Reference genomes and common file formats

Dóra Bihary
MRC Cancer Unit, University of Cambridge

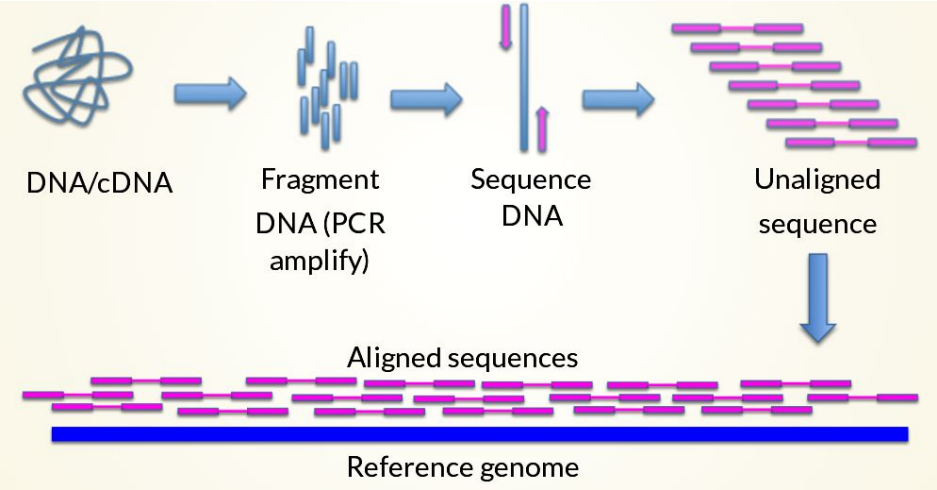
CRUK CI Bioinformatics Summer School
July 2019

Overview

- Reference genomes and GRC
- Fasta and FastQ (unaligned sequences)
- SAM/BAM/CRAM (aligned sequences)
- Summarized genomic features
  - BED (genomic intervals)
  - Wiggle files, BEDgraphs, BigWigs (genomic scores)
  - GFF/GTF (gene annotation)

Why do we need to know about reference genomes?

- Allows for genes and genomic features to be evaluated in their genomic context.
  - Gene A is close to gene B
  - Gene A and gene B are within feature C
- Can be used to align shallow targeted high-throughput sequencing to a pre-built map of an organism



Genome Reference Consortium (GRC)

- Most model organism reference genomes are being regularly updated
- Reference genomes consist of a mixture of known chromosomes and unplaced contigs called Genome Reference Assembly
- Genome Reference Consortium:
  - https://www.ncbi.nlm.nih.gov/grc
  - A collaboration of institutes which curate and maintain the reference genomes of 4 model organisms:
    - Human - GRCh38.p13 (February 2019)
    - Mouse - GRCm38.p6 (September 2017)
    - Zebrafish - GRCz11 (May 2017)
    - Chicken - GRCg6a (March 2018)
  - Latest human assembly is GRCh38
  - Patches add information to the assembly without disrupting the chromosome coordinates
    - Fix patches - will be incorporated into next major assembly
    - Novel patches - alternate sequences
- Other model organisms are maintained separately, like:
  - Drosophila - Berkeley Drosophila Genome Project

Overview

- Reference genomes and GRC
- Fasta and FastQ (unaligned sequences)
- SAM/BAM/CRAM (aligned sequences)
- Summarized genomic features
  - BED (genomic intervals)
  - Wiggle files, BEDgraphs, BigWigs (genomic scores)
  - GFF/GTF (gene annotation)

