# BIO392 file formats and tools

Izaskun Mallona

- Commands/options are in `typewriter font`
- URLs are highlighted in blue

- Lecture from SIB
  https://edu.sib.swiss/pluginfile.php/2878/mod_resource/content/4/couselab-html/content.html

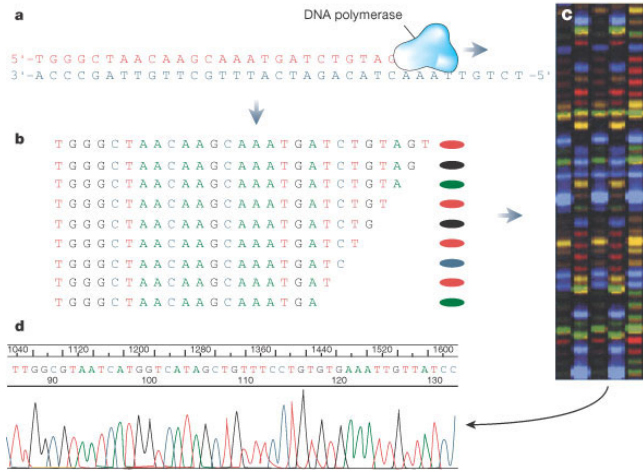- Run the exercises till number 4

# Commonly used formats

- Reference genomes
- Fasta and FastQ (Unaligned sequences)
- SAM/BAM (Alignments)
- BED (Genomic ranges)
- GFF/GTF (Gene annotation)
- BEDgraphs (Genomic ranges)
- Wiggle files, BEDgraphs and BigWigs (Genomic scores).
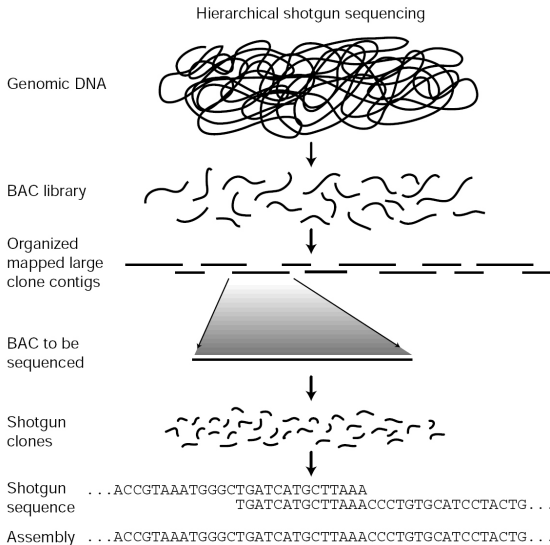- Indexed BEDgraphs/Wiggles
- VCFs (variants)

- Reference genomes describe the 'consensus' DNA sequence
- (Who is the human consensus for DNA sequencing?)
- Aside of human variation, multiple assemblies have been released

Hierarchical shotgun sequencing

Genomic DNA

BAC library

Organized
mapped large
clone contigs

BAC to be
sequenced

Shotgun
clones

Shotgun      ...ACCGTAAATGGGCTGATCATGCTTAAA
sequence                    TGATCATGCTTAAACCCTGTGCATCCTACTG...

Assembly  ...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

GRCh stands for 'Genome Reference Consortium'

- Human GRCh37 (hg19)
- Human GRCh38
- Mouse mm10
- Mouse GRCm38
- Zebrafish, chicken and others:
  https://www.ncbi.nlm.nih.gov/grc The Genome
  Reference consortium

# Reference genomes: FASTA

- A reference genome is a collection of contigs/scaffolds
- A contig is a stretch of DNA sequence encoded as A,G,C,T,N.
- Typically comes in FASTA format.
- ">" line contains the scaffold name
- Following lines contain the sequence (single line, 80 nt-column sized...)

```
>NC_009902.1 Babesia bovis T2Bo mitochondrion, complete genome
TTTAAAAAAGTGTTAAAAACTTTATACATTAAAAAATTTAAACAAGTGATCATGTATAAAGTACACTTGT
TACTGTGTAAATATCAAAAACAATTTAATTTCAAAATTTTTGAAATATGTTTTTTGTGTTGTGTTATAAA
GTTTTTTTTCAAAATTATATATGTTTGCATTTGCTGGATATAGTTCGGTCTCTGCAAACCATAAAGTCAT
CGGTATATCCTACATATGGCTTTCATATTGGTTTGGAGTTATTGGATTTTATATGAGTATTTTGATAAGA
ACAGAATTGAGTATGAGTGGTTTAAAGATTATGACAATGGATACTCTTGAGATATACAATATGATGTTTT
```

