# BIO392 file formats and tools

Izaskun Mallona

# Talk typesetting

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

# Unix: recap

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

# Exercises

- Run the exercises till number 4

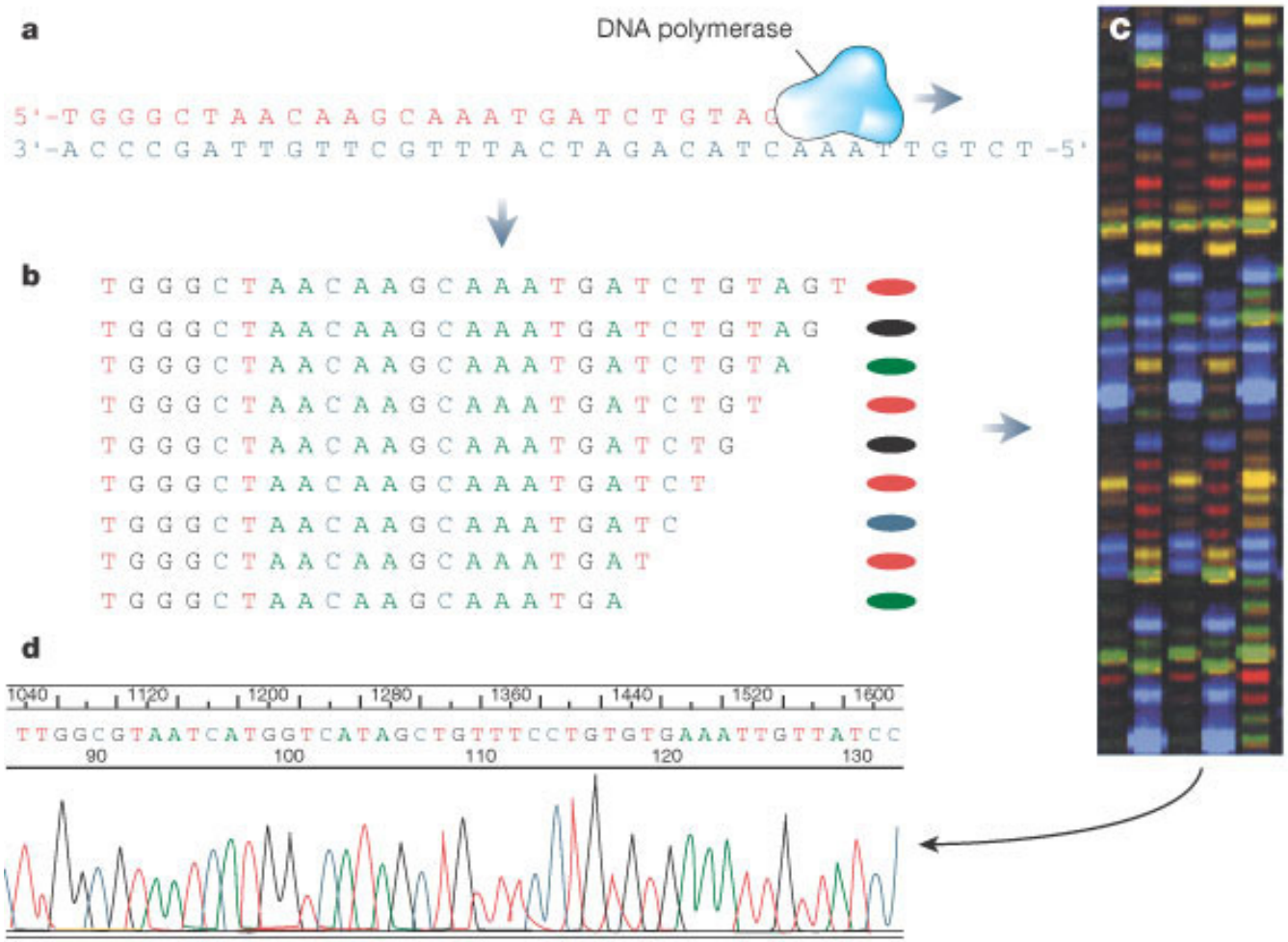
# Commonly used formats

- 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)