The reference genome

- A reference genome is a collection of contigs
- A contig refers to overlapping DNA reads encoded as A, G, C, T or N
- Typically comes in FASTA format:
  - ">" line contains information on contig
  - Following lines contain contig sequences

```
>gi|568815581:c7687550-7668402 Homo sapiens chromosome 17, GRCh38.p7 Primary
Assembly
GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGCTGGGAGCGTG
CTTTCCACGACGGTGACACGCTTCCCTGGATTGGGTAAGCTCCTGACTGAACTTGATGAGTCCTCTCTGA
GTCACGGGCTCTCGGCTCCGTGTATTTTCAGCTCGGGAAAATCGCTGGGGCTGGGGGTGGGGCAGTGGGG
ACTTAGCGAGTTTGGGGGTGAGTGGGATGGAAGCTTGGCTAGAGGGATCATCATAGGAGTTGCATTGTTG
GGAGACCTGGGTGTAGATGATGGGGATGTTAGGACCATCCGAACTCAAAGTTGAACGCCTAGGCAGAGGA
GTGGAGCTTTGGGGAACCTTGAGCCGGCCTAAAGCGTACTTCTTTGCACATCCACCCGGTGCTGGGCGTA
GGGAATCCCTGAAATAAAAGATGCACAAAGCATTGAGGTCTGAGACTTTTGGATCTCGAAACATTGAGAA
CTCATAGCTGTATATTTTAGAGCCCATGGCATCCTAGTGAAAACTGGGGCTCCATTCCGAAATGATCATT
TGGGGGTGATCCGGGGAGCCCAAGCTGCTAAGGTCCCACAACTTCCGGACCTTTGTCCTTCCTGGAGCGA
TCTTTCCAGGCAGCCCCCGGCTCCGCTAGATGGAGAAAATCCAATTGAAGGCTGTCAGTCGTGGAAGTGA
GAAGTGCTAAACCAGGGGTTTGCCCGCCAGGCCGAGGAGGACCGTCGCAATCTGAGAGGCCCGGCAGCCC
```

Unaligned sequences - FastQ

- Unaligned sequence files generated from HTS machines are mapped to a reference genome to produce aligned sequence

  FastQ (unaligned sequences) → SAM (aligned sequences)

- FastQ: FASTA with quality

```
1  @HS2000-887_89:5:1101:1595:156011/1
2  ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCGTAACAC
3  +
4  @?8BDDDABDHAFEGBGE@?BGE?FHG?GFH@@CG?;BB@B?BGGGHG9FDCF3=2CG;@=?(667?#################################
5  @HS2000-887_89:5:1101:1599:35168/1
6  CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACGCTATCCCTATCCCTAACCCTAACCATAACTCTAACCCTAACCCTCACCCTAATCCTA
7  +
8  :==<AA3?32A8?+2@A3<7+<ABAABA;@>AAB;0):=A#####################################################################
9  @HS2000-887_89:5:1101:1624:195842/1
10 CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
11 +
12 CCCFFFFFHGHHHJJJJJJJJJJIJJJJJJJJIJJJJJIJJJJJHJJJIJJJJJIIIFJJHHFFFFFFEEEEDDBCDDDBBDDDDBDDDDDDBDDDDBBC
```

  - "@" followed by identifier
  - Sequence information
  - "+"
  - Quality scores encoded as ASCII characters

Unaligned sequences - FastQ header

```
1  @HS2000-887_89:5:1101:1595:156011/1
2  ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCGTAACAC
3  +
4  @?8BDDDABDHAFEGBGE@?BGE?FHG?GFH@@CG?;BB@B?BGGGHG9FDCF3=2CG;@=?(667?################################
5  @HS2000-887_89:5:1101:1599:35168/1
6  CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACGCTATCCCTATCCCTAACCCTAACCATAACTCTAACCCTAACCCTCACCCTAATCCTA
7  +
8  :==<AA3?32A8?+2@A3<7+<ABAABA;@>AAB;0):=A############################################################
9  @HS2000-887_89:5:1101:1624:195842/1
10 CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
11 +
12 CCCFFFFFHGHHHJJJJJJJJJJIJJJJJJJJIJJJJJIJJJJJHJJJIJJJJJIIIFJJHHFFFFFFEEEEDDBCDDDBBDDDDBDDDDDDBDDDDBBC
```