- Primary assembly: the best known assembly of a haploid genome
  - Chromosome assembly
  - Unlocalized sequence (associated to a chromosome but whose order/orientation is unknown)
  - Unplaced sequence (not linked to any chromosome)
- Alternate loci: An alternate representation of a locus (usually highly polimorphic regions, such as the MHC region)
- Patches: A contig sequence that is released outside of the full assembly release
  - Fix: error correction
  - Novel: new sequences that will be included into the next full assemblty release

- ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/
  001/405/GCA_000001405.27_GRCh38.p12/README_patch_
  release.txt

# Retrieving fasta sequences manually (UCSC)

- Try to retrieve the DIEXF gene promoter
- (What is a promoter in terms of sequence?)
- Go to an assembly https: //genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg38
- Query gene symbol (i.e. DIEXF)
- Click into the gene (gencode track)
- Click into the sequence and links item
- Specify your promoter definition

**Genomes**    **Genome Browser**    **Tools**    **Mirrors**    **Downloads**    **My Data**

**Genomic Sequence Near Gene**

## Get Genomic Sequence Near Gene

Note: if you would prefer to get DNA for more than one feature of this track at a time, try the
Table Browser using the output format sequence.

**Sequence Retrieval Region Options:**

☑ Promoter/Upstream by `1000` bases
☐ 5' UTR Exons
☐ CDS Exons
☐ 3' UTR Exons
☐ Introns
☐ Downstream by `1000` bases
◉ One FASTA record per gene.
○ One FASTA record per region (exon, intron, etc.) with `0` extra bases upstream (5')
and `0` extra downstream (3')
    ☐ Split UTR and CDS parts of an exon into separate FASTA records
Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream
bases are added, they may be truncated in order to avoid extending past the edge of the
chromosome.

```
>hg38_knownGene_uc001hhr.3 range=chr1:209827007-2098
gtttctgctgttgttaaatggggaatgctggaacagatttgtttgcgggg
actcttccaatactttcagaaaatgcgagaatagggtgagggtgggaatc
tcagacttgtgggccccatgattgatataacacacacaggcggcagaccc
taatgggtaaaagcatgtggttgcatcagttaaggtttttctctcttctc
ttgctagcgtgttatcttttcttttcttttcttcttttttttttttttcg
agatggagtctagcttttgtcgcccaggctggagtaggctggagtgcagt
ggagtgatctcggctcattgcaacctccacctcccgggttccagcgattc
tcctgcctcacctcctgagtagctgggattacaggcgcccgctaccacgc
ccggctgatttttgtactttagtagagacggggtttcaccatgtttggc
catgctggtctcgaactcctgacctcaggtgatccgcccatctcggcctc
ccaaagtgttgagattacaggcgtgagccaccgcgcccggccgctagcgt
gttatcttttctaagcatcagtttccttatctgcaacaccaggcttatta
acaagacctatctgtacactgttgtggtgatgaagtgagatgttcaggca
cccttaaatgttggttgatatttttattgcagtatactgtaaagtcactg
cattcgactatctccgctactacacatttacgcagactgatttccataac
caaaacacaagcacaaagctcatgcccccgactcacgcaacccgggaagc
ccagctgcccacgttctagggctctgagaacactagtgaacgaactcccg
tgctttcaaagagctgcggtaggggggcagaaccgggaaccggatgttcta
agcctgtcgtacgagcgcgacgtaaagcggatctgctttatggcaccttg
ctttcgccgtaaagcgcagtcagcgagcccacgtgcttgtgttgactgga
```

- Next step to FASTA: including quality data
- Standard de facto for high-throughput sequencing instruments such (i.e. Illumina)

- Sequence quality is represented using Phred scores
- The sequencing quality score of a given base Q is defined by as
- $Q = -10 \ \log_{10} P$

Phred quality scores are logarithmically linked to error probabilities

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
| --- | --- | --- |
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |
| 60 | 1 in 1,000,000 | 99.9999% |