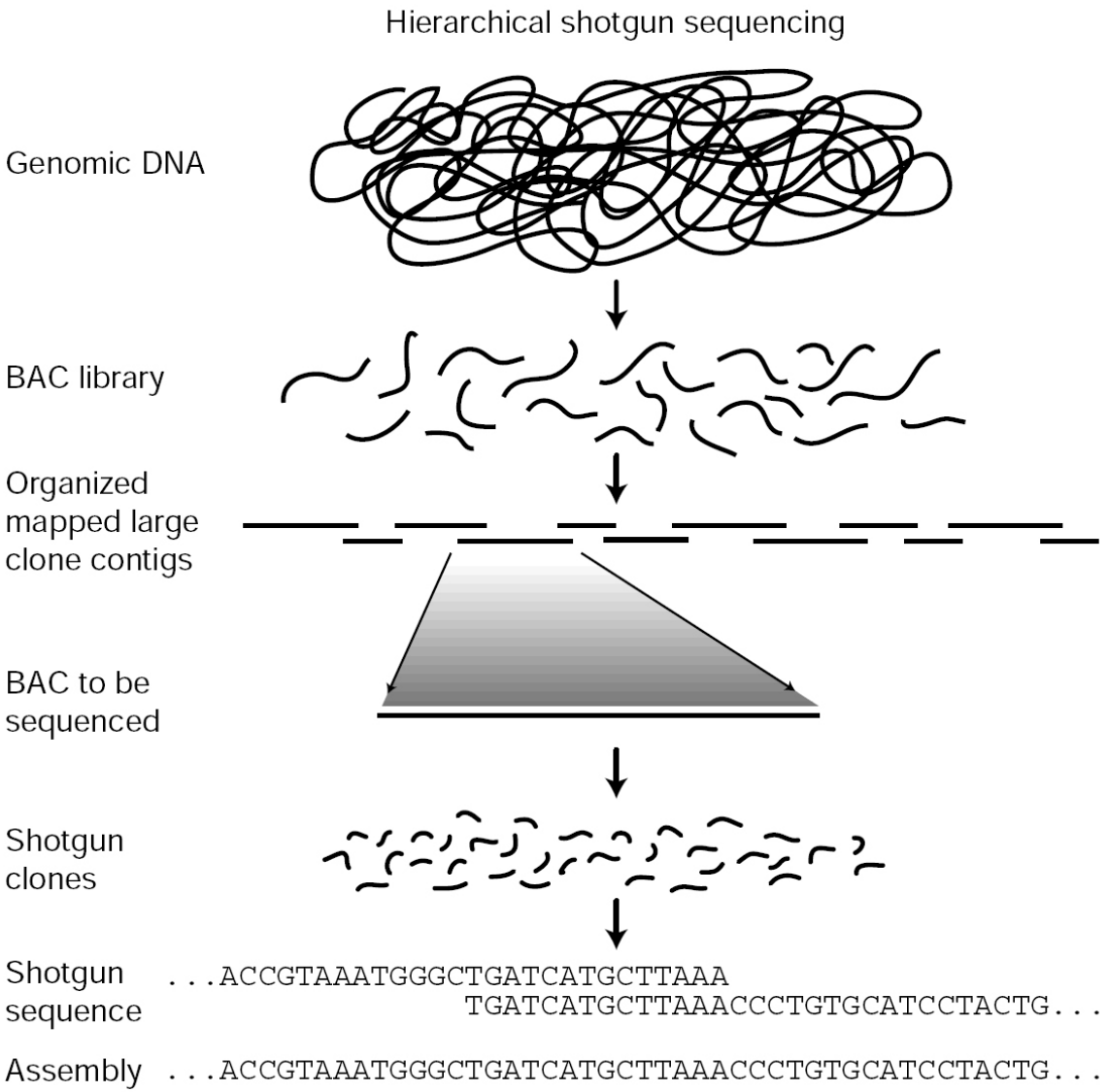
# Reference genomes

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

# Sanger sequencing Nature 409, 863 (2001)



# Hierarchical shotgun Nature 409, 863 (2001)



# Reference genomes

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/grcThe 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...)

