

Canadian Bioinformatics Workshops

www.bioinformatics.ca bioinformaticsdotca.github.io

Supported by

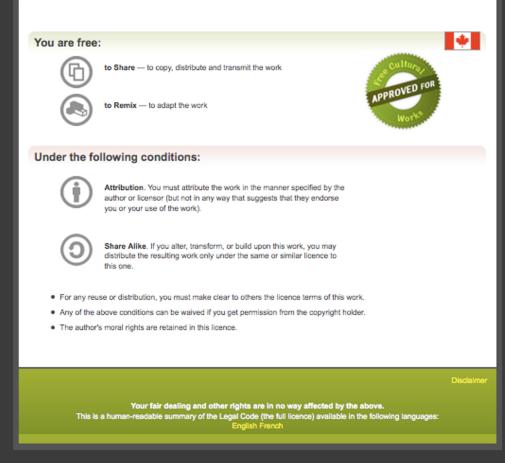


Creative Commons

This page is available in the following languages: Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Cestellano Castellano (AR) Espeñol (CL) Castellano (CO) Espeñol (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomeksi français français (CA) Galego ארער hrvatski Magyar Italiano 日本語 한국어 Macedonian Metayu Nederlands Norsk Sesotho sa Lebos polski Português română slovenski gizîk cpncxv srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada



Learn how to distribute your work using this licence

RNA: Module 2

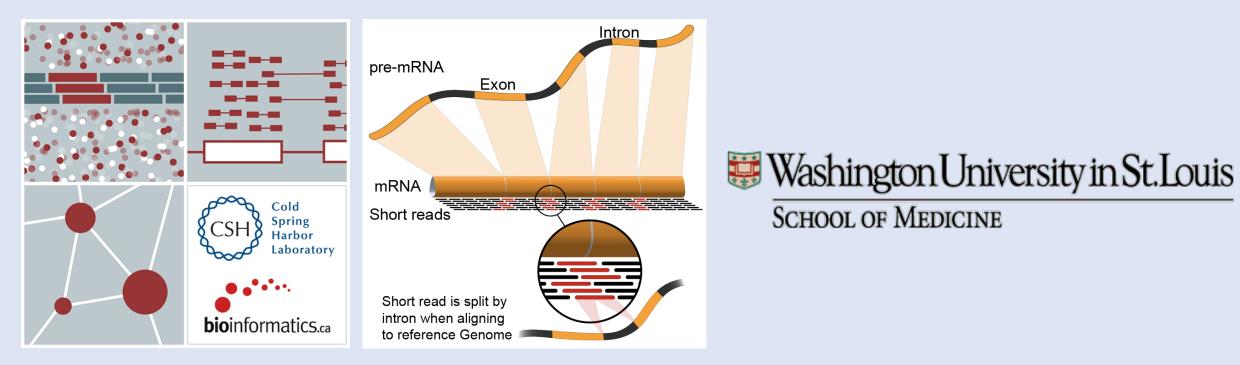
SAM/BAM/BED file formats

Emma Bell, Felicia Gomez, Obi Griffith, Malachi Griffith, Huiming Xia

RNA-Seq Analysis

Sep 8th-10th, 2021





bioinformatics.ca

RNA: Module 2

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"

@HD VN:1.4 SO:coordinate

@SQ 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

@RG 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 SM:H_KA-452198-0817007 CN:WUGSC

@PG ID:2888721359 VN:2.0.8 CL:tophat --library-type fr-secondstrand --bowtie-version=2.1.0

@PG
ID:MarkDuplicates
PN:MarkDuplicates
PP:2898721359
VN:1.85(exported)
CL:net.sf.picard.sam.MarkDuplicates
INPUT=[/gscmnt/gc13001/info/build_merged_alignments/m

