Reference genomes and Common file formats

The Rockefeller University

Bioinformatics Resource Centre

Overview

- Reference genomes and GRC.
- Fasta and FastQ (Unaligned sequences).
- SAM/BAM (Aligned sequences).
- BED (Genomic Intervals).
- GFF/GTF (Gene annotation).
- Wiggle files, BEDgraphs and BigWigs (Genomic scores).

Are there we there yet?

- The human genome isnt complete!
- In fact, most model organisms's reference genomes are being regularly updated.
- Reference genomes consist of mixture of known chromosomes and unplaced contigs called a " Genome Reference Assembly".
- Major revisions to assembies result in change of co-ordinates.
  - Requires conversion between revisions.
  - The latest genome assembly for humans is GRCh38.
- Patches add information to the assembly without disrupting the chromosome coordinates . i.e GRCh38.p3

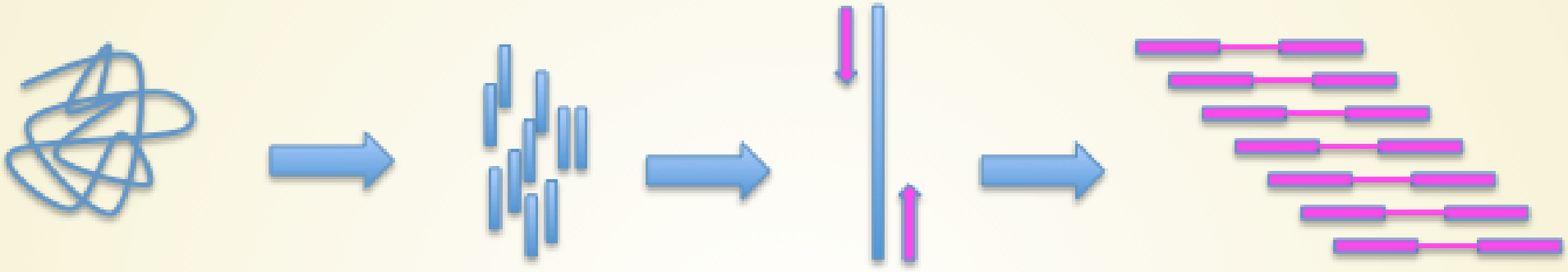
Genome Reference Consortium

- GRC is collaboration of institutes which curate and maintain the reference genomes for 3 model organims.
  - Human - GRCh38.p3
  - Mouse - GRCm38.p3
  - Zebrafish - GRCz10
- Other model organisms are maintained separately.
  - Drosophila - Berkeley Drosophila Genome Project, BDGP36

Why do we need to know about reference genomes

- Allows for genes and genomic features to be evaluated in their linear 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-thoughput sequencing to a pre-built map of an organisms genome.

Aligning to a reference genomes

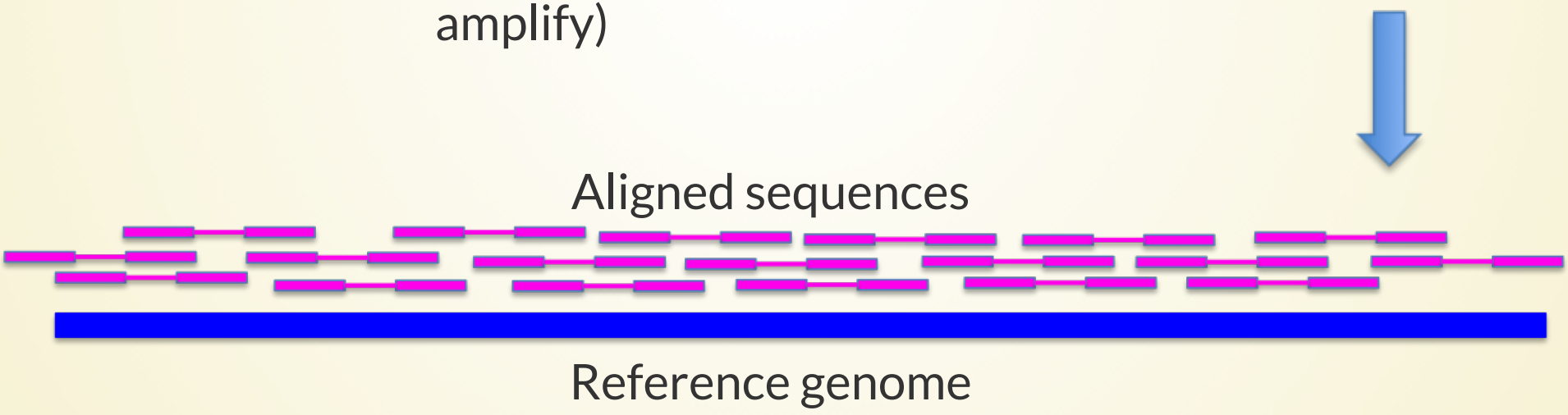


DNA/cDNA

Fragment DNA (PCR amplify)