- Header for each read can contain additional information
  - HS2000-887_89 - Machine name
  - 5 - Flowcell lane
  - /1 - Read 1 or 2 of pair

Unaligned sequences - FastQ qualities

```
1  @HS2000-887_89:5:1101:1595:156011/1
2  ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCGTAACAC
3  +
4  @?8BDDDABDHAFEGBGE@?BGE?FHG?GFH@@CG?;BB@B?BGGGHG9FDCF3=2CG;@=?(667?#################################
5  @HS2000-887_89:5:1101:1599:35168/1
6  CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACGCTATCCCTATCCCTAACCCTAACCATAACTCTAACCCTAACCCTCACCCTAATCCTA
7  +
8  :==<AA3?32A8?+2@A3<7+<ABAABA;@>AAB;0):=A#####################################################################
9  @HS2000-887_89:5:1101:1624:195842/1
10 CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
11 +
12 CCCFFFFFHGHHHJJJJJJJJJJIJJJJJJJJJIJJJJJIJJJJJHJJJIJJJJJIIIFJJHHFFFFFFEEEEDDBCDDDBBDDDDBDDDDDDBDDDDBBC
```

- Quality scores come after the "+" line
- Quality (*Q*) is proportional to -$\log_{10}$ probability of sequence base being wrong (*e*):

$$Q = - 10 * \log_{10}(e)$$

- Encoded in ASCII to save space
- Used in quality assessment and downstream analysis
- For further information: https://en.wikipedia.org/wiki/FASTQ_format

Overview

- Reference genomes and GRC
- Fasta and FastQ (unaligned sequences)
- SAM/BAM/CRAM (aligned sequences)
- Summarized genomic features
  - BED (genomic intervals)
  - Wiggle files, BEDgraphs, BigWigs (genomic scores)
  - GFF/GTF (gene annotation)

Aligned sequences - SAM format

- SAM - Sequence Alignment Map
- Standard format for sequence data
- Recognised by majority of software and browsers