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

CC>4C>DCCCACACDCC?BDCEE@ECFFFFHHHHHIJJJIIJJIIHHEHIIGJIJJJIGHIIJJJJJIIJJJIJJJJIJ		
CCFFFFFHH#GH1JJJJJJJJMCJJJMCDDDDDDDDDDDDDDDDDDDDDDDD		
$ \begin{array}{c} 1 & \text{Niie} \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \$		
$ \begin{split} \label{eq:heat-stability} \\ \hline heat-stability} \\ \hline heat-stability \\ \hline heat-stability} \\ \hline heat-stability} \\ \hline heat-stability \\ \hline heat-stability} \\ \hline heat-stability} \\ \hline heat-stability} \\ \hline heat-stability} \\ \hline heat-stability \\ \hline heat-stabili$		CC:2:15 MD:2:5A94 PG:2:MarkDuplicates RG:2:2888/21359 XG:1:0 NH:1:2 H1:1:0 NM:1:1 XM:1:
1:5:00CDCCECCPC0g6=77EFIIIHEGG1JJJIJJIJIJIHF0780IHHFG674JJIJGHGEIJJJJJJJJHCEHGGGGJJJJGT26HIHFG674JJIJGHGEIJJJJJJJJHCEHGGGGJJJJGT26HIHFG674JGT2674 CC:2:15 MD:2:34A65 PG:2:MarkDuplicates RG:2:2888721359 XG:10 NH:1:2 HI:10 NM:1:1 XM:1: NH:1: XM:1:1 VM:1:0 X0:1:0 CP:1:102519261 AS:1:0 XS:A:+ YT:2:UU VM:1:0 X0:1:0 CP:1:102519261 AS:1:0 XS:A:+ YT:2:UU VM:1:10 X0:1:0 CP:1:102519261 AS:1:0 XS:A:+ YT:2:UU CC:2:15 MD:2:100 PG:2:MarkDuplicates RG:2:2888721359 XG:1:0 NH:1:0 N		
$ \begin{array}{llllllllllllllllllllllllllllllllllll$		
$ \begin{split} Hul = 5495, 129147882; 3:1216:1257; 15283 163 1 11818 3 100M = 12855 345 \\ CTGCATGTAGTTAACTCACACCTGGTTAACTCACACCTGTTGTTAACTGCGCTTAGCTGCCCCCAC; 2&=QAAQ; AA-AA \\ MI = 5195, 129147882; 3:1210; 1257; 15283 83 1 1285 5 100M = 11810 - 345 \\ O & Mi: 10 & Mi$		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:
CC:FFFFHHFiAFGGIIJJJGT/@EHIGIDGHIHIGGIJJJJJIJGHIHHGHFFFCDDDDDDDCDCCCCCA:>@-@AA@:AA>AA CC:Z:IS MD:Z:100 PG:Z:MarkDuplicates RG:Z:2888721359 XG::0 NH::0 NH::0 NH::0 NH::0 NH::10 <		
0 XN:10 XO:110 CP:1102519261 AS:10 XS:1- YT:2:UU HMI-ST495_129147882:3:12914257:15208 B3 1 12955 3 100M 1181 -345 GACACTGGAGTGGACTGCCTAGAGGAGGAGGGGCATGGCTCAGGTGGGACTGGTGCACTGGTGCACGTGGCACTGGCCCTGGTGCTGCAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGCATGGCTCAGGTGGAGGAGGGGGCCGCCTTGGATGACCACTCTTGTGAGGGCCCGCCC		CCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTCTGCC
HWI-57495_129147882:3:1219:1257:16283 83 1 12855 3 0 M = 11810 -345 GC>4C>CCAC>DCCCRACCDCC7BDCEE@CFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	CCFFFFHFHAFGGIIIJJJEHGIGGGIJIJJGI?@EHIGIJDGHIHIGGIJJJJJJJJJJJJGHHHGHFFFCDDDDDDCDCCCCCCA;>@>@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:
CC-4C>DCCCACDCC2RD0CEgeCrFFFHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH	0 XN:i:0 X0:i:0 CP:i:102519261 AS:i:0 XS:A:- YT:Z:UU	
0 XN::10 CP::1:102519016 AS::10 XS::4: YT:Z:UU HMI-ST495_129147882:3:2111:3117:78828 163 12634 100M 12746 212 GGCFFFFFHHHHH9HGIFLEFEACPDEEDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD	HWI-ST495_129147882:3:1210:1257:16203 83 1 12055 3 100M = 11810 -345	GAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATAC
HMI-ST495_129147882:3:2111:3117:78828 1 12634 3 100M = 12746 212 G@CFFFFFDHHHM9FHGITGAPOHEGIT>GHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	CC>4C>DCCCACACDCC?BDCEE@ECFFFFHHHHHIJJJIIJJIIIHHEHIIGJIJIJJJIGHIIIJJJJJIIJJJJJIJJJJJJJJJJ	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:
@@FFFFDHHHH9FHGIIFGAFDHEGI>GHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	0 XN:i:0 X0:i:0 CP:i:102519016 AS:i:0 XS:A:+ YT:Z:UU	
1 XN:i:0 X0:i:0 CP:i:102518437 AS:i:-5 XS:A:- YT:Z:UU HMI-ST495_129147882:3:2111:317:78828 83 1 12746 3 100M = 12634 -212 GGGAGTGGCGTCTACGGGGCTCTACGGGGACTCTCTGGAGGAGGCTTCGATGGCCCCTCACACCCCCTCTGATGTCT DCABB0BDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD	HWI-ST495_129147882:3:2111:3117:78828 163 1 12634 3 100M = 12746 212	GCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCACAGGTGAGAGGAGAGA<
HWI-ST495_129147882:3:2111:3117:78828B31127463100M=12634-212DCABDBDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD	@@FFFFDHHHH9FHGIIFGAFDHEGII>GHIIIIIIIIIIIIIIIIIFHDDFFEEECEECCCACCCCCC:AADCCBCC>CAC <cccccc:@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></cccccc:@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:
DCABDBDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD	1 XN:i:0 X0:i:0 CP:i:102518437 AS:i:-5 XS:A:- YT:Z:UU	
1 XN:i:0 X0:i:0 CF:i:102518325 AS:i:-5 XS:a:- YT:Z:UU HWI-ST495_129147882:3:1102:4242:2638 99 1 13503 100M 13779 376 CCFFFFFHHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJ	HWI-ST495 129147882:3:2111:3117:78828 83 1 12746 3 100M = 12634 -212	GGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCATCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTD
HWI-ST495_129147882:3:1102:4242:26638 99 1 13503 100M = 13779 376 CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJ	DCABDBDDDDDDDDDDDDDDDDDDDBDB@BDDDB@;CCCCCDEFD@;.? <higgeigehigjjjiigigiihegfehfjiiiiigjjjjhhhhhffffc@@< 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></higgeigehigjjjiigigiihegfehfjiiiiigjjjjhhhhhffffc@@<>	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:
CCFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	1 XN:i:0 X0:i:0 CP:i:102518325 AS:i:-5 XS:A:- YT:Z:UU	
CCFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	HWI-ST495 129147882:3:1102:4242:26638 99 1 13503 3 100M = 13779 376	CGCTGTGCCCTTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAC
0 XN:i:0 CP:i:114357414 AS:i:0 XS:A:+ YT:Z:UU HWI-ST495_129147882:3:1309:15328:74082 99 1 13534 100M = 13780 346 AGACGGTGTTTGCCAGGGGCCCGGGGCTGGCAGGGGGGGG		
HWI-ST495_129147882:3:1309:15328:74082 99 1 13534 3 100M = 13780 346 AGACGGTGTTTGTCATGGGGCTCGCAGGGATCCTGCAAGGGATGTGGAGACCCAGGAGGTGTGGAGGCCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGAGGTGTGGAGGTCCAGGGGTGCCAGGGCACAGGG VN:i:0 XN:i:0 CP:i:114357383 AS:i:0 XS:A:+ YT::0 VM:i:0 XM:i:0 PG:Z:MarkDuplicates RG:Z:2888721359 XG:i:0 NH:i:0 XM:i:0 XM:i:0 XN:i:0 XN		
CCFFFADHHHHFIJJJJJIJJJJJJJJJJJJJJJJJJJJJJJJJJJ		AGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGGACCCAGGG
0 XN:i:0 XD:i:0 CP:i:114357383 AS:i:0 XS:A:+ YT:Z:UU HMI-57495_129147882:3:1308:1026:19636 99 1 13779 3 100M = 14027 348 CCTCTGCAGGAGGCTGCCATTTGTCTGCCACCTCTGTACTGCCACTGCTACTGCCACTGCTACTGCCACTGCTACTAAAGTTAGCTGC CCFFFFFHKH1JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ		
HWI-ST495_129147882:3:1308:10126:19636 99 1 13779 3 100M = 14027 348 CCCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCGCTCTTTAGAAGCGAGACGGAGCAGACCCAACCGCCCTTCTATAATAACTAAAGTTAGCTGC CCFFFFHHGHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ		
CCFFFFHHGHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ		
0 XN:i:0 X0:i:0 CP:i:114357140 AS:i:0 XS:A:+ YT:Z:UU HWI-5T495_129147882:3:1102:424:26638 147 1 13779 3 100M = 13503 -376 CCTCTGCAGGAGGCTGCCATTGTCCTGCCCACCTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTCTATAATAACTAAAGTTAGCTG# ##DCCDDDCCBBBABCCDDDCBDDBDHC?=GIIJIIIJIGIIIIJJHJJIJJIGGIJJJJJJJJJJJJJ		
HWI-ST495_129147882:3:1102:422:26638 147 1 13779 3 100M = 13503 -376 CCTCTGCAGGAGGCGCCGCCATTTGTCCTGCCACCGCGAGGAGGCGGAGCCGAGACCGAGACCGAGACCCATCTGCTACTGCCCTTCTATAATAACTAAAGTTAAGTTAGCTG# ##DCCDDDCCBBBABCCDDDCBDDBBDHC?=GIJJIIIJJGIIIJJJJJJJJJGCIJJJJJJJJJJJJJJ		
##DCCDDDCCBBBABCCDDDCBDDBBDHC?=GIIJIIIIJIGIIIIJJHJJIJJJJJJJJJJJJJJJJJJ		
0 XN:i:0 X0:i:0 CP:i:114357140 AS:i:0 XS:A:+ YT:Z:UU		
	mgranatetta	