# Reference genomes: FASTA

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

# Patches, alternate loci and primary assembly

- 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

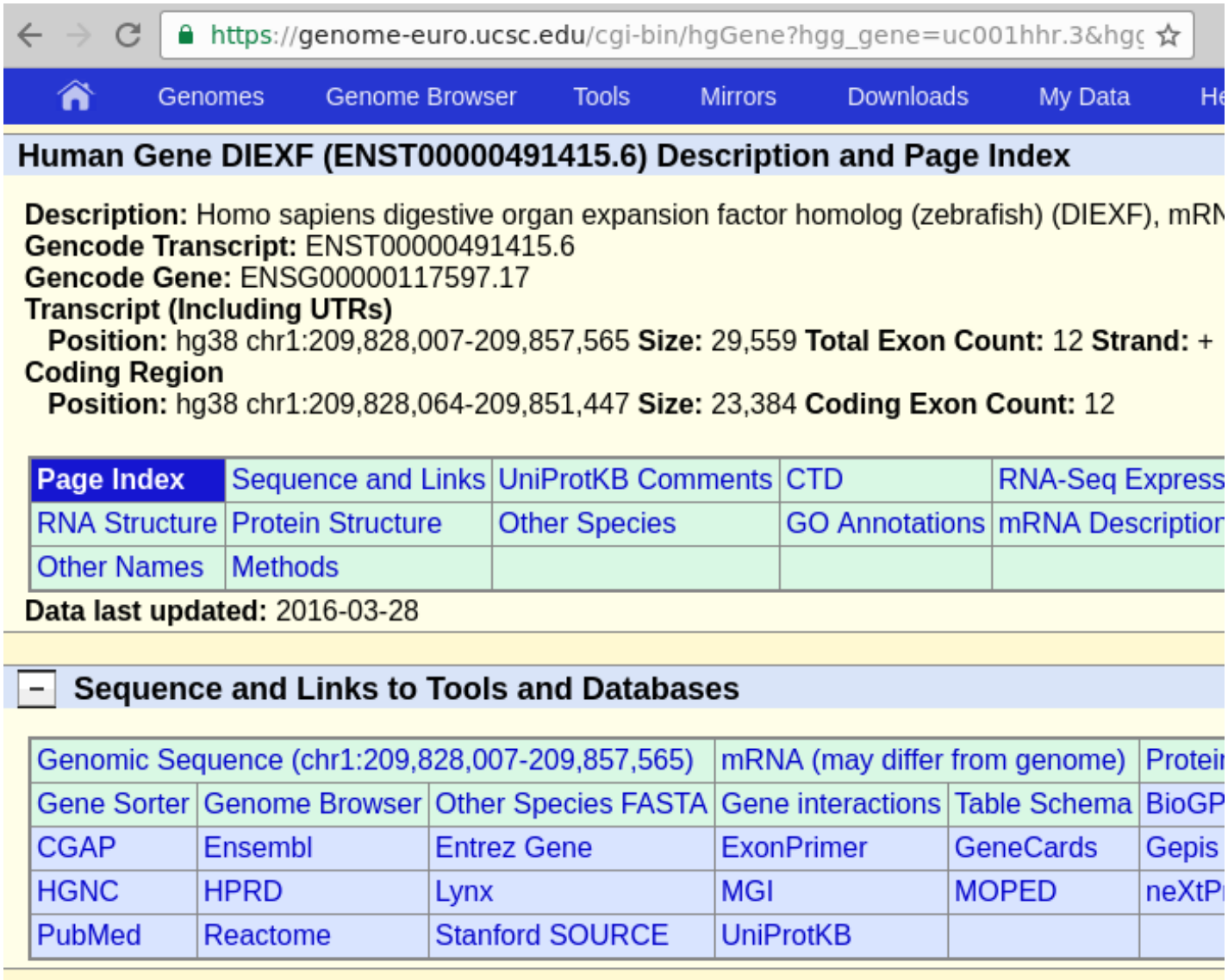
# Browsing genomic patches

- 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

# Manually downloading the DIEXF promoter



https://genome-euro.ucsc.edu/cgi-bin/hgGene?hgg_gene=uc001hhr.3&hgc

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | He

**Human Gene DIEXF (ENST00000491415.6) Description and Page Index**

**Description:** Homo sapiens digestive organ expansion factor homolog (zebrafish) (DIEXF), mRN
**Gencode Transcript:** ENST00000491415.6
**Gencode Gene:** ENSG00000117597.17
**Transcript (Including UTRs)**
**Position:** hg38 chr1:209,828,007-209,857,565 **Size:** 29,559 **Total Exon Count:** 12 **Strand:** +
**Coding Region**
**Position:** hg38 chr1:209,828,064-209,851,447 **Size:** 23,384 **Coding Exon Count:** 12

| Page Index | Sequence and Links | UniProtKB Comments | CTD | RNA-Seq Express |
|---|---|---|---|---|
| RNA Structure | Protein Structure | Other Species | GO Annotations | mRNA Descriptior |
| Other Names | Methods | | | |

**Data last updated:** 2016-03-28

**Sequence and Links to Tools and Databases**

| Genomic Sequence (chr1:209,828,007-209,857,565) | | | mRNA (may differ from genome) | | Protei |
|---|---|---|---|---|---|
| Gene Sorter | Genome Browser | Other Species FASTA | Gene interactions | Table Schema | BioGP |
| CGAP | Ensembl | Entrez Gene | ExonPrimer | GeneCards | Gepis |
| HGNC | HPRD | Lynx | MGI | MOPED | neXtP |
| PubMed | Reactome | Stanford SOURCE | UniProtKB | | |