# phred scores
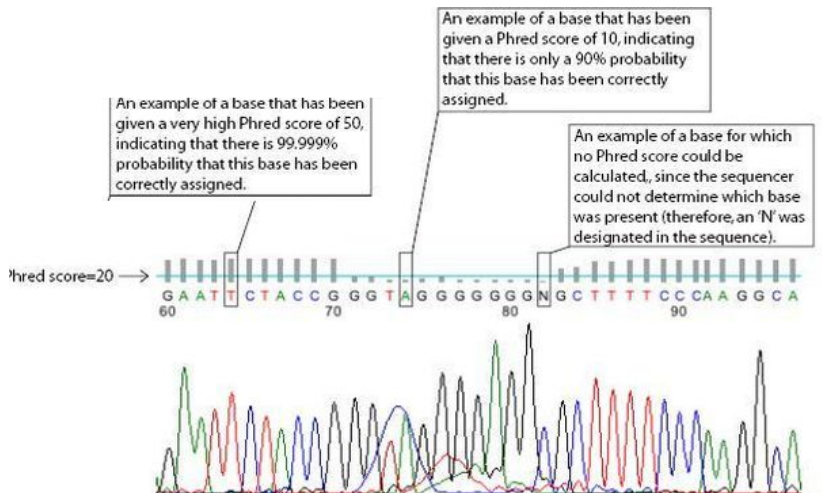
- FASTQs stands for FASTA with Qualities
- Plain text files with chunks of four lines:
  - @ identifier line
  - Sequence
  - "+"
  - Quality scores (different encodings exist)

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

- There are several Phred score encodings:
- https://wiki.bits.vib.be/index.php/Identify_the_Phred_scale_of_quality_scores_used_in_fastQ

```
## retrieving an example fastq file
curl  https://molb7621.github.io/workshop/_downloads/SP1.fq \
 > file.fastq

## counting number of reads
awk 'END{print NR/4}' file.fastq

## transforming into fasta
awk 'NR % 4 == 1 {print ">"$1};
     NR % 4 == 2 {print}' file.fasta
```

- https://en.wikipedia.org/wiki/AWK
- AWK essentials manual

- `awk 'ENDprint NR/4' file.fastq`
- NR gives the number of records (line numbers)
- FASTQ are chunks of 4 lines for each sequence
- NR/4 at the END of the file indicates the number of sequences

```
## retrieving an example fasta file
curl  https://molb7621.github.io/workshop/_downloads/SP1.fq \
 > file.fastq

## counting number of reads
awk 'END{print NR/4}' file.fastq

## transforming into fasta
awk 'NR%4==1{a=substr($0,2);} NR%4==2 {print ">"a"\n"$0}' \
  file.fastq
```

```
awk 'NR%4==1{a=substr($0,2);} NR%4==2 {print ">"a"\n"$0}'
file.fastq
```

- % is a modulo operator
- NR%4==1 will retrieve the first line of a fastq chunk (header)
- NR%4==2 will retrieve the second line (the sequence)
- the id line will be preprended with the $>$ and reduced to a substring (chopped)
- This will be applied to all the lines!

- Activity: FASTQ/A exercises (exercises 5 to 14)

# Alignment file format

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

- Sequence alignment: arrange a set of sequences to identify regions of similarity/identity
- Mapping short reads against a reference genome: aligning large amounts short reads to a reference genome

```
     1  2  3  4  5  6  7  8  9  10 11
     C  G  T  C  C  G  A  A  G  T  G
           |  .  |  |  |  |
     ⋆  ⋆  T  A  C  G  A  A  ⋆  ⋆  ⋆
```

(a) Global alignment

```
  3  4  5  6  7  8
  T  C  C  G  A  A
  |  .  |  |  |  |
  T  A  C  G  A  A
```

(b) Local alignment

```
  1  2  3  4  5  6  7  8
  C  G  T  C  C  G  A  A
        |  .  |  |  |  |
  ⋆  ⋆  T  A  C  G  A  A
```
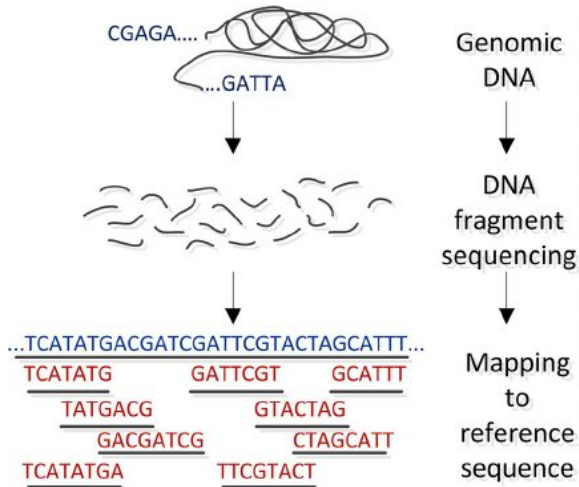
(c) Semi-global alignment

Alachiotis et al, 2013

- Chromosome
- Locus (coordinate)
- CIGAR string, i.e.
- 30M1D2M - 30 bases continuously match, 1 deletion from reference, 2 base match
- Flags (https://broadinstitute.github.io/picard/explain-flags.html)