Sequence DNA

Unaligned sequence



Aligned sequences

Reference genome

A reference genome

- A reference genome is a collection of contigs.
- A contig is a stretch of DNA sequence encoded as A,G,C,T,N.
- Typically comes in FASTA format.
  - ">" line contains information on contig
  - Lines following contain contig sequence

```
>gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCC
TAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCC
CAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACG
CGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCG
CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCG
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAG
CGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGA
GCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGG
AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCGTGG
GCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACG
GCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGGCGC
GGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCACTACA
GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCT
GAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAG
```

High-throughput Sequencing formats

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

- **FASTQ - Unaligned sequences**
- **SAM - Aligned sequences**

Unaligned Sequences

FastQ (FASTA with Qualities)

```
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 encodes as ASCI.

Unaligned Sequences

FastQ - Header

```
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  CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACGCTATCCCTATCCCTAACCCTAACCATAACTCTAACCCTAACCCTCACCCTAATCCTA
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
```

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

Unaligned Sequences

FastQ - Qualities

```
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
```

- Qualities follow "+" line.
- -log10 probability of sequence base being wrong.
- Encoded in ASCI to save space.
- Used in quality assessment and downstream analysis

Aligned sequences

SAM format

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

Aligned sequences

SAM - Header

```
@HD VN:1.4  SO:coordinate
@SQ SN:chr10    LN:130694993
@SQ SN:chr11    LN:122082543
@SQ SN:chr12    LN:120129022
@SQ SN:chr13    LN:120421639
@SQ SN:chr14    LN:124902244
@SQ SN:chr15    LN:104043685
@SQ SN:chr16    LN:98207768
@SQ SN:chr17    LN:94987271
@SQ SN:chr18    LN:90702639
@SQ SN:chr19    LN:61431566
@SQ SN:chr1 LN:195471971
@SQ SN:chr2 LN:182113224
@SQ SN:chr3 LN:160039680
@SQ SN:chr4 LN:156508116
@SQ SN:chr5 LN:151834684
@SQ SN:chr6 LN:149736546
@SQ SN:chr7 LN:145441459
@SQ SN:chr8 LN:129401213
@SQ SN:chr9 LN:124595110
@SQ SN:chrM LN:16299
@SQ SN:chrX LN:171031299
@SQ SN:chrY LN:91744698
```

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

Aligned sequences

SAM - Aligned Reads

```
13894   HS2000-905_68:3:1307:14091:6825 137 chr2    92045101    254 28M1D72M    *   0   0
        ATAGACAACTAACAGAGTGGGAACCCTGCCCCTGAACCCTGACCCTGACCCCTAACCCCTGACCCTGACCACTAACCCCTGGCCATAACCCTAACCCCTA
        CCCFFFFFHHHHHJJJJFHIGIJJJJIJJJJJJJJJJJJJJIIJJJJIIJJHIIJJIIJJHHHHHFFFFCECDECDDDDDDDDDDDDADDDDDDDDDDDBB
        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
```

- Contains read and alignment information and location

Aligned sequences

SAM

```
13894   HS2000-905_68:3:1307:14091:6825 137 chr2    92045101    254 28M1D72M    *   0   0
        ATAGACAACTAACAGAGTGGGAACCCTGCCCCTGAACCCTGACCCTGACCCCTAACCCCTGACCCTGACCACTAACCCCTGGCCATAACCCTAACCCCTA
        CCCFFFFFHHHHHJJJJFHIGIJJJJIJJJJJJJJJJJJJJIIJJJJIIJJHIIJJIIJJHHHHHFFFFCECDECDDDDDDDDDDDDADDDDDDDDDDBB
        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
```

- Read name.
- Sequence of read.
- Encoded sequence quality.

Aligned sequences

SAM

```
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 read aligns.
- Position in chromosome to which 5' of 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

```
13894   HS2000-905_68:3:1307:14091:6825 137 chr2    92045101    254 28M1D72M    *   0   0
        ATAGACAACTAACAGAGTGGGAACCCTGCCCCTGAACCCTGACCCTGACCCCTAACCCCTGACCCTGACCACTAACCCCTGGCCATAACCCTAACCCCTA
        CCCFFFFFHHHHHJJJJFHIGIJJJJIJJJJJJJJJJJJJJIIJJJJIIJJHIIJJIIJJHHHHHFFFFCECDECDDDDDDDDDDDDADDDDDDDDDDDBB
        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
```

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

Summarised Genomic Features formats

Post alignment, sequences reads are typically summarised into scores over/within genomic intervals.