# Manually downloading the DIEXF promoter

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:

- [x] 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/downstrea bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

# Manually downloading the DIEXF promoter

```
>hg38_knownGene_uc001hhr.3 range=chr1:209827007-209
gtttctgctgttgttaaatggggaatgctggaacagatttgtttgcgggg
actcttccaatactttcagaaaatgcgagaatagggtgagggtgggaatc
tcagacttgtgggccccatgattgatataacacacacaggcggcagaccc
taatgggtaaaagcatgtggttgcatcagttaaggtttttctctcttctc
ttgctagcgtgttatctttttctttttcttttcttttttttttttttcg
agatggagtctagcttttgtcgcccaggctggagtaggctggagtgcagt
ggagtgatctcggctcattgcaacctccacctcccgggttccagcgattc
tcctgcctcacctcctgagtagctgggattacaggcgcccgctaccacgc
ccggctgatttttgtacttttagtagagacggggtttcaccatgtttggc
catgctggtctcgaactcctgacctcaggtgatccgcccatctcggcctc
ccaaagtgttgagattacaggcgtgagccaccgcgcccggccgctagcgt
gttatcttttctaagcatcagtttccttatctgcaacaccaggcttatta
acaagacctatctgtacactgttgtggtgatgaagtgagatgttcaggca
cccttaaatgttggttgatattttattgcagtatactgtaaagtcactg
cattcgactatctccgctactacacatttacgcagactgatttccataac
caaaacacaagcacaaagctcatgcccccgactcacgcaacccgggaagc
ccagctgcccacgttctagggctctgagaacactagtgaacgaactcccg
tgctttcaaagagctgcggtaggggcagaaccgggaaccggatgttcta
agcctgtcgtacgagcgcgacgtaaagcggatctgctttatggcaccttg
ctttcgccgtaaagcgcagtcagcgagcccacgtgcttgtgttgactgga
```

# How do we do this in a reproducible manner?

- Querying an API
- Scripting
  - Storing an up-to-date reference genome in our computer (once) and using specific file standards to specify the genome annotation (i.e. GTF, BED files)

# Commonly used formats

- Reference genomes
- **Fasta and FastQ**
- 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)

