Data standards in genomics

Izaskun Mallona

COST Project Epichembio - Introduction to NGS data analysis

13th March 2019

Acknowledgements

- EU Horizon 2020 COST Project Epichembio
- IJC Carreras Foundation
- Organizers: Sarah, Marguerite-Marie, David, Roberto
- SIB Swiss Institute of Bioinformatics and Univ. Zurich

Notes

- Talk typesetting
  - Commands/options are in `typewriter font`
  - URLs are highlighted in blue
- Exercises
  - Available at the course GitHub repo
  - Please use ad libitum (caution: there are 38 of them, exceeding the workshop workload)

Commonly used formats

Meant to provide an usable information representation for each NGS processing data step

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

Reference genomes: FASTA

- Reference genomes describe the 'consensus' DNA sequence
- 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 (edited)
TTTAAAAAAGTGTTAAAAACTTTATACATTAAAAAATTTAAACAAGTGATCATGTATAAA
TACTGTGTAAATATCAAAAACAATTTAATTTCAAAATTTTTGAAATATGTTTTTTGTGTT
GTTTTTTTTCAAAATTATATATGTTTGCATTTGCTGGATATAGTTCGGTCTCTGCAAACC
CGGTATATCCTACATATGGCTTTCATATTGGTTTGGAGTTATTGGATTTTATATGAGTAT
ACAGAATTGAGTATGAGTGGTTTAAAGATTATGACAATGGATACTCTTGAGATATACAAT
```

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

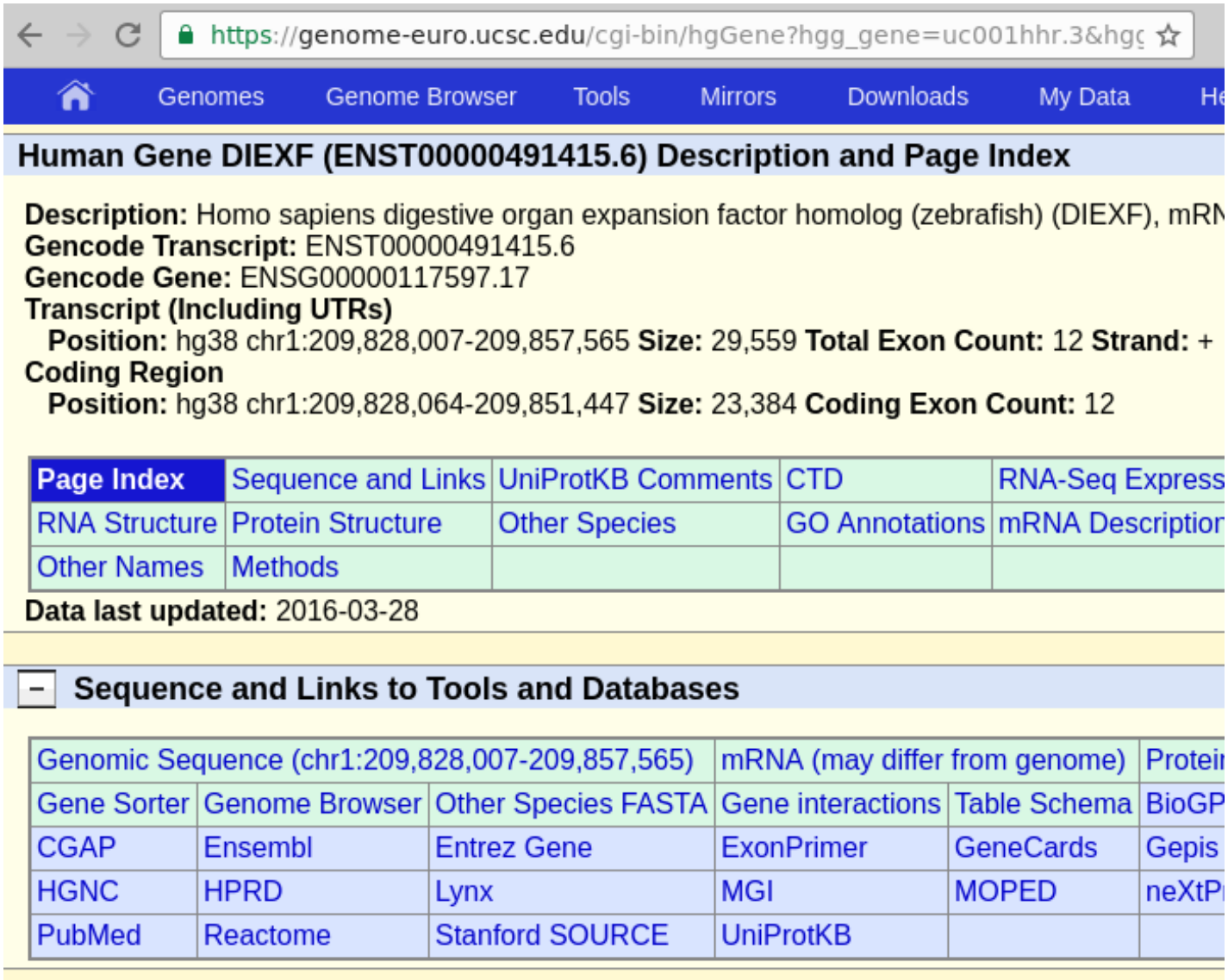
Browsing genomic patches

- Activity: browse genomic patches, i.e. 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

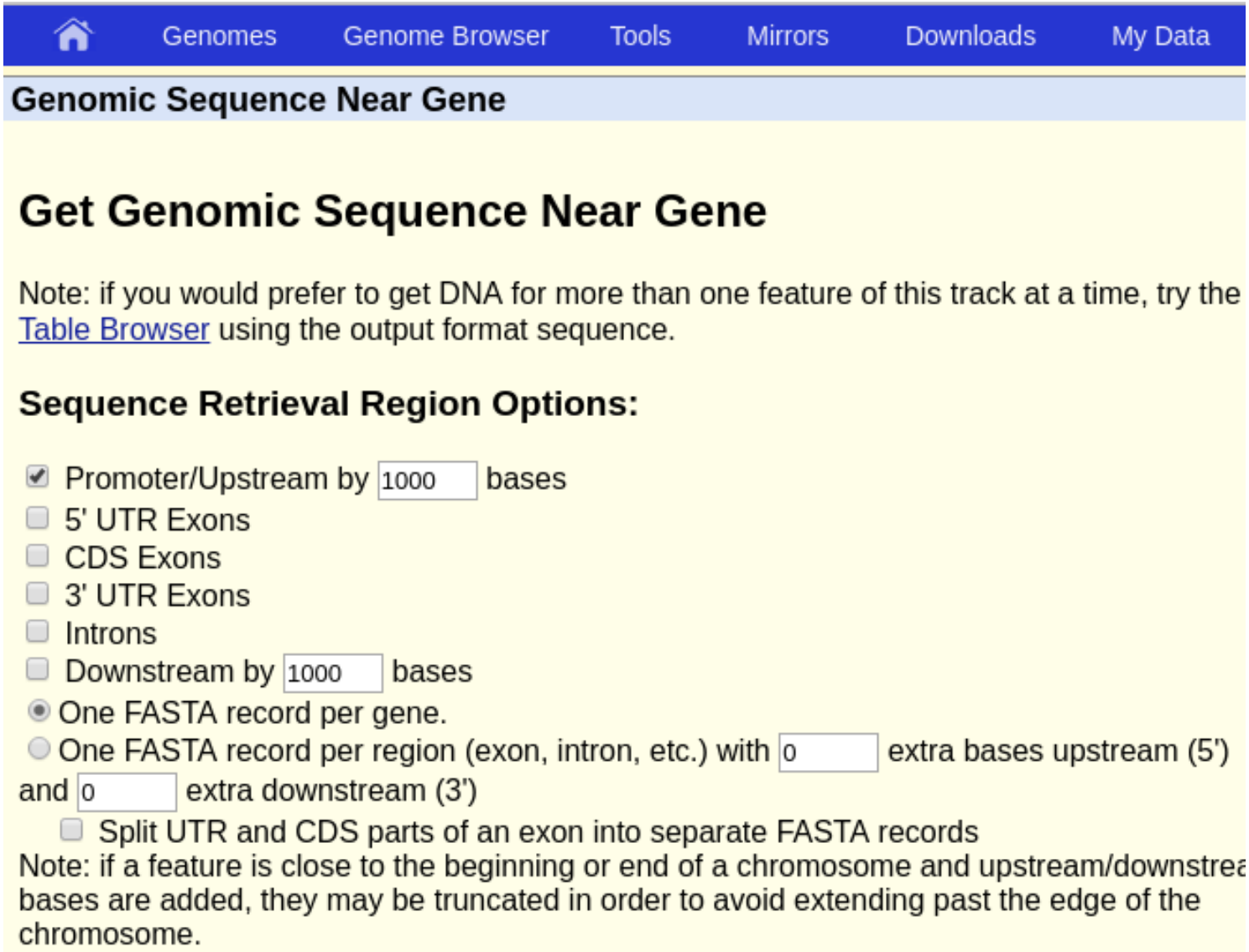
Page Index	Sequence and Links	UniProtKB Comments	CTD	RNA-Seq Express
RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptior
Other Names	Methods			

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



Manually downloading the DIEXF promoter