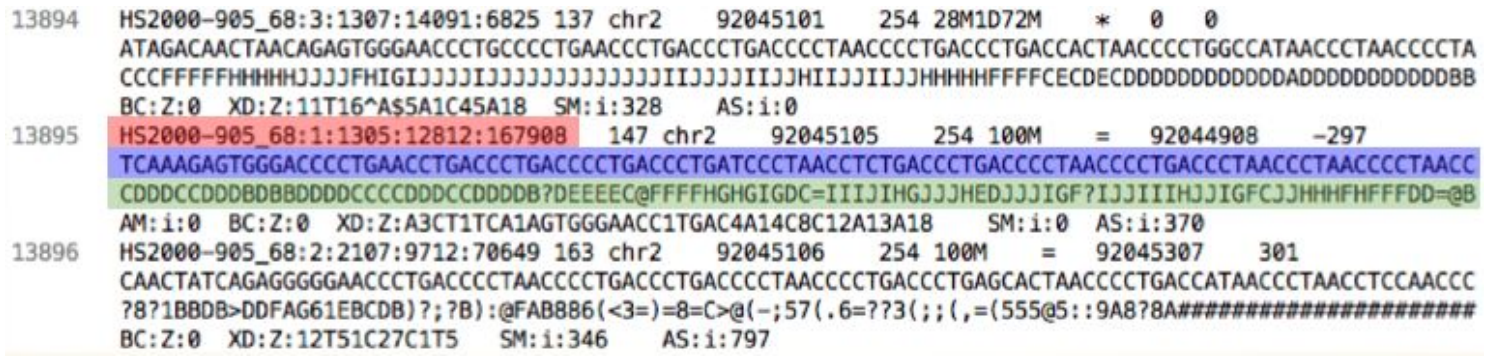
SAM header

- SAM header contains information on alignment and contigs used
- @HD - Version number and sorting information
- @SQ - Contig/Chromosome name and length of sequence

```
1  @HD VN:1.4  SO:coordinate
2  @SQ SN:chr10    LN:130694993
3  @SQ SN:chr11    LN:122082543
4  @SQ SN:chr12    LN:120129022
5  @SQ SN:chr13    LN:120421639
6  @SQ SN:chr14    LN:124902244
7  @SQ SN:chr15    LN:104043685
8  @SQ SN:chr16    LN:98207768
9  @SQ SN:chr17    LN:94987271
10 @SQ SN:chr18    LN:90702639
11 @SQ SN:chr19    LN:61431566
12 @SQ SN:chr1 LN:195471971
13 @SQ SN:chr2 LN:182113224
14 @SQ SN:chr3 LN:160039680
15 @SQ SN:chr4 LN:156508116
16 @SQ SN:chr5 LN:151834684
17 @SQ SN:chr6 LN:149736546
18 @SQ SN:chr7 LN:145441459
19 @SQ SN:chr8 LN:129401213
20 @SQ SN:chr9 LN:124595110
21 @SQ SN:chrM LN:16299
22 @SQ SN:chrX LN:171031299
23 @SQ SN:chrY LN:91744698
```

Aligned sequences - SAM format

SAM aligned reads



- Contains read and alignment information and location
  - Read name
  - Sequence of read
  - Encoded sequence quality

Aligned sequences - SAM format

SAM aligned reads

```
13894   HS2000-905_68:3:1307:14091:6825 137 chr2    92045101    254 28M1D72M    *   0   0
        ATAGACAACTAACAGAGTGGGAACCCTGCCCCTGAACCCTGACCCTGACCCCTAACCCCTGACCCTGACCACTAACCCCTGGCCATAACCCTAACCCCTA
        CCCFFFFFFHHHHHJJJJFHIGIJJJJIJJJJJJJJJJJJJJIIJJJJIIJJHIIJJIIJJHHHHHFFFFCECDECDDDDDDDDDDDDADDDDDDDDDDDBB
        BC:Z:0  XD:Z:11T16^A$5A1C45A18  SM:i:328    AS:i:0
13895   HS2000-905_68:1:1305:12812:167908   147 chr2    92045105    254 100M    =   92044908    -297
        TCAAAGAGTGGGACCCCTGAACCTGACCCTGACCCCTGACCCTGATCCCTAACCTCTGACCCTGACCCCTAACCCCTGACCCTAACCCTAACCCCTAACC
        CDDDCCDDDBDBBDDDDCCCCDDDCCDDDDB?DEEEEC@FFFFHGHGIGDC=IIIJIHGJJJHEDJJJIGF?IJJIIIHJJIGFCJJHHHFHFFFDD=@B
        AM:i:0  BC:Z:0  XD:Z:A3CT1TCA1AGTGGGAACC1TGAC4A14C8C12A13A18    SM:i:0  AS:i:370
13896   HS2000-905_68:2:2107:9712:70649 163 chr2    92045106    254 100M    =   92045307    301
        CAACTATCAGAGGGGGAACCCTGACCCCTAACCCCTGACCCTGACCCCTAACCCCTGACCCTGAGCACTAACCCCTGACCATAACCCTAACCTCCAACCC
        ?8?1BBDB>DDFAG61EBCDB)?;?B):@FAB886(<3=)=8=C>@(-;57(.6=??3(;;(,=(555@5::9A8?8A######################
        BC:Z:0  XD:Z:12T51C27C1T5   SM:i:346    AS:i:797
```

- Chromosome to which the read aligns
- Position in chromosome to which 5' end of the read aligns
- Alignment information - "Cigar string"
  - 100M - Continuous match of 100 bases
  - 28M1D72M - 28 bases continuously match, 1 deletion from reference, 72 base match