# FASTQ: Short read sequencing

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

# phred scores

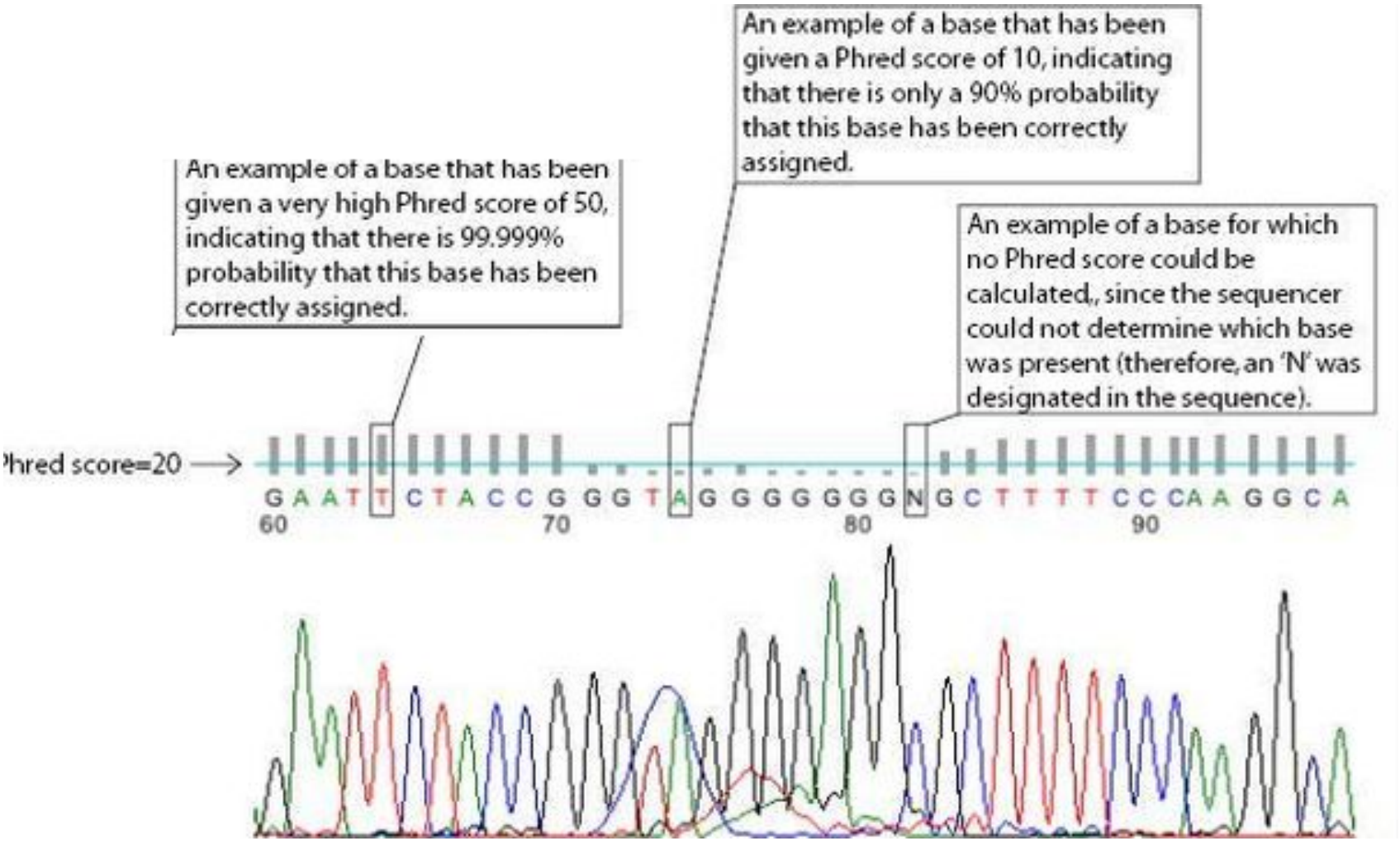
- Sequence quality is represented using Phred scores
- The sequencing quality score of a given base Q is defined by as
- $Q = -10 \ \log_{10} P$

# phred scores

**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



# Unaligned sequences (from sequencers): FASTQ

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

# Example FASTQ entry

```
@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
```

# Phred scores encoding

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

# Working with fastq files

```
## 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
```

# Further manuals on awk

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

# awk: Counting the number of items in a fastq

```
awk 'END{print 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

# Working with fastq files