Introduction to the SAM/BAM format

- The specification
 - <u>http://samtools.sourceforge.net/SAM1.pdf</u>
- 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
$\star 2$	FLAG	\mathbf{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	\mathbf{Int}	[0,2 ⁸ -1]	MAPping Quality
$\star 6$	CIGAR	String	$\times ([0-9] + [MIDNSHPX=]) +$	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

1	QNAME	e.g.	HWI-ST495 129147882:1:2302:10269:12362
2	FLAG	e.g.	99
3	RNAME	e.g.	1
4	POS	e.g.	11623
5	MAPQ	e.g.	3
6	CIGAR	e.g.	100M
7	RNEXT	e.g.	=
8	PNEXT	e.g.	11740
9	TLEN	e.g.	217
10	SEQ	e.g.	CCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGT
11	QUAL	e.g.	CCCFFFFFHHHHHJJIJFIJJJJJJJJJJJJJJJJJJJJJ

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. 00100000000
- 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).

bioinformatics.ca

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	
М	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)	
Η	5	hard clipping (clipped sequences NOT present in SEQ)	' that
Р	6	padding (silent deletion from padded reference)	llial
=	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.

bioinformatics.ca

*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
 - <u>https://genome.ucsc.edu/FAQ/FAQformat.html#format1</u>
- Many BAM manipulation tools accept regions of interest in BED format
- Basic BED format (tab separated):
 - Chromosome name, start position, end position (BED3)
 - Coordinates in BED format are 0 based

Introduction to the BED format

- There are several flavors of BED format: BED3, BED4, BED6, BED8, etc
- First 3 fields always required: chr, start, stop
- Followed by up to 9 additional optional fields: name, score, strand, thickStart, thickEnd, itemRGB, blockCount, blockSizes, blockStarts

chr7	127471196	127472363	Pos1	0	+
chr7	127472363	127473530	Pos2	0	+
chr7	127473530	127474697	Pos3	0	+
chr7	127474697	127475864	Pos4	0	+
chr7	127475864	127477031	Neg1	0	-
chr7	127477031	127478198	Neg2	0	-
chr7	127478198	127479365	Neg3	0	-
chr7	127479365	127480532	Pos5	0	+
chr7	127480532	127481699	Neg4	0	-

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		Т	А		С		G	Т		С	А	
1-based		 1	2		 3		 4	5		6	 7	
0-based	0		1	2		3	2	1	5	e	5	7
							1-base	ed		0-base	ed	

chr1:4-4 G

chr1:2-4 ACG

chr1:5-5 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, ...

Indicate a single nucleotide

Indicate a range of nucleotides

Indicate a single nucleotide variant

chr1:3-4 G

chr1:4-5

chr1:1-4 ACG

T/A

Genome builds

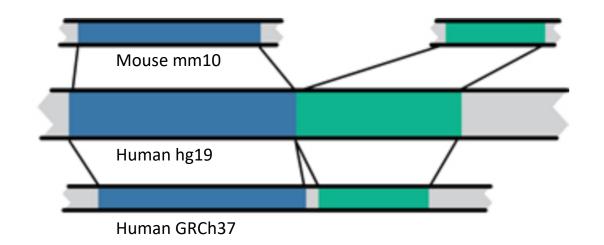
Reference Genome builds

Current human: GRCh38, hg38, b38 alternates: GRCh38v2_ccdg, GRCh38_full_analysis_set_plus_decoy_hla

Previous human: GRCh37, hg19, b37

Current mouse: GRCm38, mm10

Lift-over



For a detailed discussion of various human reference genome flavors refer here: https://pmbio.org/module-02-inputs/0002/02/01/Reference_Genome/

Variant shifting (alignment) and parsimony/trimming

Reference a alleles of a (tandem repe		REF ALT	GGGCACA GGGCACA		GG K	leletion from th	ne referer	nce
	Genome R	eferenc	e I t	Varia	ant Call	Format		
	GGGCACAC	ACAGGG		POS	REF	ALT	i	
REF	С	A	1	8	CA			Not left aligned and alternate
ALT								allele is empty
REF	CAC			6	CAC	С		Not left aligned
ALT	С		1				- 1	but parsimonious
REF	GCACA		1	3	GCACA	GCA		Not right trimmed
ALT	GCA							
REF	GGCA		1	2	GGCA	GG		Not left trimmed
ALT	GG						- 1	
REF	GCA		1.3	3	GCA	G	- 1	Normalized
ALT	G		- i -				i	(left aligned & parsimonious)

Alleles represented against the human genome reference. Allele pairs are colored the same, all are representations of the same variant. Alleles represented in Variant Call Format, all are representations of the same variant. **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 position
 - 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



We are on a Coffee Break & Networking Session

Workshop Sponsors:



McGill initiative in Computational Medicine



RNA: Module 2