- **BED - Genomic intervals and information.**
- **Wiggle/BedGraph - Genomic intervals and scores.**
- **GFF - Genomic annotation with information and scores**

Summarising in genomic intervals.

BED format (BED)

```
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
```

- Simple format
- 3 tab separated columns
- Chromsome, start, end

Summarising in genomic intervals.

BED format (BED6)

```
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  -
10
```

- Chromosome, start, end
- Identifier
- Score
- Strand ("." for strandless)

Summarising in genomic intervals.

narrowPeak and broadPeak

- narrowPeak and broadPeak are extensions to BED6 used in Encode's peak calling.
- Contains p-values, q-values.
- narrowPeak - BED 6+4
- broadPeak - BED6+3

Signal at genomic positions

- Common practice to review signal over genome.
- Special formats exist for this
  - Wiggle
  - bedGraph

Signal at genomic positions

Wiggle

```
variableStep chrom=chr21 span=5
9411191 50
9411196 40
9411201 60
9411206 20
9411211 20
9411216 20
9411221 40
9411226 60
9411231 40
9411236 40
9411241 40
9411246 40
9411251 40
9411256 60
9411261 20
9411266 60
9411271 60
9411276 40
9411281 20
9411286 40
9411291 60
9411296 60
9411301 60
9411306 20
```

- Information line
  - Chromosome
  - Step size
- Step start position
- Score

Signal at genomic positions

bedGraph

```
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
```

- BED 3 format
  - Chromosome
  - Start
  - End
- 4th column - Score

Genomic Annotation

GFF

```
1  ##gff-version 3
2  chr1  BLAST  exon  1300  1500  .  +  .  ID=exon00001;PARENT=Gene1
3  chr1  BLAST  exon  1050  1500  .  +  .  ID=exon00002;PARENT=Gene1
4  chr1  BLAST  exon  3000  3902  .  +  .  ID=exon00003;PARENT=Gene1
5  chr1  BLAST  exon  5000  5500  .  +  .  ID=exon00004;PARENT=Gene1
6  chr1  BLAST  exon  7000  9000  .  +  .  ID=exon00005;PARENT=Gene1
```

- Used to genome annotation.
- Stores position, feature (exon) and meta-feature (transcript/gene) information.

Genomic Annotation

GFF

```
1  ##gff-version 3
2  chr1  BLAST  exon  1300  1500  .  +  .  ID=exon00001;PARENT=Gene1
3  chr1  BLAST  exon  1050  1500  .  +  .  ID=exon00002;PARENT=Gene1
4  chr1  BLAST  exon  3000  3902  .  +  .  ID=exon00003;PARENT=Gene1
5  chr1  BLAST  exon  5000  5500  .  +  .  ID=exon00004;PARENT=Gene1
6  chr1  BLAST  exon  7000  9000  .  +  .  ID=exon00005;PARENT=Gene1
```

- Chromosome
- Start of feature
- End of Feature
- Strand

Genomic Annotation

GFF

```
1  ##gff-version 3
2  chr1  BLAST  exon  1300  1500  .  +  .  ID=exon00001;PARENT=Gene1
3  chr1  BLAST  exon  1050  1500  .  +  .  ID=exon00002;PARENT=Gene1
4  chr1  BLAST  exon  3000  3902  .  +  .  ID=exon00003;PARENT=Gene1
5  chr1  BLAST  exon  5000  5500  .  +  .  ID=exon00004;PARENT=Gene1
6  chr1  BLAST  exon  7000  9000  .  +  .  ID=exon00005;PARENT=Gene1
```

- Source
- Feature type
- Score

Genomic Annotation

GFF

```
1  ##gff-version 3
2  chr1  BLAST  exon  1300  1500  .  +  .  ID=exon00001;PARENT=Gene1
3  chr1  BLAST  exon  1050  1500  .  +  .  ID=exon00002;PARENT=Gene1
4  chr1  BLAST  exon  3000  3902  .  +  .  ID=exon00003;PARENT=Gene1
5  chr1  BLAST  exon  5000  5500  .  +  .  ID=exon00004;PARENT=Gene1
6  chr1  BLAST  exon  7000  9000  .  +  .  ID=exon00005;PARENT=Gene1
```

- Column 9 contains key pairs (ID=exon01), separated by semi-colons ";"
- ID - Feature name.
- PARENT- Meta-feature name.

Saving time and space
bigWig, bigBED and TABIX

- Many programs and browsers deal better with compressed, indexed versions of genomic files
  - SAM -> BAM (.bam and index file of .bai)
  - Wiggle and bedGraph -> bigWig (.bw/.bigWig)
  - BED -> bigBed (.bb)
  - 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
  - https://www.broadinstitute.org/igv/FileFormats
- Sanger (GFF)
  - https://www.sanger.ac.uk/resources/software/gff/spec.html