Studying the transcriptome using RNA-seq

Cecilia Coimbra Klein
Computational Biology of RNA Processing, CRG
Departament de Genètica, IBUB, UB

Master in Omics Data Analysis
Jan. 2019

IBUB
Institut de Biomedicina
de la Universitat de Barcelona

UNIVERSITAT DE BARCELONA

CRG Centre for Genomic Regulation

UVIC UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL DE CATALUNYA

Master in Omics Data Analysis

Outline

Outline

- Basic concepts
- Reference gene annotation
- Next generation sequencing
- RNA-seq experimental protocols
- Short-read RNA-seq data processing
    - mapping
    - visualisation of gene expression signal
    - gene expression quantification
- RNA-seq data analysis
    - sample clustering based on gene expression
    - differential gene expression
    - gene ontology (GO) term enrichment
    - differential splicing analysis

Outline

- ChIP-seq data processing
  - mapping
  - peak calling
  - visualisation of signal
- ChIP-seq data analysis
  - genomic locations
  - differential peaks per tissue
  - BED files in UCSC browser
- Integrative data analysis
  - promoter regions of differentially expressed genes
  - ATAC-seq signal in the UCSC genome browser
  - promoter regions of differentially spliced genes
  - omics portals

Outline

- Basic concepts
- Reference gene annotation
- Next generation sequencing
- RNA-seq experimental protocols
- Short-read RNA-seq data processing
- RNA-seq data analysis
- ChIP-seq data processing
- ChIP-seq data analysis
- Integrative data analysis

Data Analysis Hands-on

Hands-on

- Forebrain, heart and liver of 12.5 days mouse embryos
  - 2 bio replicates
  - RNA-seq, ChIP-seq and ATAC-seq

- References:
  - mouse genome – mm10 assembly
  - gene annotation – gencode vM4

- Processing:
  - References: a small sample of the genome and annotation (21 chromosomes, 1Mb long)
  - Data: one sample only (100,000 alignment-based pre-filtered reads)

