_INDEX=false CREATE_MD5_FILE=false
mgriffit@linus270 ~>
```

Example SAM/BAM/CRAM alignment section (only 10 alignments shown)

<pre>mgriffit@linus270 -> samtools view -f 3 -F 1804 /gscmnt/gc13001/info/model_data/2891632684/build136494</pre>	4552/alignments/136080019.bam head
HWI-ST495_129147882:3:2114:15769:38646_991	ACTGCGGGGCCCTCTTGCTTACTGATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCAGCACGC
CCFFFFHHHGHJJJJJJJJJJJJJJJJJHIJJJJJJHFDDDDDDDDDDDD	CC;Z:15 MD:Z:5A94 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:
1 XN:i:0 XO:i:0 CP:i:102519765 AS:i:-5 XS:A:+ YT:Z:UU	
HWI-ST495_129147882:3:2114:15769:38646 147 1 11508 3 100M = 11306 -302	ACTCCTAAATATGGGATTCCTGGGTTTAAAAGTATAAAATAAAT
;5:CDCDCDECEFCD@9E=?7EEIIIIHCEGGIJJJJIIJJIHF@?00IHHFFGG?*JJJIJJJHGEIJJJJJJJJHHCIEJJJHFHHGHFFEDFCCB	CC:Z:15 MD:Z:34A65 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:
1 XN:i:0 XO:i:0 CP:i:102519563 AS:i:-6 XS:A:+ YT:Z:UU	
HWI-ST495_129147882:3:1210:1257:16203	CCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTTCTTTC
CCFFFFFHFHAFGGIIIJJJEEHGIGGGIJIJJGI?@EHIGIJDGHIHIGGIJJJJJJJJJJJJGHHHGHFFFCDDDDDDCDCCCCCA;>@-@AA@:AA>AA	CC:Z:15 MD:Z:100 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
0 XN:i:0 XO:i:0 CP:i:102519261 AS:i:0 XS:A:- YT:Z:UU	
HWI-ST495_129147882:3:1210:1257:16203 83 1 12055 3 100M = 11810 -345	GAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATAC
CC>4C>DCCCACACDCC?BDCEE@ECFFFFHHHHIJJJIIJJIIJHEHIIGJIJJJJIGHIIIJJJJJJIIJJJJJJJJJJ	CC:Z:15 MD:Z:100 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
0 XN:i:0 X0:i:0 CP:i:102519016 AS:i:0 XS:A:+ YT:Z:UU	
HWI-ST495_129147882:3:2111:3117:78828 163 1 12634 3 100M = 12746 212	GCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCACAGGTGAGAGGAGAGG
@@FFFFFDHHHH9FHGIIFGAFDHEGII>GHIIIIIIIIIIIIIIIIIIIHDDFFEEECEECCACCCCCC:AADCCBCC>CAC <ccccc:@cb@@bab##< td=""><td>CC:Z:15 MD:Z:85G14 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:</td></ccccc:@cb@@bab##<>	CC:Z:15 MD:Z:85G14 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:
1 XN:i:0 X0:i:0 CP:i:102518437 AS:i:-5 XS:A:- YT:Z:UU	
HWI-ST495_129147882:3:2111:3117:78828 83 1 12746 3 100M = 12634 -212	GGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCATCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTD
DCABDBDDDDDDDDDDDDDDBB@BDDDB@;CCCCDEFD@;.7 <higgeigehigjjjiigigiihegfehfjiiiiigjjjjhhhhhfffffc@< td=""><td>CC:Z:15 MD:Z:37G62 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:</td></higgeigehigjjjiigigiihegfehfjiiiiigjjjjhhhhhfffffc@<>	CC:Z:15 MD:Z:37G62 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:
1 XN:i:0 X0:i:0 CP:i:102518325 AS:i:-5 XS:A:- YT:Z:UU HWI-ST495 129147882:3:1102:4242:26638 99 1 13503 3 100M = 13779 376	CGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAC
HWI-ST495_129147882:3:1102:4242:26638	CC:Z:2 MD:Z:100 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
0 XN:i:0 XO:i:0 CP:i:114357414 AS:i:0 XS:A:+ YT:Z:UU	CC:2:2 PD:2:100 PG:2:MarkDupticates RG:2:2000/21559 AG:1:0 Nn:1:2 H1:1:0 AH:1:
HWI-ST495 129147882:3:1389:15328:74982 99 1 13534 3 100M = 13780 346	AGACGGTGTTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCACCAGGCACAGGG
CCFFFADHHHHIJJJJJJIJJIJJHJJJJJJJJJJJJJJJJJJJJ	CC:Z:2 MD:Z:100 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
0 XN:i:0 X0:i:0 CP:i:114357383 AS:i:0 XS:A:+ YT:Z:UU	CCIZIZ NO.Z.100 TG.Z.Natroopercares No.Z.2000/21333 AG.110 Nilli Nilli Nilli Nilli
HWI-ST495 129147882:3:1308:10126:19636 99 1 13779 3 100M = 14027 348	CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCCGATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTGC
CCFFFFFHGHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	CC:Z:2 MD:Z:100 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
0 XN:i:0 XO:i:0 CP:i:114357140 AS:i:0 XS:A:+ YT:Z:UU	
HWI-ST495 129147882:3:1102:4242:26638 147 1 13779 3 100M = 13503 -376	CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGACGAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTG#