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

How do we do this in a reproducible manner?

- Scripting. We store an up-to-date reference genome in our computer (once)...
- ... and then use specific file standards to specify the genome annotation (i.e. GTF, BED files).
- Activity: read the documentation of UCSC on how to download sequences http://genome.ucsc.edu/FAQ/FAQdownloads.html (section Extracting sequence in batch from an assembly)

How?

Same concept that when we did manually:

1. Download the human genome sequence
2. Download a file with all the genes (transcripts) locations (not sequences, but their coordinates)
3. Then select the gene we are looking for (DIEXF)
4. Decide what a promoter is (i.e. 2 kb upstream of the gene) and update the coordinates accordingly
5. Then use a specific tool to slice the full genome to only report the DIEXF promoter

How? using data standards

Same concept that when we did manually + some standards

1. Download the human genome sequence (fasta)
2. Download a file with all the genes (transcripts) locations (GTF)
3. Then select the gene we are looking for (DIEXF) (grep/awk)
4. Decide what a promoter is (i.e. 2 kb upstream of the gene) and update the coordinates accordingly (BEDfile, bedtools/awk)
5. Then use a specific tool to slice the full genome to only report the DIEXF promoter (bedtools)

Standards

- This we can do because the genome consortia and the science community released open, free data and software/toolsets
- To handle them we benefit from the Unix-like operating systems
- We still need to use the same lingua francah: the need for data standards
  - Open
  - Efficent
  - Structured

Commonly used formats

- **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)