Aligned sequences - SAM format

SAM aligned reads

```
13894   HS2000-905_68:3:1307:14091:6825 137 chr2    92045101    254 28M1D72M    *   0   0
        ATAGACAACTAACAGAGTGGGAACCCTGCCCCTGAACCCTGACCCTGACCCCTAACCCCTGACCCTGACCACTAACCCCTGGCCATAACCCTAACCCCTA
        CCCFFFFFFHHHHHJJJJFHIGIJJJJIJJJJJJJJJJJJJJIIJJJJIIJJHIIJJIIJJHHHHHFFFFCECDECDDDDDDDDDDDDADDDDDDDDDDDBB
        BC:Z:0  XD:Z:11T16^A$5A1C45A18  SM:i:328    AS:i:0
13895   HS2000-905_68:1:1305:12812:167908   147 chr2    92045105    254 100M    =   92044908    -297
        TCAAAGAGTGGGACCCCTGAACCTGACCCTGACCCCTGACCCTGATCCCTAACCTCTGACCCTGACCCCTAACCCCTGACCCTAACCCTAACCCCCTAACC
        CDDDCCDDDDBDBBDDDDCCCCDDDCCDDDDB?DEEEEC@FFFFHGHGIGDC=IIIJIHGJJJHEDJJJIGF?IJJIIIHJJIGFCJJHHHFHFFFDD=@B
        AM:i:0  BC:Z:0  XD:Z:A3CT1TCA1AGTGGGAACC1TGAC4A14C8C12A13A18    SM:i:0  AS:i:370
13896   HS2000-905_68:2:2107:9712:70649 163 chr2    92045106    254 100M    =   92045307    301
        CAACTATCAGAGGGGGAACCCTGACCCCTAACCCCTGACCCTGACCCCTAACCCCTGACCCTGAGCACTAACCCCTGACCATAACCCTAACCTCCAACCC
        ?8?1BBDB>DDFAG61EBCDB)?;?B):@FAB886(<3=)=8=C>@(-;57(.6=??3(;;(,=(555@5::9A8?8A######################
        BC:Z:0  XD:Z:12T51C27C1T5   SM:i:346    AS:i:797
```

- Bit flag - TRUE/FALSE for pre-defined read criteria, like: is it paired? duplicate?
  - https://broadinstitute.github.io/picard/explain-flags.html
- Paired read position and insert size
- User defined flags

[1] Li H et al.,The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

Compressed aligned sequences - BAM and CRAM format

- SAM files can be large, so to save space people usually store some compressed versions of them instead:
  - BAM files
    - Binary SAM files
    - You also need to store an index file
  - CRAM files
    - Another way to compress alignment files designed by the EBI
    - The compression is driven by the reference the sequence data is aligned to, so it is very important that the exact same reference sequence is used for compression and decompression
    - Typically 40-50% space saving compared to BAM files
    - Full compatibility with BAM files
    - For further information: http://samtools.github.io/hts-specs/

Overview

- Reference genomes and GRC
- Fasta and FastQ (unaligned sequences)
- SAM/BAM/CRAM (aligned sequences)
- Summarized genomic features
  - BED (genomic intervals)
  - Wiggle files, BEDgraphs, BigWigs (genomic scores)
  - GFF/GTF (gene annotation)

Summarised genomic features formats

- After alignment, sequence reads are typically summarised into scores over/within genomic intervals
  - BED - genomic intervals with additional information
  - Wiggle files, BEDgraphs, BigWigs - genomic intervals with scores
  - GFF/GTF - genomic annotation with information and scores