##DCCDDDCCBBBABCCDDDCBDDBBDHC?=GIIJIIIIJIGIIIIJJHJJJJJJIGCIIJJJJJJJIGHGJJIJJJJJJJIIIIIGGFGHHHHFFFFFCCC	CC:Z:2 MD:Z:100 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
0 XN:i:0 XO:i:0 CP:i:114357140 AS:i:0 XS:A:+ YT:Z:UU	
mgriffit@linus270 <>	
·	

Introduction to the SAM/BAM format

- The specification
 - http://samtools.sourceforge.net/SAM1.pdf
- SAM is uncompressed text data
- BAM is a compressed version of SAM
 - lossless BGZF format
- BAM files are usually 'indexed'
 - A '.bai' file will be found beside the '.bam' file
 - Indexing provides fast retrieval of alignments overlapping a specified region without going through all alignments.
 - BAM must be sorted by the reference ID and then the leftmost coordinate before indexing

SAM/BAM header section

- Used to describe source of data, reference sequence, method of alignment, etc.
- Each section begins with character '@' followed by a two-letter record type code. These are followed by two-letter tags and values:
 - @HD The header line
 - VN: format version
 - SO: Sorting order of alignments
 - @SQ Reference sequence dictionary
 - SN: reference sequence name
 - LN: reference sequence length
 - SP: species

- @RG Read group
 - ID: read group identifier
 - CN: name of sequencing center
 - SM: sample name
- @PG Program
 - PN: program name
 - VN: program version

SAM/BAM alignment section

Col	Field	Type	Regexp/Range	Brief description		
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME		
$\bigstar 2$	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		
$\star 6$	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		

Example values

```
QNAME e.g. HWI-ST495 129147882:1:2302:10269:12362
   FLAG
        e.g.
   RNAME e.g.
   POS
        e.g. 11623
   MAPQ
        e.g.
        e.g. 100M
   CIGAR
   RNEXT
        e.g.
        e.g. 11740
   PNEXT
             217
   TLEN
        e.g.
10
        e.q. CCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTTGGGCTGGGCCTTGCCATGT
   SEQ
11
    QUAL
```

SAM/BAM flags explained

- 12 bitwise flags describing the alignment
- Stored as a binary string of length 12 instead of 12 columns of data
- Value of '1' indicates the flag is set. e.g. 001000000000
- All combinations can be represented as a number from 0 to 4095 (i.e. 2¹²-1). This number is used in the BAM/SAM file.
- You can specify 'required' or 'filter' flags in samtools view using the '-f' and '-F' options respectively

	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

Note that to maximize confusion, each bit is described in the SAM specification using its hexadecimal representation (i.e., '0x10' = 16 and '0x40' = 64).

http://broadinstitute.github.io/picard/explain-flags.html

CIGAR strings explained

• The CIGAR string is a sequence of base lengths and associated 'operations' indicating which bases align to the reference (either a match or mismatch), are deleted, are inserted, represent introns, etc.

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)	;' that
P	6	padding (silent deletion from padded reference)) liidl
=	7	sequence match	
X	8	sequence mismatch	_

•e.g. 81M859N19M

•A 100 bp read consists of: 81 bases of alignment to reference, 859 bases skipped (an intron), 19 bases of alignment

CRAM files

- CRAM is an ultra-compressed version of a BAM file
 - Usually between 30-60% smaller than the corresponding BAM
- Stores "diffs" from the reference genome
 - requires the matching reference genome to restore original data!
- Base quality binning may be used as well
- Some tools still require conversion back to bam

Quality Score Bins	Example of Empirically Mapped Quality Scores*		
N (no call)	N (no call)		
2–9	6		
10–19	15		
20–24	22		
25–29	27		
30–34	33		
35–39	37		
≥ 40	40		
	·		

By replacing the quality scores between 19 and 25 with a new score of 22, data storage space is conserved.