Short reads sequencing

- Sequencing very short reads (50 to 150 nucleotides) is common practice
- We get hundreds of millions of short reads for each experiment
- Instead of assembling them, we map them into a reference genome
- Activity: read https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2836519/
- Sequencers provide sequence and error rates assessment: `fasta` format is not suitable, but `fastq` is

FASTQ: Short read sequencing

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

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

phred scores

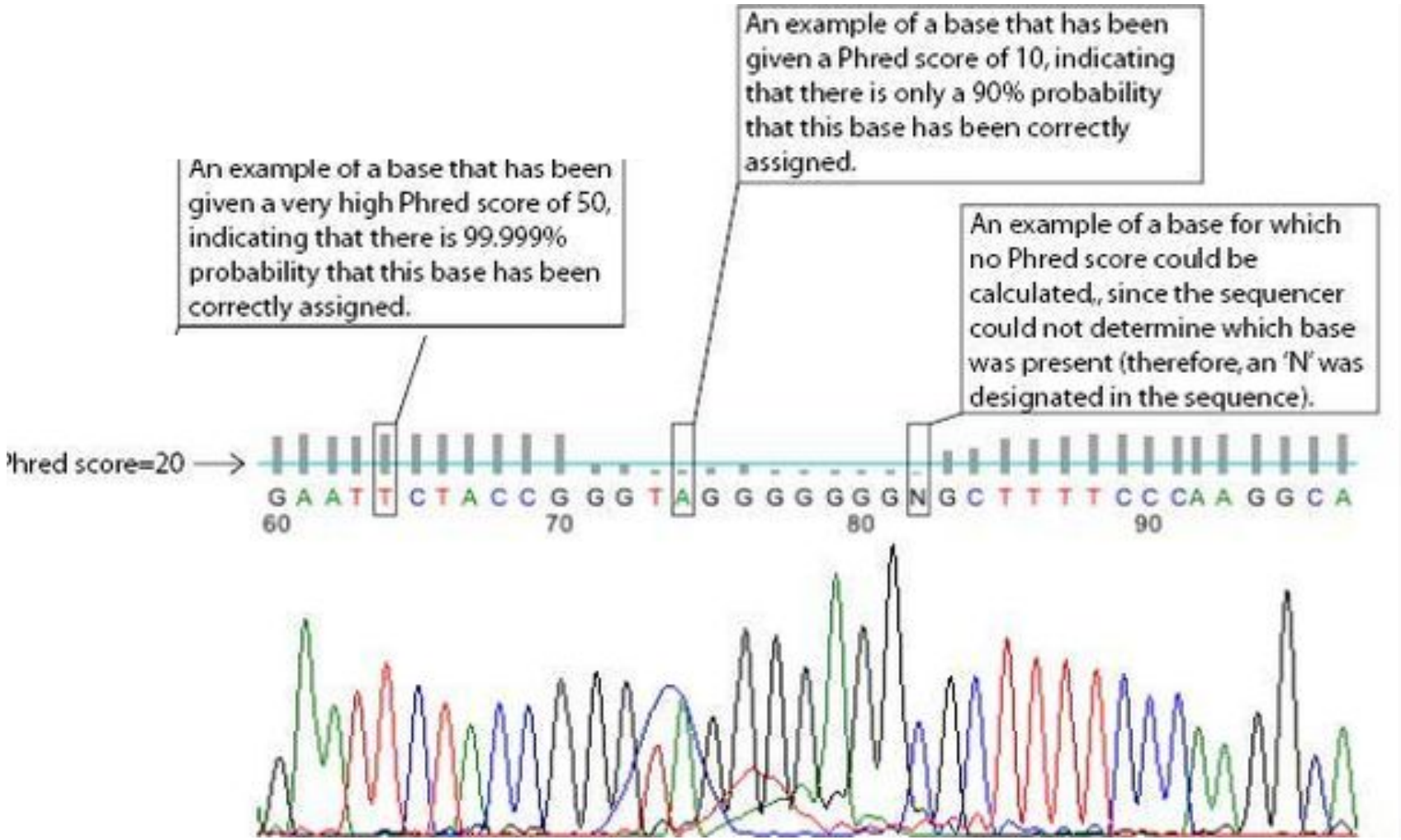
- Sequence quality is represented using Phred scores
- The sequencing quality score of a given base Q is defined by as
- $Q = -10 \ \log_{10} P$

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%

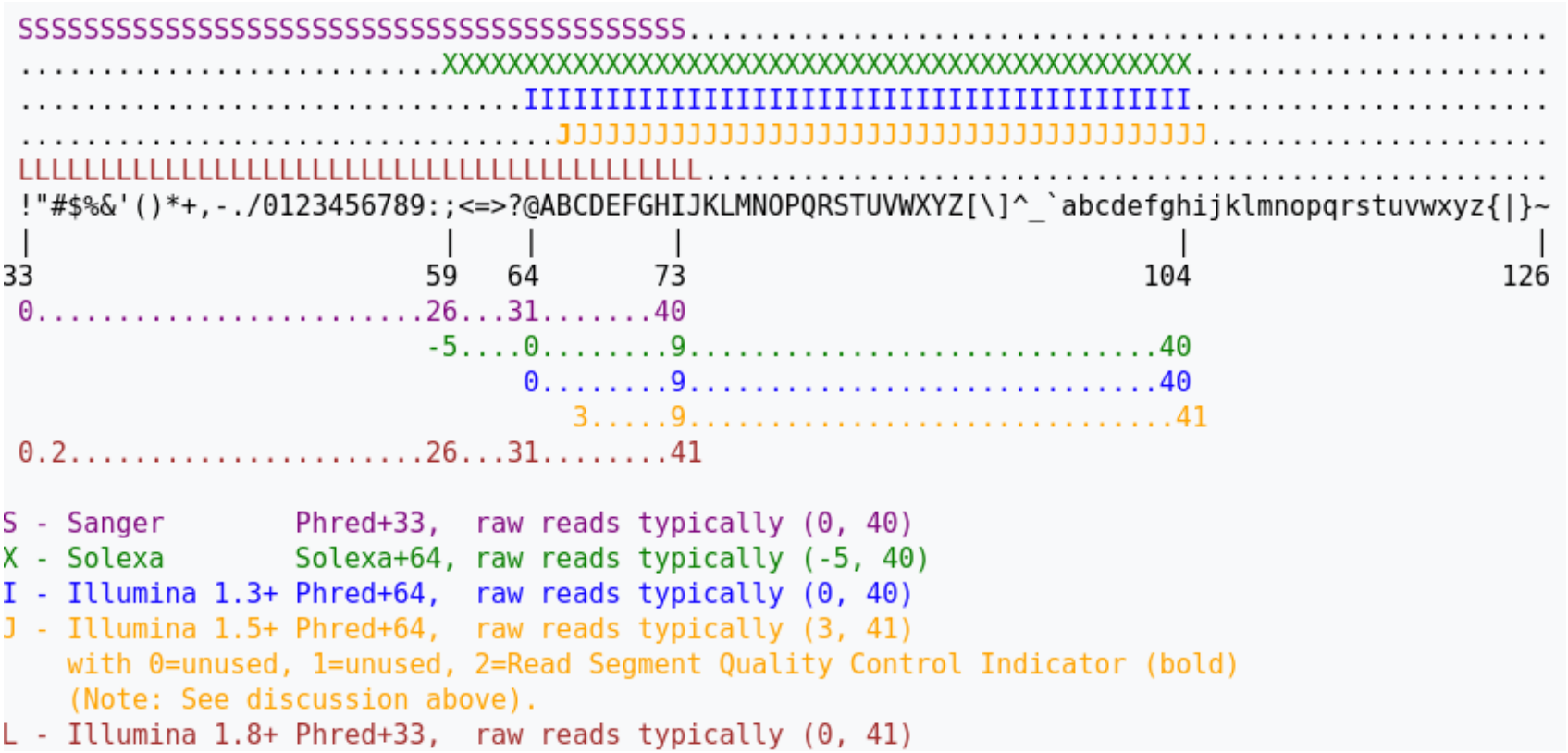
phred scores (old school Sanger electrophoretogram)



Phred scores encoding

- There are several Phred score encodings:
- Activity: read about the Quality scores offsets at https://en.wikipedia.org/wiki/FASTQ_format and https://wiki.bits.vib.be/index.php/Identify_the_Phred_scale_of_quality_scores_used_in_fastQ

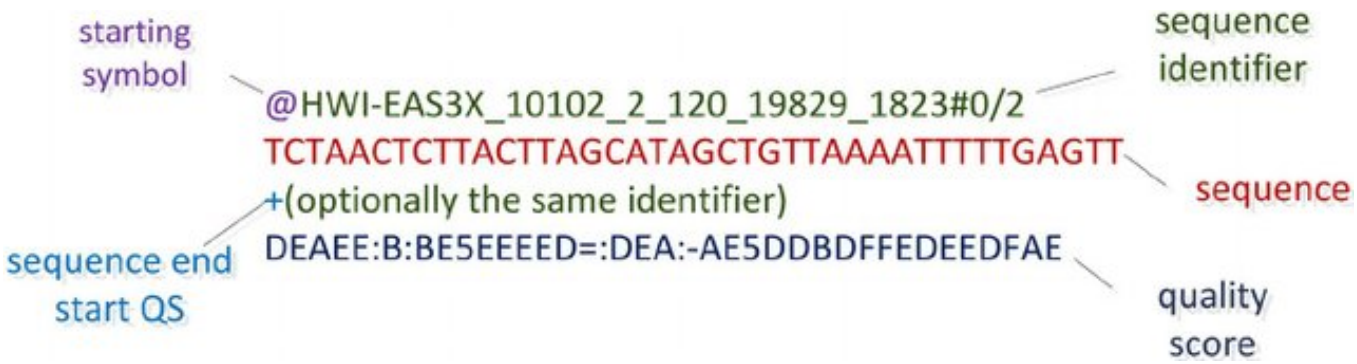
Phred scores encoding (Wikipedia)



Unaligned sequences (from sequencers): FASTQ

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

Example FASTQ entry



Pavlopoulos et al 2013

In class exercises for fastq

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

awk: Counting the number of items in a fastq

So in fastq each data chunk is stored in four different lines. We'll need to be able to extract the first, second, third of fourth line for each block of four lines. Using awk,