```
## 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: fastq to fasta

```
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!

# In class exercises for fastq

- 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

# What is an alignment?

- 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

# Local alignments vs global alignment


(a) Global alignment


(b) Local alignment

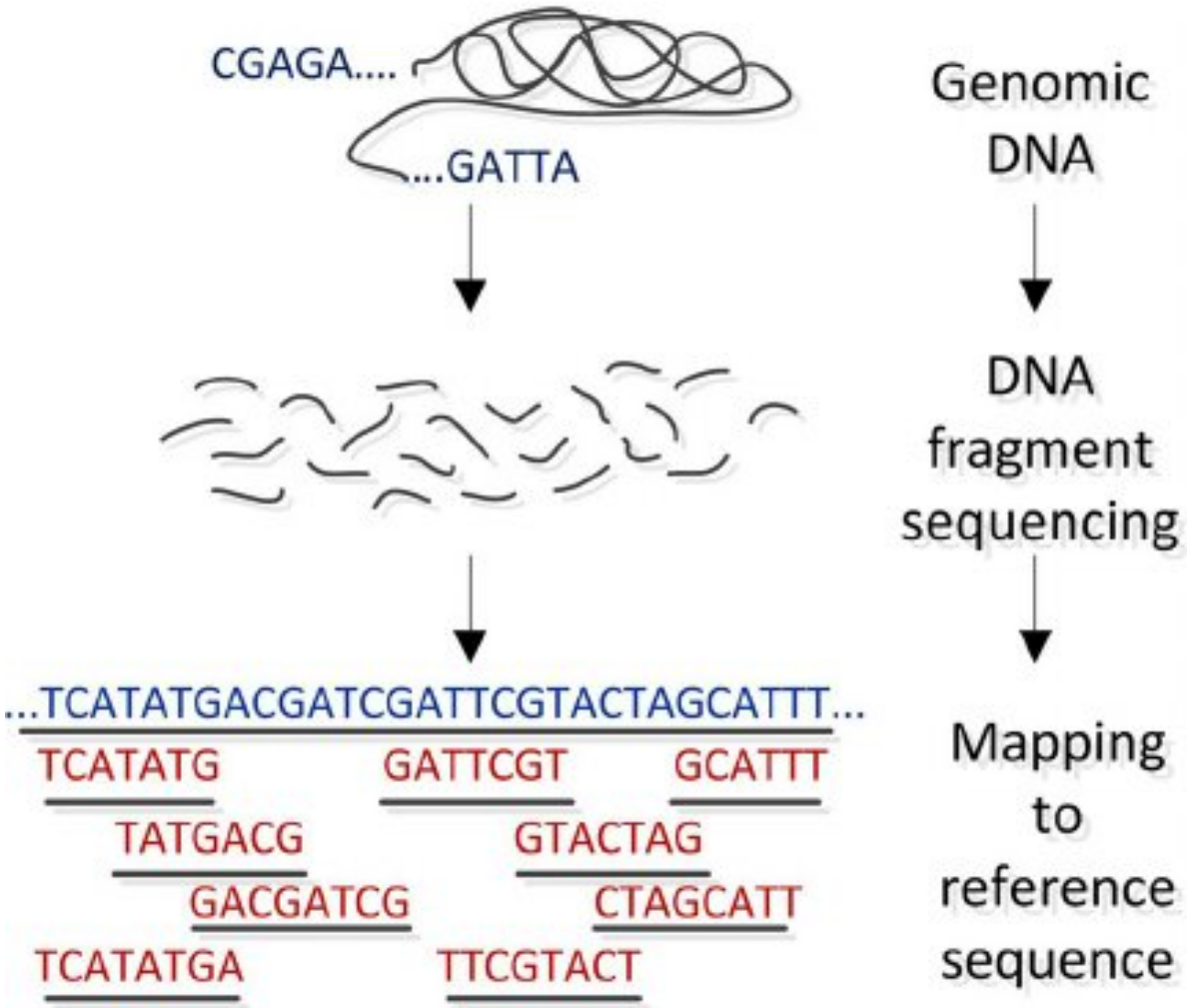

(c) Semi-global alignment

Alachiotis et al, 2013

# SAM format

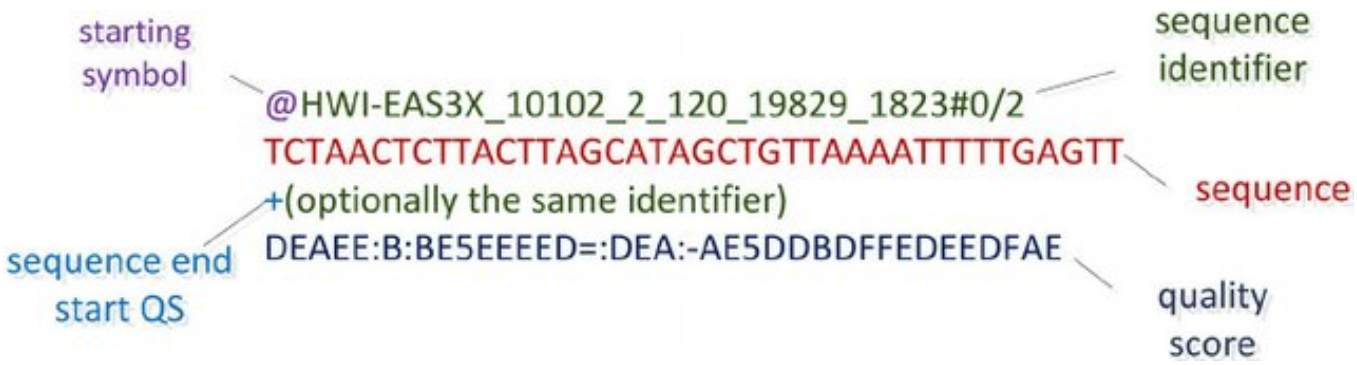
- Chromosome
- Locus (coordinate)
- CIGAR string, i.e.
- 30M1D2M - 30 bases continuously match, 1 deletion from reference, 2 base match
- Flags (https://broadinstitute.github.io/picard/explain-flags.html)

# Next generation sequencing to SAM



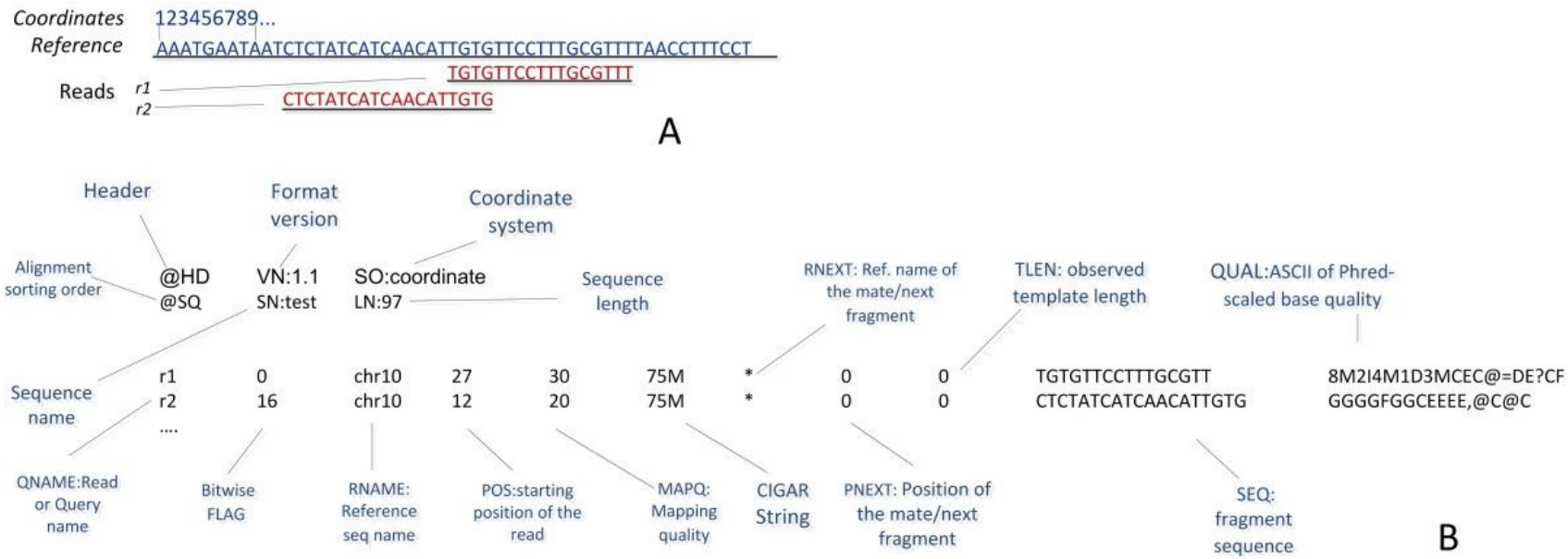

Pavlopoulos et al 2013

# Next generation sequencing to SAM




Pavlopoulos et al 2013

# Next generation sequencing to SAM




Pavlopoulos et al 2013

# SAM format

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

# Exercises on SAM files

- Exercise number 15

# From SAM to BED: counts

- BED files are simpler data representations, usually the next step after getting the SAM files