- Analysis:
  - all samples

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

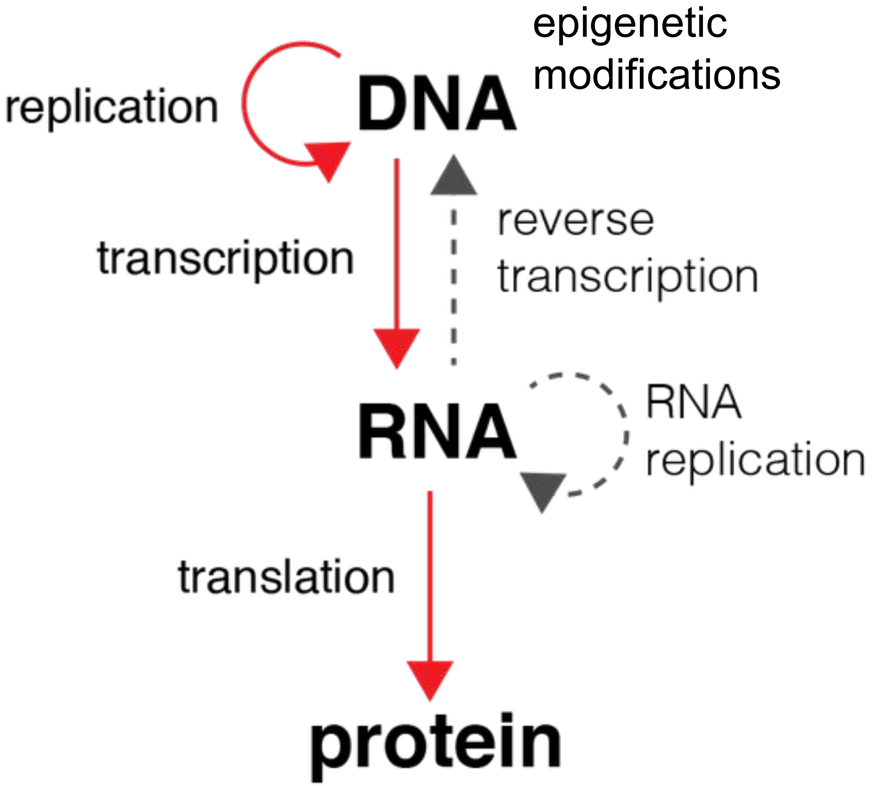
Hands-on

Setup environment 1

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Basic concepts

Molecular biology dogma



epigenetic modifications

replication

DNA

transcription

reverse transcription

RNA

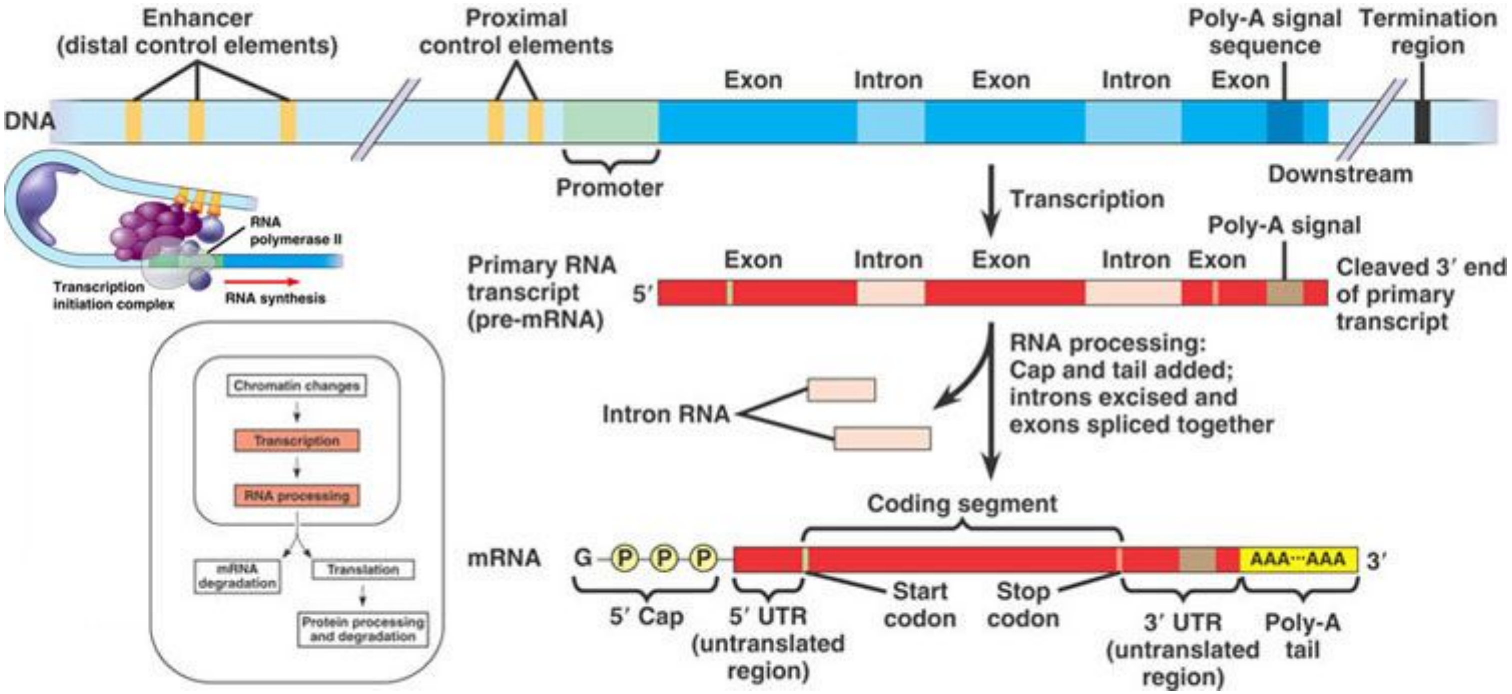
RNA replication

translation

protein

- Only ~1% of the human genome produces proteins, although much more is transcribed (~60%).
- The genome is identical in all cell types, however not all cell types have the same function. That’s why the transcriptome (and the epigenome) becomes also relevant.

RNA transcription and processing



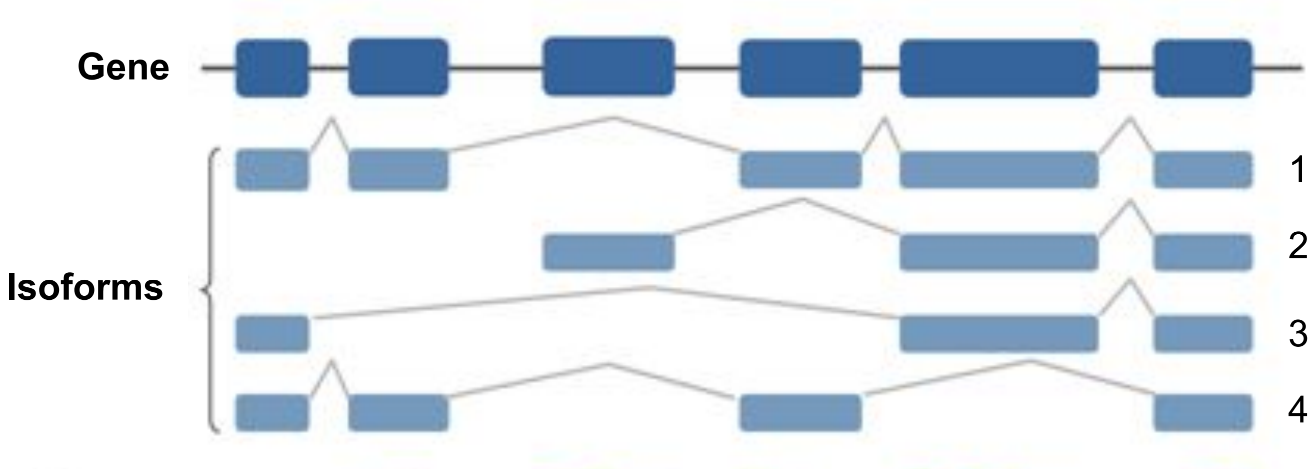
Primary RNA transcripts are extensively processed: capping, splicing, polyadenylation, editing