```
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 lines!

Still need to align the FASTQ reads to the reference genome

- Discussion: how to get rid of the sequences and to have a smaller data representation?
- Trying to transform sequence to reference genome coordinates (= aligning to the genome/mapping)
- i.e. transforming `ACGCACGCACGCACGCCCC` to human genome hg19 ‘chr10:10010-10030‘

Alignment file format

- SAM - Sequence Alignment Map.
- The standard stores where the reads (i.e. the ones we had as FASTQs) map in the reference genome
- 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

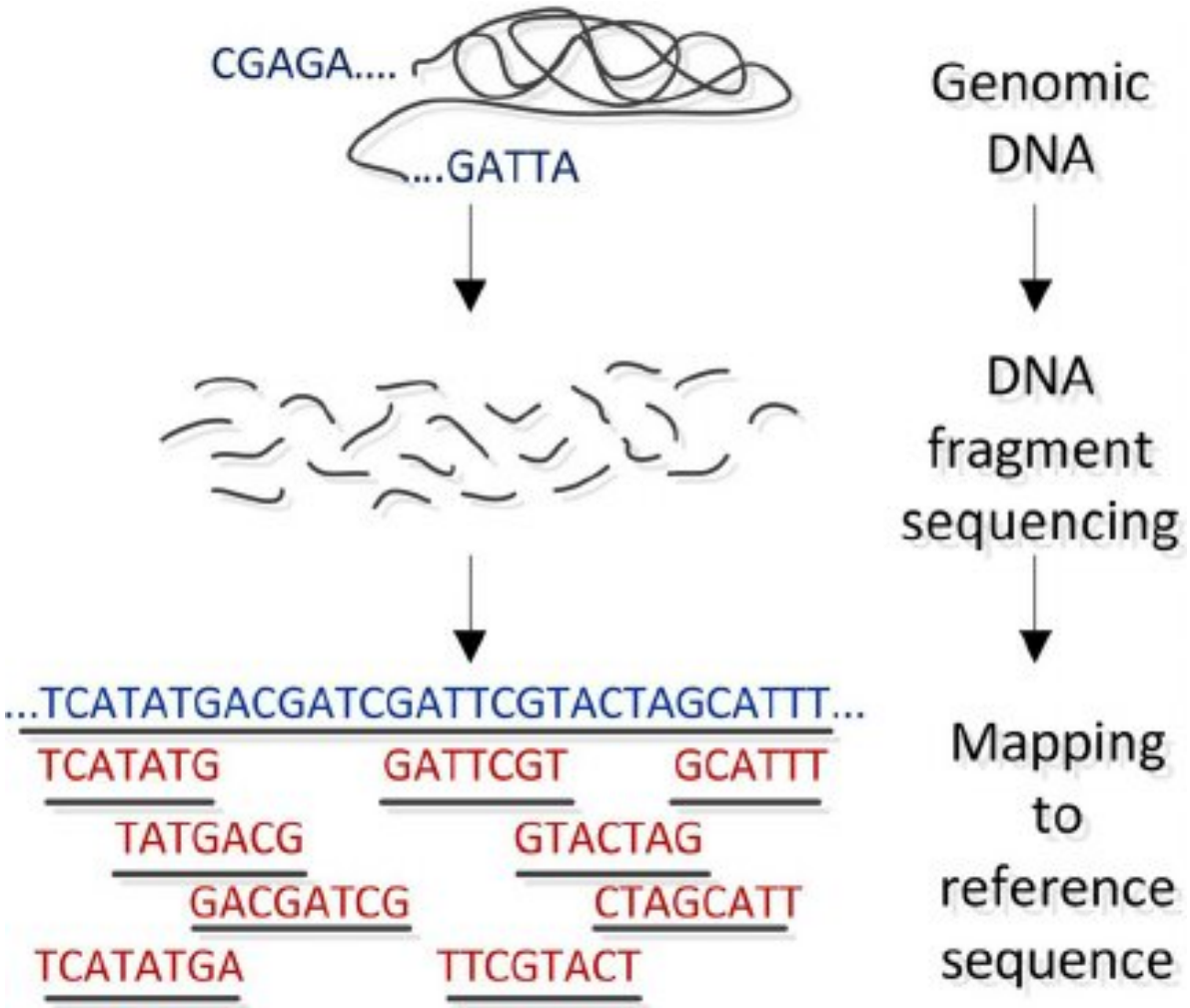
SAM format

- Chromosome
- Locus (coordinate)
- CIGAR string, i.e.
- 30M1D2M - 30 bases match (actually can be a mismatch, but present in the reference), 1 deletion from reference, 2 base match