BED format - genomic intervals

```
1  chr7    127471196  127472363
2  chr7    127472363  127473530
3  chr7    127473530  127474697
4  chr7    127474697  127475864
5  chr7    127475864  127477031
6  chr7    127477031  127478198
7  chr7    127478198  127479365
8  chr7    127479365  127480532
9  chr7    127480532  127481699
```

```
1  chr7    127471196  127472363  Pos1  10  +
2  chr7    127472363  127473530  Pos2  11  +
3  chr7    127473530  127474697  Pos3  20  +
4  chr7    127474697  127475864  Pos4  10  +
5  chr7    127475864  127477031  Neg1  98  -
6  chr7    127477031  127478198  Neg2  10  -
7  chr7    127478198  127479365  Neg3  67  -
8  chr7    127479365  127480532  Pos5  20  +
9  chr7    127480532  127481699  Neg4  50  -
```

- BED3 - 3 tab separated columns
  - Chromosome
  - Start
  - End
- Simplest format

- BED6 - 6 tab separated columns
  - Chromosome, start, end
  - Identifier
  - Score
  - Strand ("." stands for strandless)

bedGraph format - genomic scores (eg. signal intensity)

- BED-like format
- Starts as a 3 column BED file (chromosome, start, end)
- 4th column: score value

```
1   chr1    10001   10002   1
2   chr1    10003   10010   10
3   chr1    10011   10020   11
4   chr1    10021   10040   10
5   chr1    10041   10050   2
6   chr1    10051   99999   0
```

Wiggle format - genomic scores (eg. signal intensity)

Variable step Wiggle format

```
1   variableStep chrom=chr2
2   300701 12.5
3   300702 12.5
4   300703 12.5
5   300704 12.5
6   300705 12.5

9   variableStep chrom=chr2 span=5
10  300701 12.5
```

- Information line:
  - Chromosome
  - (Span - default=1, to describe continuous positions with same value)
- Each line contains:
  - Start position of the step
  - Score value

Fixed step Wiggle format

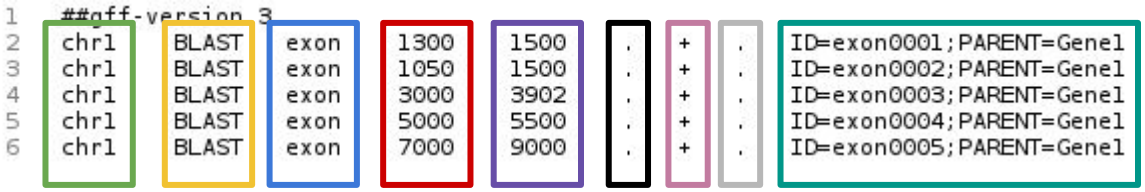
```
15  fixedStep chrom=chr3 start=400601 step=100
16  11
17  22
18  33

21  fixedStep chrom=chr3 start=400601 step=100 span=5
22  11
23  22
24  33
```

- Information line:
  - Chromosome
  - Start position of first step
  - Step size
  - (Span - default=1, to describe continuous positions with same value)
- Each line contains:
  - Score value

GFF/GTF files - genomic annotation

- Stores position, feature (exon) and meta-feature (transcript/gene) information



- Columns:
  - Chromosome
  - Source
  - Feature type
  - Start position
  - End position
  - Score
  - Strand
  - Frame - 0, 1 or 2 indicating which base of the feature is the first base of the codon
  - Semicolon separated attribute: ID (feature name);PARENT (meta-feature name)

Saving time and space - compressed file formats

- Many programs and browsers deal better with compressed, indexed versions of genomic files
  - SAM -> BAM (.bam and index file of .bai)
  - SAM/BAM -> CRAM (.cram file with the reference)
  - BED -> bigBed (.bb)
  - Wiggle and bedGraph -> bigWig (.bw/.bigWig)
  - BED and GFF -> (.gz and index file of .tbi)

Getting help and more information

- UCSC file formats
  - https://genome.ucsc.edu/FAQ/FAQformat.html
- IGV file formats
  - http://software.broadinstitute.org/software/igv/FileFormats
- Sanger file formats
  - http://gmod.org/wiki/GFF3

Acknowledgement

- Tom Carroll

  http://mrccsc.github.io/genomic_formats/genomicFileFormats.html#/