Pavlopoulos et al 2013

Pavlopoulos et al 2013

Pavlopoulos et al 2013

- Activity: read the SAM format specification
- https:
  //www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/

- Exercises from no 15 on

- Read https://bedtools.readthedocs.io/en/latest/content/tools/genomecov.html

- BED (Browser Extensible Data)
- BED3: 3 tab separated columns, chromosome (scaffold), start, end
- BED6: BED3 plus name, score, strand

```
chr22 1000 5000
chr22 2000 6000
```

```
chr22 1000 5000 cloneA 960 +
chr22 2000 6000 cloneB 900 -
```

```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

- Which fraction of the human genome is covered by exons?
- http://pedagogix-tagc.univ-mrs.fr/courses/
  jgb53d-bd-prog/practicals/03_bedtools/

- Adding layers to genomic coordinates

Henrik Lantz, BILS/SciLifeLab

# What is genomic annotation?

| Seqid | source | type | start | end | score | strand | phase | attributes |
|-------|--------|------|-------|-----|-------|--------|-------|------------|
| Chr1 | Snap | gene | 234 | 3657 | . | + | . | ID=gene1; Name=Snap1; |
| Chr1 | Snap | mRNA | 234 | 3657 | . | + | . | ID=gene1.m1; Parent=gene1; |
| Chr1 | Snap | exon | 234 | 1543 | . | + | . | ID=gene1.m1.exon1; Parent=gene1.m1; |
| Chr1 | Snap | CDS | 577 | 1543 | . | + | 0 | ID=gene1.m1.CDS; Parent=gene1.m1; |
| Chr1 | Snap | exon | 1822 | 2674 | . | + | . | ID=gene1.m1.exon2; Parent=gene1.m1; |
| Chr1 | Snap | CDS | 1822 | 2674 | . | + | 2 | ID=gene1.m1.CDS; Parent=gene1.m1; |
| | | start_ codon | | | | | | Alias, note, ontology_term ... |
| | | stop_c odon | | | | | | |

Henrik Lantz, BILS/SciLifeLab

| Seqid | source | type | start | end | score | strand | phase | attributes |
|-------|--------|------|-------|-----|-------|--------|-------|------------|
| Chr1 | Snap | exon | 234 | 1543 | . | + | . | gene_id "gene1"; transcript_id "transcript1"; |
| Chr1 | Snap | CDS | 577 | 1543 | . | + | 0 | gene_id "gene1"; transcript_id "transcript1"; |
| Chr1 | Snap | exon | 1822 | 2674 | . | + | . | gene_id "gene1"; transcript_id "transcript1"; |
| Chr1 | Snap | CDS | 1822 | 2674 | . | + | 2 | gene_id "gene1"; transcript_id "transcript1"; |
| | | start_codon | | | | | | |
| | | stop_codon | | | | | | |

Henrik Lantz, BILS/SciLifeLab

Steven M. Carr

- http://mblab.wustl.edu/GTF22.html

# BEDgraph

```
chromA    chromStartA    chromEndA    dataValueA
chromB    chromStartB    chromEndB    dataValueB

chr19 49303800 49304100 0.50
chr19 49304100 49304400 0.75
chr19 49304400 49304700 1.00
```

# BEDgraphs

- To display continuous-valued data in track format.
- Uuseful for probability scores

- BED, BED, BED12?
- Advantages: the coordinates are specified, so sparsity is allowed

# Wig files

- To display continuos-value data
- GC percent, probability scores, and transcriptome data.
- Data is not sparse! Wiggle data elements must be equally sized (step)

# Wig

```
variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
```

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

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

# VCF

- Standard file format for storing variation data
- Unambiguous, scalable and flexible
- 8 columns:
    1. CHROM
    2. POS
    3. ID
    4. REF
    5. ALT
    6. QUAL
    7. FILTER
    8. INFO

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | NA19909 |
|--------|-----|-----|-----|-----|------|--------|------|--------|---------|
| 11 | 5248232 | rs334 | T | A | 100 | PASS | AA=T\|\|\|;AC=1;AF=0.0273562;AFR_AF=0.0998;AMR_AF=0.0072;AN=2;DP=22876;EAS_AF=0;EUR_AF=0;EX_TARGET;NS=2504;SAS_AF=0;VT=SNP | GT | 0\|1 |

EMBL/EBI training

- Phred-scaled quality score for the assertion made in ALT. i.e. $Q = -10 \, \log_{10} P$ being $P(\text{call in ALT is wrong})$
- Read quality
- Mapping quality
- Variant calling quality

# Variant calling

- Lecture by Michael Lawrence (VariantExplore package)
- https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/3_Wednesday/lectures/VariantCallingLecture.pdf