- Some flags (https://broadinstitute.github.io/picard/explain-flags.html)

CIGAR strings

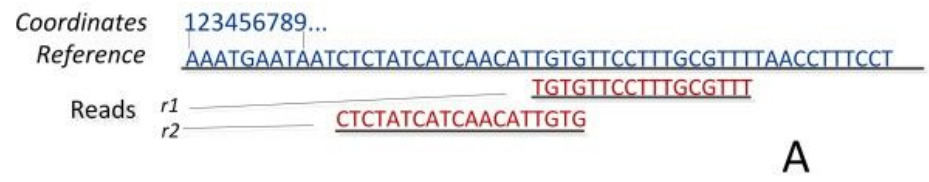
- Activity: read a post on CIGAR encoding http://bioinformatics.cvr.ac.uk/blog/tag/cigar-string/ (please skip the Java code)

Aligning NGS reads to a reference

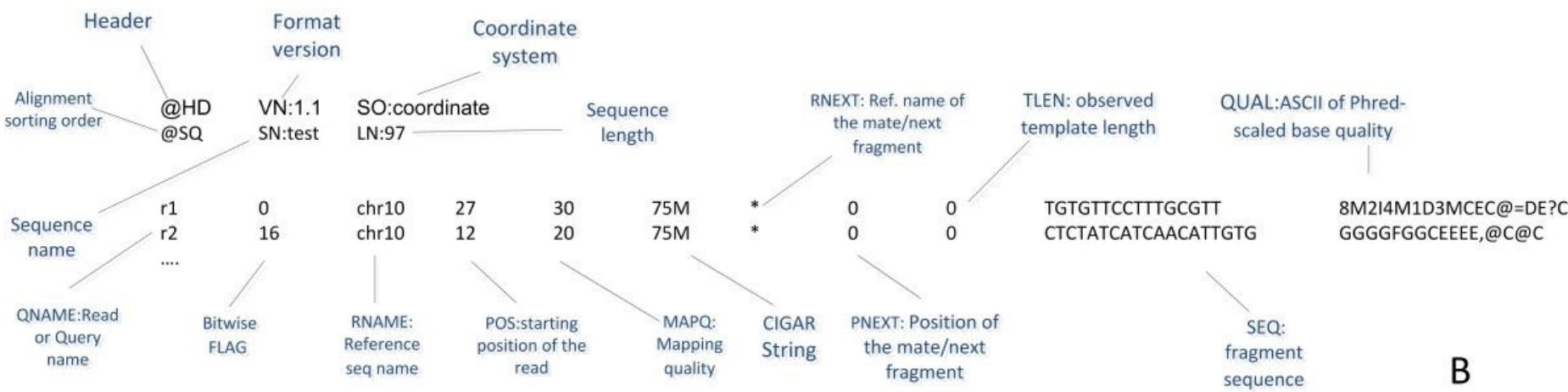


Pavlopoulos et al 2013

Next generation sequencing to SAM



A



B

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
- Why? they are smaller and easier to handle
- For instance, after mapping a new genome-wide sequencing BED files with the genomic coverages are generated
- Discussion: how to handle expression data, i.e. transcripts without introns etc? how do we count them?
- Activity: read https://bedtools.readthedocs.io/en/latest/content/tools/genomecov.html

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 -
```

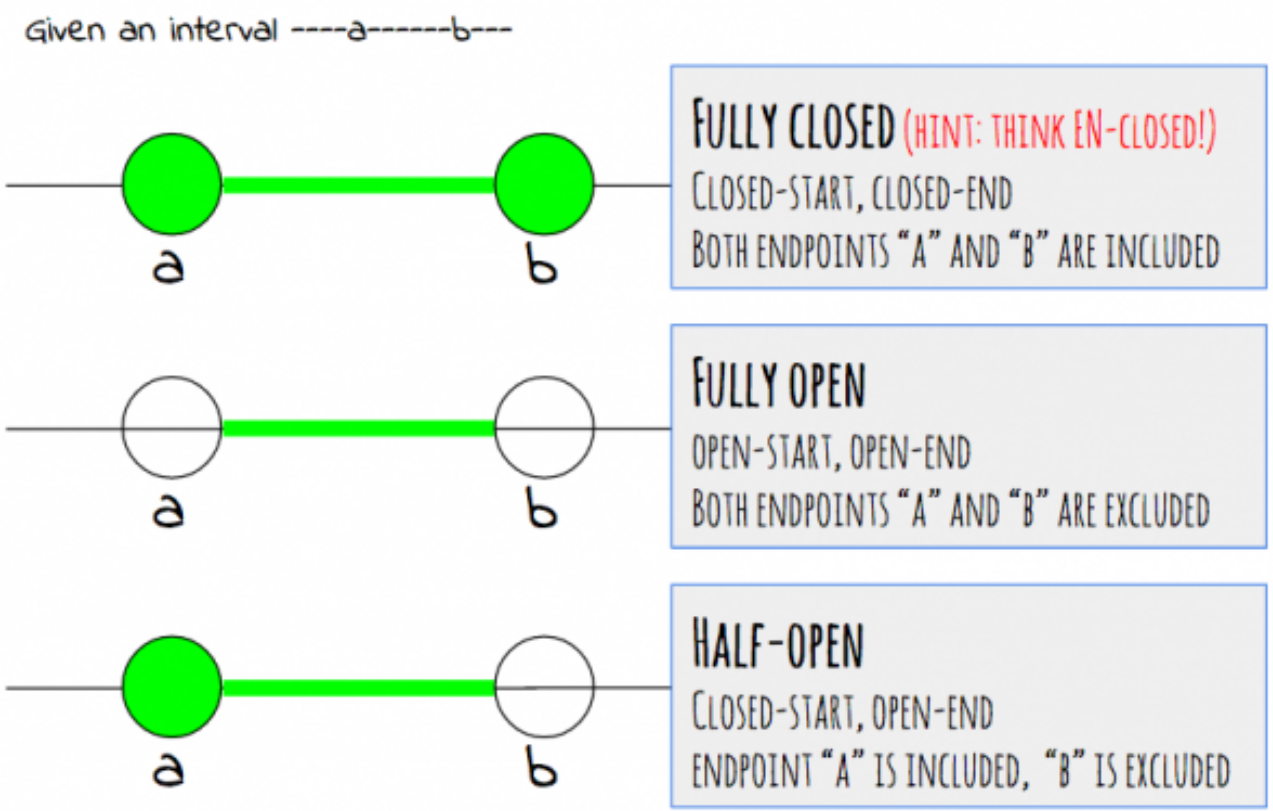
BED12

