

Reference genomes and common file formats

Dóra Bihary

MRC Cancer Unit, University of Cambridge

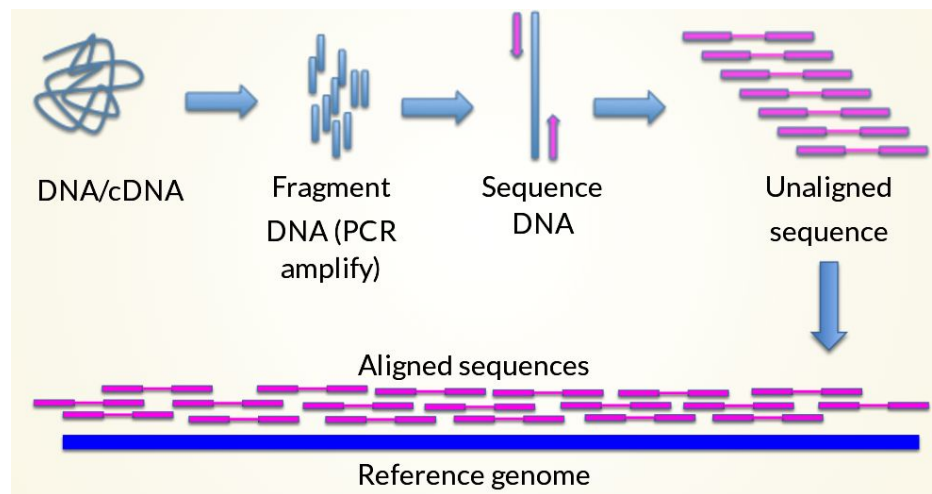
CRUK CI Bioinformatics Summer School
July 2018

Overview

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

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.p12 (December 2017)
 - Mouse - GRCm38.p6 (September 2017)
 - Zebrafish - GRCz11 (May 2017)
 - Chicken - GRCg6a (May 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)
 - GFF/GTF (gene annotation)
 - Wiggle files, BEDgraphs, BigWigs (genomic scores)

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  
GAT GGGATT GGGGTTTT CCCCT CCCAT GT GCT CAAGACT GGCGCT AAAAGTTTT GAGCTT CT CAAAAGT C  
TAGAGCCACCGT CCAGGGAGCAGGT AGCT GCT GGGCT CCGGGGACACTTT GCGTT CGGGCT GGGAGCGT G  
CTTT CCACGACGGT GACACGCTT CCCT GGATT GGGT AAGCT CCT GACT GAACTT GAT GAGT CCT CT CT GA  
GT CACGGGCT CT CGGCT CCGT GT ATTTT CAGCT CGGGAAAAT CGCT GGGGCT GGGGGT GGGGCAGT GGGG  
ACTT AGCGAGTTT GGGGGT GAGT GGGAT GGAAGCTT GGCT AGAGGGAT CAT CAT AGGAGTT GCATT GTT G  
GGAGACCT GGGT GT AGAT GAT GGGGAT GTT AGGACCAT CCGAACT CAAAGTT GAACGCCT AGGCAGAGGA  
GT GGAGCTTT GGGGAACCTT GAGCCGGCCT AAAGCGT ACTT CTTT GCACAT CCACCCGGT GCT GGGCGT A  
GGGAAT CCCT GAAAT AAAAGAT GCACAAAGCATT GAGGT CT GAGACTTTT GGAT CT CGAAACATT GAGAA  
CT CAT AGCT GT AT ATTTT AGAGCCCAT GGCAT CCT AGT GAAAACCT GGGGCT CCATT CCGAAAT GAT CATT  
T GGGGGT GAT CCGGGGAGCCCAAGCT GCT AAGGT CCCACAACCTT CCGGACCTTT GT CCTT CCT GGAGCGA  
T CTTT CCAGGCAGCCCCCGGCT CCGCT AGAT GGAGAAAAT CCAATT GAAGGCT GT CAGT CGT GGAAGT GA  
GAAGT GCT AAACCAGGGGTTT GCCCGCCAGGCCGAGGAGGACCGT CGCAAT CT GAGAGGCCCGGCAGCCC
```


Overview

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

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 S0: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
```


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)
 - GFF/GTF (gene annotation)
 - Wiggle files, BEDgraphs, BigWigs (genomic scores)

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)

Wiggle format - genomic scores

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 contiguous positions with same value)
- Each line contains:
 - Start position of the step
 - Score

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 contiguous positions with same value)
- Each line contains:
 - Score

bedGraph format - genomic scores

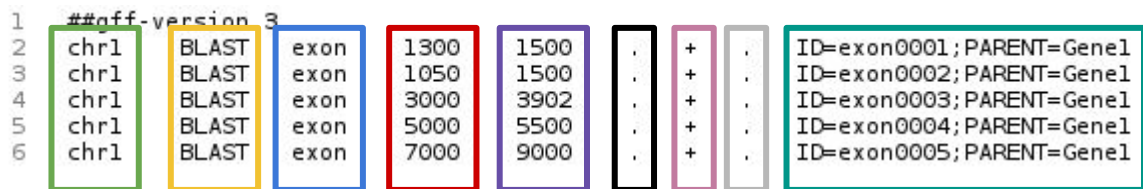
- BED-like format
- Starts as a 3 column BED file (chromosome, start, end)
- 4th column: score value

```
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
```

GFF/GTF files - genomic annotation

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

```
1 ##gff-version 3
2 chr1 BLAST exon 1300 1500 . + . ID=exon0001;PARENT=Gene1
3 chr1 BLAST exon 1050 1500 . + . ID=exon0002;PARENT=Gene1
4 chr1 BLAST exon 3000 3902 . + . ID=exon0003;PARENT=Gene1
5 chr1 BLAST exon 5000 5500 . + . ID=exon0004;PARENT=Gene1
6 chr1 BLAST exon 7000 9000 . + . ID=exon0005;PARENT=Gene1
```



- 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#/