^{*}The mapped quality score of each bin (except "N") is subject to change depending on individual Q-tables.

Introduction to the BED format

- When working with BAM files, it is very common to want to examine a focused subset of the reference genome
 - e.g. the exons of a gene
- These subsets are commonly specified in 'BED' files
 - https://genome.ucsc.edu/FAQ/FAQformat.html#format1
- Many BAM manipulation tools accept regions of interest in BED format
- Basic BED format (tab separated):
 - Chromosome name, start position, end position
 - Coordinates in BED format are 0 based

Manipulation of SAM/BAM and BED files

 Several tools are used ubiquitously in sequence analysis to manipulate these files

- SAM/BAM files
 - samtools
 - bamtools
 - Picard
- BED files
 - bedtools
 - bedops



Common sources of confusion

Genomic coordinate systems

Genome builds

Variant representation

Genomic coordinates – 1 vs 0 based

chr1		Τ	Α	С		G	Т	C	Α	
1-based		1	2	3		4	5	6		
0-based	0	1		2	3		4	5	6	7

	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

- 1-based : Single nucleotides, variant positions, or ranges are specified directly by their corresponding nucleotide numbers
 - GFF, SAM, VCF, Ensembl browser, ...
- O-based: Single nucleotides, variant positions, or ranges are specified by the coordinates that flank them
 - BED, BAM, UCSC browser, ...

Genome builds

Reference Genome builds

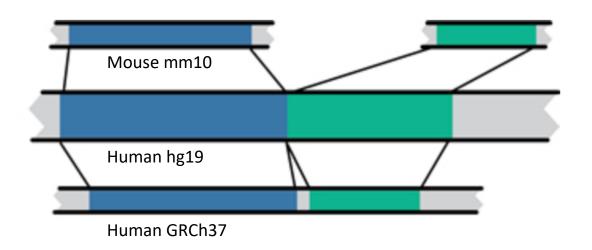
Current human: GRCh38, hg38, b38

alternate: GRCh38v2_ccdg

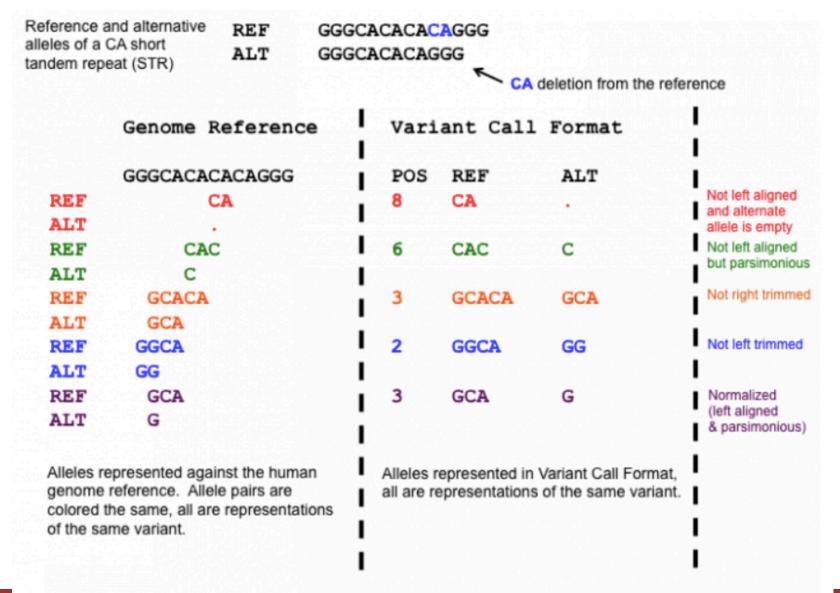
Previous human: GRCh37, hg19, b37

Current mouse: GRCm38, mm10

Lift-over



Variant shifting (alignment) and parsimony/trimming



Parsimony: representing variant in as few nucleotides as possible without reducing the length of any allele to 0

Left (right) aligning =
shifting the start position of
a variant as far to the left
(right) as possible

How should I sort my SAM/BAM file?

- Generally BAM files are sorted by <u>position</u>
 - This is for performance reasons
 - When sorted and indexed, arbitrary positions in a massive BAM file can be accessed rapidly
- Certain tools require a BAM sorted by <u>read name</u>
 - Usually this is when we need to easily identify both reads of a pair
 - The insert size between two reads may be large
 - In fusion detection we are interested in read pairs that map to different chromosomes