```
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
```

How do we count? 0s and 1s

- Even though BED files are standard how to count nucleotides is not
- 0-start vs. 1-start : Does counting start at 0 or 1?
- For a counted range, is the specified interval fully-open, fully-closed, or a hybrid-interval (e.g., half-open)?

On coordinates, 0s vs 1s and open and closed intervals



http://genome.ucsc.edu/blog/the-ucsc-genome-browser-coordinate-counting-systems/

UCSC Genome Browser web interface = 1-start, fully-closed

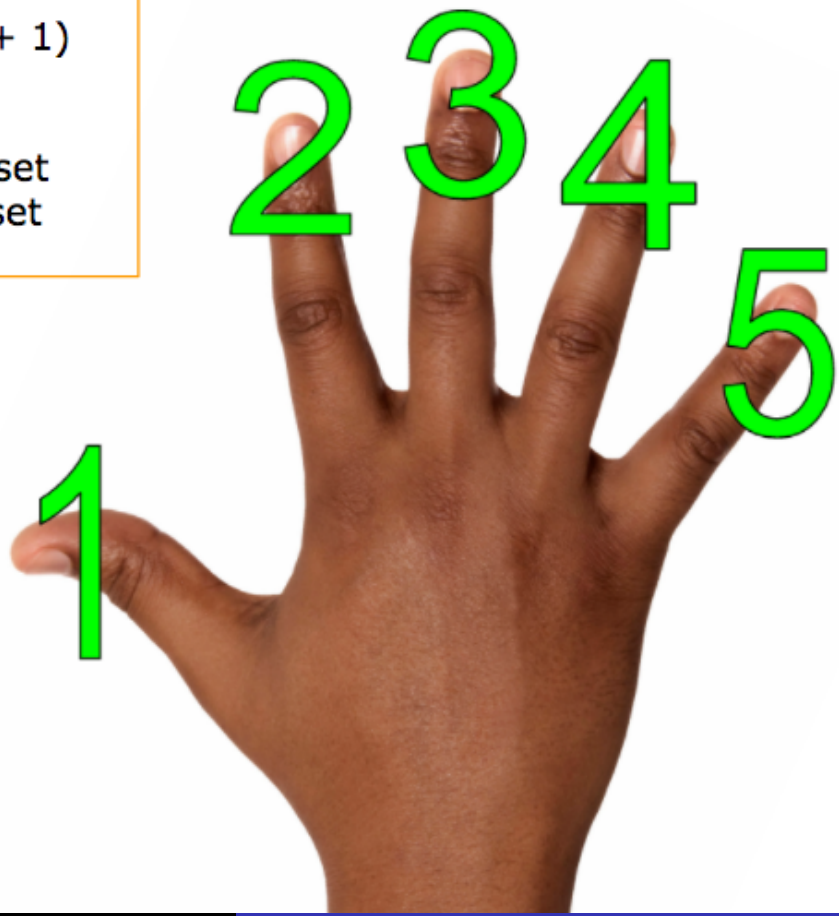
1-Start, Fully-Closed
closed-start (included)
closed-end (included)

Range= 1-5

Size = Stop - Start (+ 1)
5= 5 - 1 (+1)

Start (1) included in set
Stop (5) included in set

Coordinates positioned within the UCSC Genome Browser Web Interface



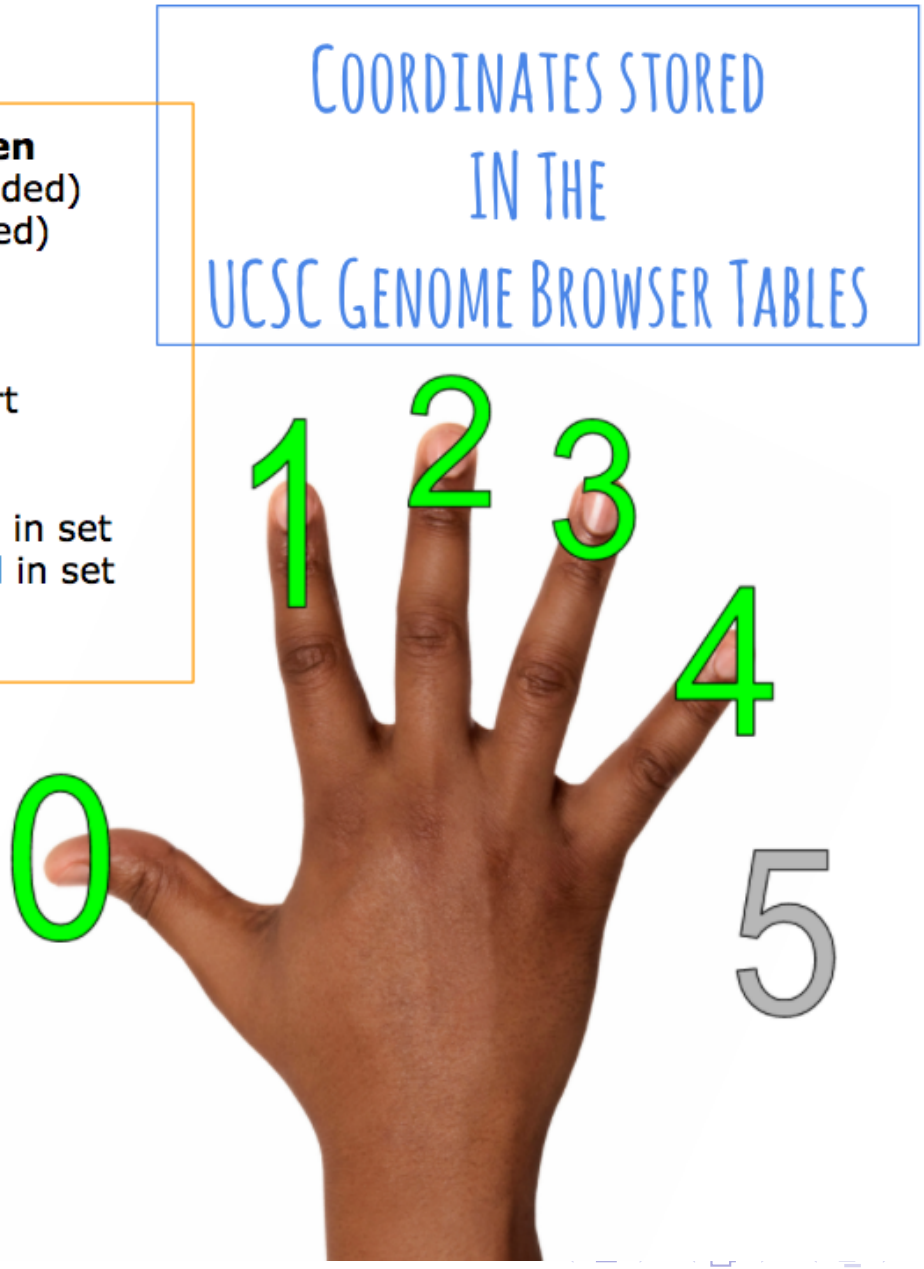
UCSC Genome Browser tables = 0-start, half-open

0-Start, half-open
closed-start (included)
open-end (excluded)

Range= 0-5

Size = Stop - Start
5 = 5 - 0

Start (0) **in**cluded in set
Stop (5) **ex**cluded in set



Activities

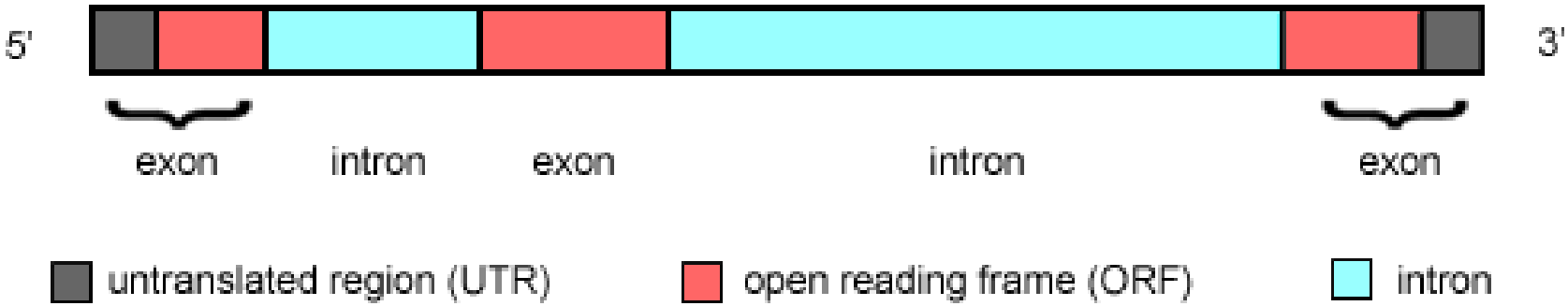
- BEDfiles are one of the most usual intermediate data files to look for genomic associations
- BEDtools and other tools integrate BED files
- Exercises 16 to 24

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
- So, at last, we can answer questions without the hurdle of dealing with sequences, i.e.
  - 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?

Moving forward: what to deal with annotation?

