

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Supported by



This page is available in the following languages:

Afrikaans Ελληνικά Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
 Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
 Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
 Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik срpski (latinica) Sotho svenska
 中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to **Share** — to copy, distribute and transmit the work



to **Remix** — to adapt the work



Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.
 This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
[English](#) [French](#)

[Learn how to distribute your work using this licence](#)

SAM/BAM/BED file formats

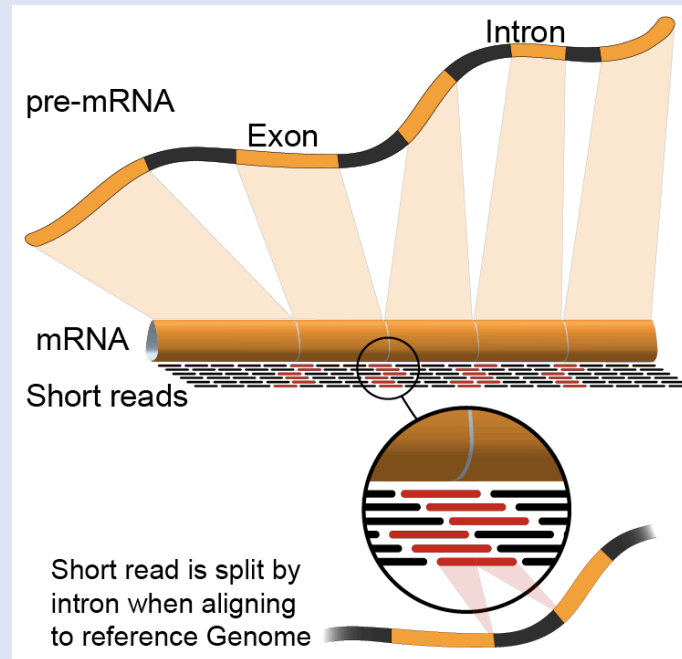
Emma Bell, Felicia Gomez, Obi Griffith, Malachi Griffith, Huiming Xia

RNA-Seq Analysis

Sep 8th-10th, 2021



A composite image containing four panels. The top-left panel shows a grid of red and white dots. The top-right panel shows a diagram of a gene with exons and introns. The bottom-left panel shows a network diagram with red nodes and white lines. The bottom-right panel contains the logos for Cold Spring Harbor Laboratory (CSH) and bioinformatics.ca.



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 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. $2^{12}-1$). 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) |
| P | 6 | padding (silent deletion from padded reference) |
| = | 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 (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

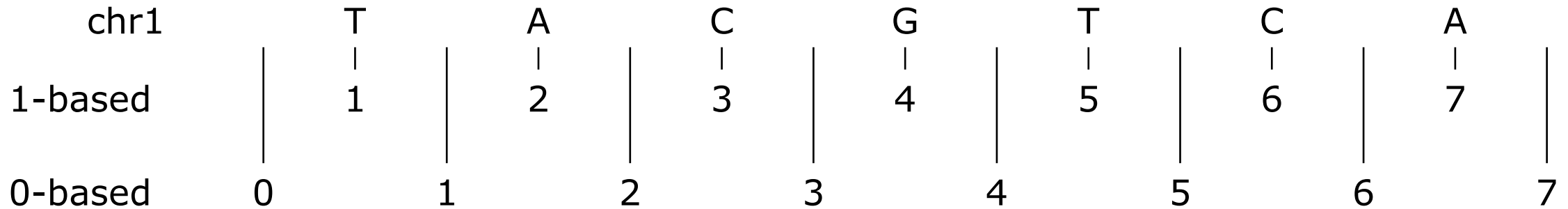
- Several tools are used ubiquitously in sequence analysis to manipulate these files
- SAM/BAM files
 - samtools
 - bamtools
 - Picard
- BED files
 - bedtools
 - bedops



Common sources of confusion

- Genomic coordinate systems
- Genome builds
- Variant representation

Genomic coordinates – 1 vs 0 based



| | 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, ...
- 0-based: Single nucleotides, variant positions, or ranges are specified by the coordinates that flank them
 - BED, BAM, UCSC browser, ...

Genome builds

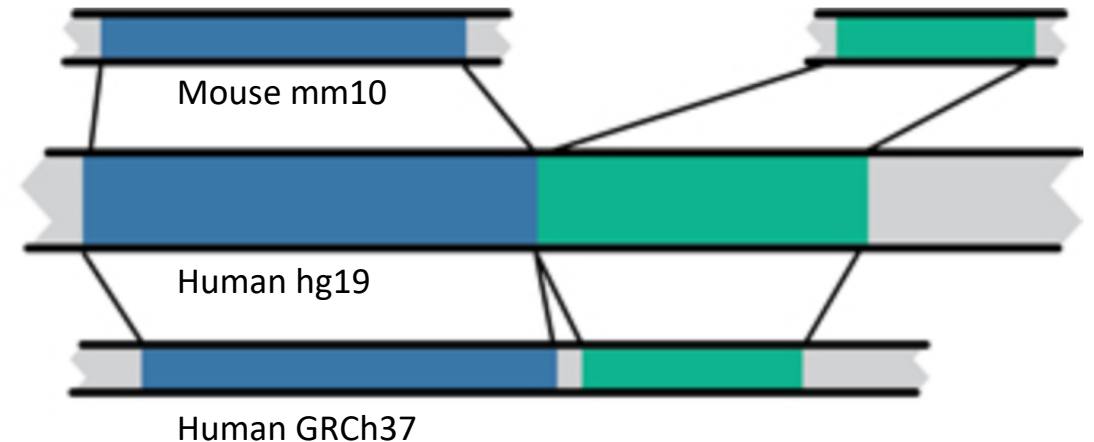
Reference Genome builds

Current human: GRCh38, hg38, b38
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 and alternative alleles of a CA short tandem repeat (STR)

REF GGGCACACACAGGG
ALT GGGCACACAGGG

← CA deletion from the reference

| | Genome Reference | | Variant Call Format | | | |
|-----|------------------|--|---------------------|-------|-----|--|
| | GGGCACACACAGGG | | POS | REF | ALT | |
| REF | CA | | 8 | CA | . | Not left aligned and alternate allele is empty |
| ALT | . | | | | | |
| REF | CAC | | 6 | CAC | C | Not left aligned but parsimonious |
| ALT | C | | | | | |
| REF | GCACA | | 3 | GCACA | GCA | Not right trimmed |
| ALT | GCA | | | | | |
| REF | GGCA | | 2 | GGCA | GG | Not left trimmed |
| ALT | GG | | | | | |
| REF | GCA | | 3 | GCA | G | Normalized (left aligned & parsimonious) |
| ALT | G | | | | | |

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 read name
 - 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:



Canadian Centre for
Computational
Genomics



neuro

Institut-Hôpital
neurologique de Montréal
Montreal Neurological
Institute-Hospital



OICR

Ontario Institute
for Cancer Research

MiC:M McGill initiative in
Computational Medicine



HPC4Health