- For instance, after mapping a new genome-wide sequencing BED files with the genomic coverages are generated
- Read https://bedtools.readthedocs.io/en/latest/content/tools/genomecov.html
- Discussion: how to handle expression data, i.e. transcripts without introns etc? how do we count them?

# Keep it simple: count and transform into BED files

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

# BED3

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

# BED6

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

# BED12

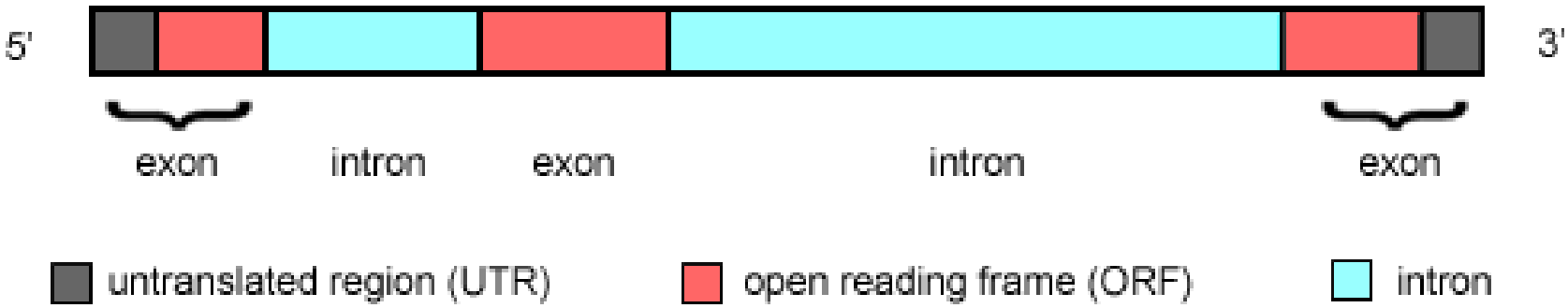
```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

# The need for further data formats

- So to sum up until now generally we have a reference genome, reads that were retrieved as FASTQ files, mapped and transformed to SAM files
- (Which step was the last one with actual sequences on them?)
- So, at last, we can answer questions
- Which fraction of the human genome is covered by exons?
- Genomic locations of SNPs associated with prostate cancer?
- Are gene bodies more variable (in terms of SNPs) than intergenic regions?
- For this we need further data analysis tools (i.e. BEDtools) on BED files and other data formats (annotations and variation)

# What is genomic annotation?

- Adding layers to genomic coordinates




Henrik Lantz, BILS/SciLifeLab

# What is genomic annotation? GFF3

| 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_codon | | | | | | |

Henrik Lantz, BILS/SciLifeLab

# What is genomic annotation? GTF

| 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

# Open reading frames

# GTF and GFF3 fileformats

- Reading: https://www.ensembl.org/info/website/upload/gff.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
```

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

# Which are the differences between BEDgraphs and BED?

- 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
```

# Wig with added span

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

# Wig with fixedStep and span

```
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

# Quality values: which one?

- Phred-scaled quality score for the assertion made in ALT. i.e. $Q = -10 \log_{10} P$ being $P$(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