- Genomic annotations are layers to genomic coordinates specifying their nature



Henrik Lantz, BILS/SciLifeLab

How to store genomic annotations? 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

How to store genomic annotations? 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

Why so complex? Open reading frames

Steven M. Carr

Why so complex? CDS, exons, introns, stop codons

- How many transcript does a gene have?
- Are they tissue dependent?
- How can we annotate the different between transcription and translation?

GTF and GFF3 fileformats

- Reading: https://www.ensembl.org/info/website/upload/gff.html
- Columns
  1. seqname (chr/scaffold)
  2. source - name of the program that generated this feature
  3. feature - e.g. Gene, Variation, Similarity
  4. start - with sequence numbering starting at 1
  5. end - end position of the feature, with sequence numbering starting at 1
  6. score - A floating point value
  7. strand - defined as + or -
  8. frame - codon-related
  9. attribute - additional information

Activity

- Run exercises 25 to 27

On data optimization

- Till now: short NGS sequences (FASTQ) get mapped into reference genomes (FASTA) giving rise to alignments (SAM) that are summarized as BED files.
- Next step is basically data mining and visualization
- Let's focus on the latter: genomic tracks for genome browser-based visualization
- Track formats with increased efficiency
  - BEDgraph
  - bigBed (indexed BED file, not plain text!)
  - Wiggle (Wig)
  - bigWig (indexed Wiggle file, not plain text!)

BEDgraph

Conceptually:

```
chromA   chromStartA   chromEndA   dataValueA
chromB   chromStartB   chromEndB   dataValueB
```

That with real data looks like this:

```
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
- Next step in file formats: trying to cover all the genome (that is, no sparsity anymore)
- Example: does it make sense to generate a BED file with GC content? (GC content at Wikipedia)
- How can we store features with definite start and ends but for which the value is the primary purpose, but not their starts and ends?

Wig files

- Imagine we'd like to visualize a track with the GC percent of 10nt bins
- Would it be a good idea to store it as a four-column BED with `chr`, `start`, `end` filling 3/4s of the file?
- Wiggle format deals with it: ideal for continuous data measured at a given step (bin size, length)

Wig

- How to store the GC content of 10nt sized bins in less than 4 columns?
- Specifying the value, the span and the step

```
variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
300706 12.5
300707 12.5
300708 12.5
300709 12.5
300710 12.5
```

Wig with added span

- Can we reduce it further? Two columns it's too much!
- Specifying the value and the step with a fixed span (10 nt)

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

Wig with fixedStep and span

- Can we reduce it further? That was less rows, but still three columns!
- Specifying the value and the step with a fixed span and step

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

- This format reports a score of 11, 22, 33 to 5nt-long bins that are 100 nt apart, starting from the nt 0 of chromosome 3

Activity

- Read more on Wig files at https://genome.ucsc.edu/goldenpath/help/wiggle.html

Variant Call Format

- Standard file format for storing variation data
- Unambiguous, scalable and flexible
- Not suprisingly, structured text file
- 8 columns:
  1. CHROM
  2. POS
  3. ID
  4. REF
  5. ALT
  6. QUAL
  7. FILTER
  8. INFO

VCF

#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

The VCF format

- Activity: read http://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40

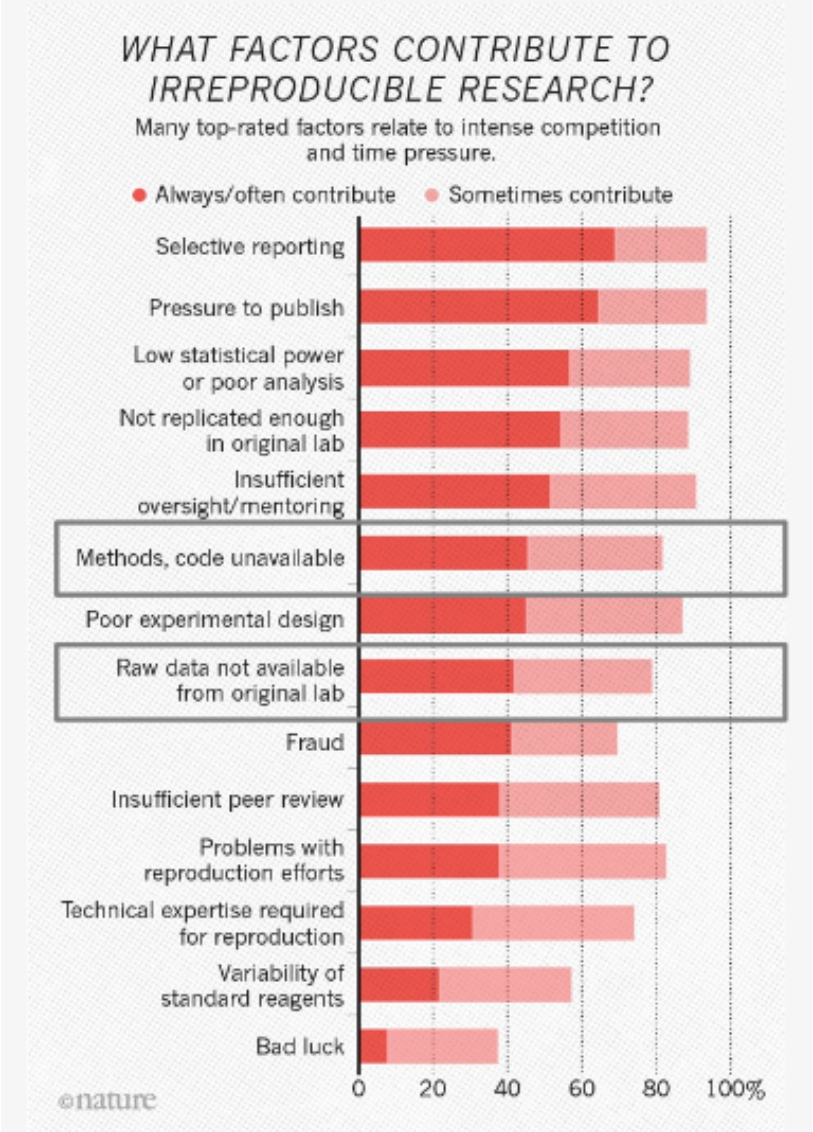
Further activities

- To sum-up, coordinate-based files allow to answer quite complex biological questions.
- For instance, if checking real somatic transposon insertions from the 1000 genomes project variants (VCFs) can we detect any enrichment for certain chromatin states?
- (If you have the time/interest run) exercises from 28 on

Indexed binary file formats

- Most of the data formats can be indexed for fast accessing using data information tricks
- Standard toolsets, i.e. samtools, vcftools etc can convert between plain text and indexed formats
  - SAM, alignments: BAM
  - BED or BedGraph, coordinate-based data: bigBed
  - Wiggle, compact coordinate-based data: bigWig
  - VCF, binary: BCF

Sum-up I



Baker M (2016) Is there a reproducibility crisis? Nature 533:452–454 9

Sum-up II

- Data
  1. Using data standards
  2. Raw data availability
  3. Metadata
  4. Intermediate datasets availability (mid-processed, i.e. BED files)
- Analysis
  1. Scripting everything
  2. Version control
  3. Trace software versions/automate installs
  4. Release all code as supplementary information

Sum-up III

- Genomic data formats are structured and are suited to the different steps of NGS data analysis
- There are open source toolsets for any of them
- They are either plain text files or indexed versions of them that using nonpropietary formats
- As text files, they can be fastly processed using Unix
  - FASTA, FASTQ sequences
  - SAM, alignments (binary: BAM)
  - GTF and GFF, annotations
  - BED, BedGraph, coordinate-based data (binary: bigBed)
  - Wiggle, compact coordinate-based data (binary: bigWig)
  - VCF, variants (binary: BCF)