This process is highly regulated and results in a gene producing many distinct transcript isoforms: one gene, many transcripts

The transcriptome is distinct from and more complex than the genome

The transcriptome cannot be predicted from the genome sequence alone: it must be measured

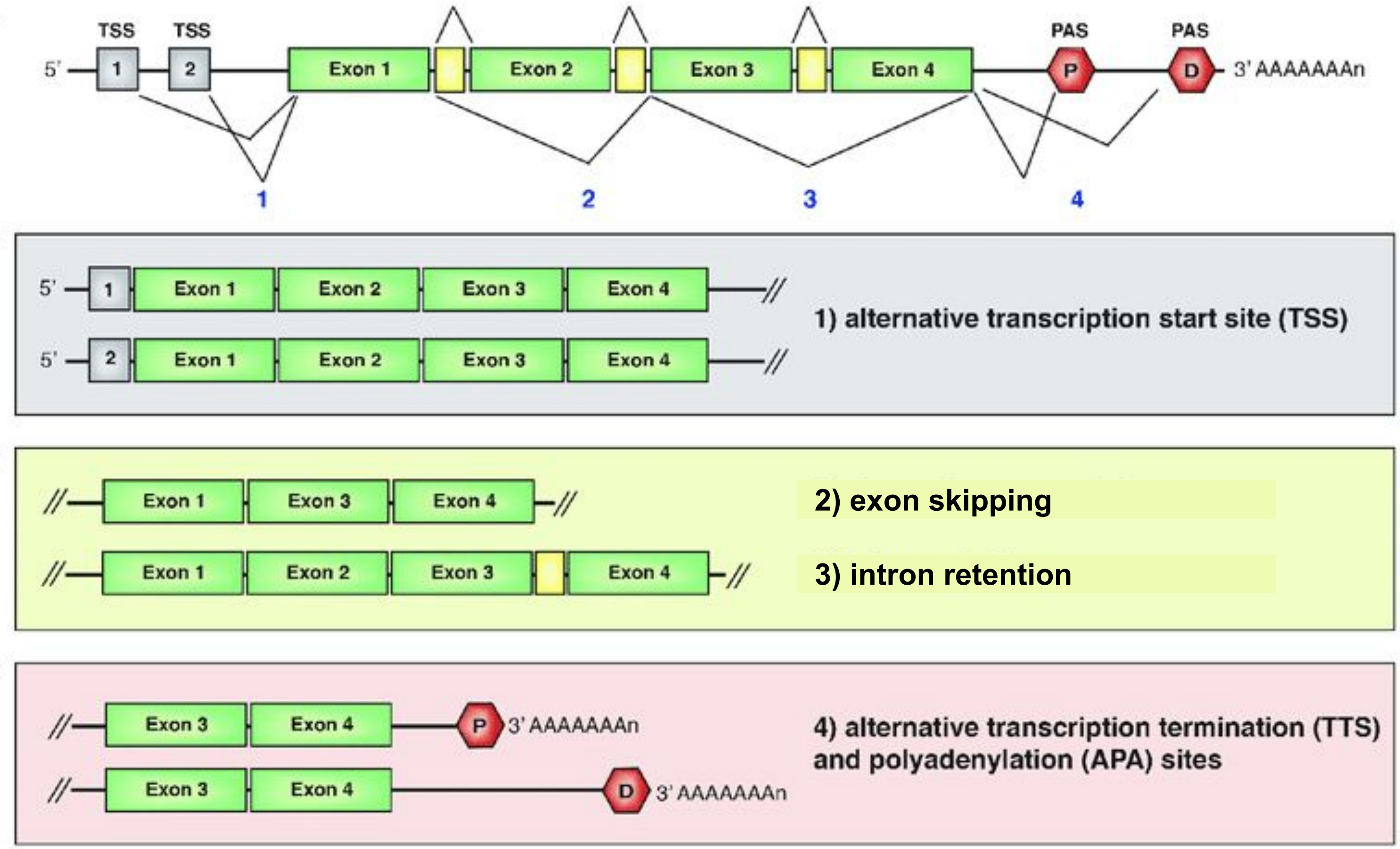
Genome and transcriptome



Some definitions:

- Genome: the full DNA complement of a species’ cell
- Gene: the physical region of a chromosome producing some kind or RNA transcript
- Isoforms: distinct RNAs arising from the gene, through differential exon inclusion, transcription start or termination sites.
- Transcript: The RNA molecule corresponding to one of the isoforms
- Transcriptome: the full RNA complement of a species' cell

Complexity arising from differential processing



1) alternative transcription start site (TSS)

2) exon skipping

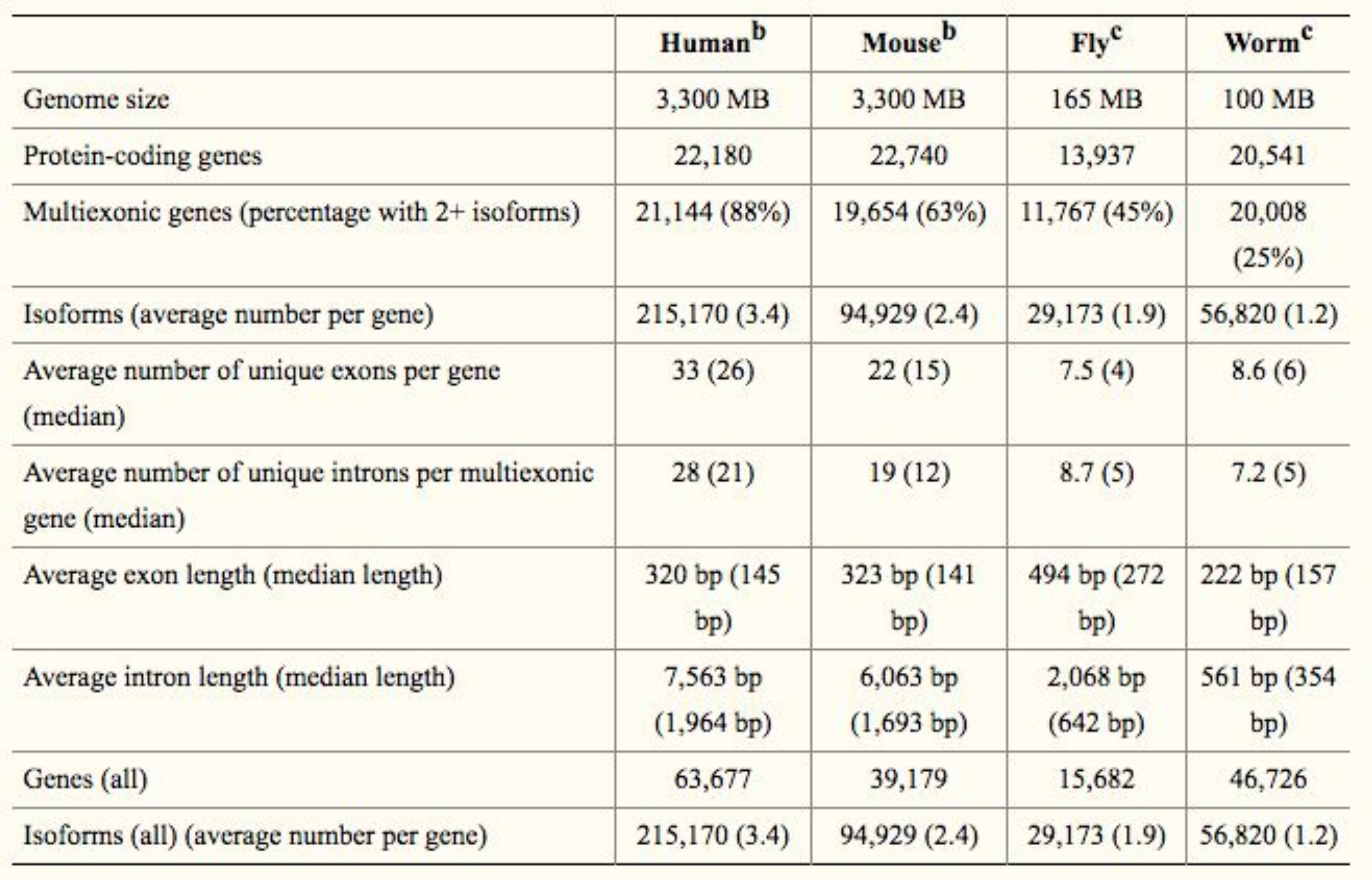
3) intron retention

4) alternative transcription termination (TTS) and polyadenylation (APA) sites

These processing events can result in different protein products, differentially (post-) transcriptionally regulated mRNAs or non-protein coding isoforms.

Andreassi, C. *et al.* (2018). doi: 10.3389/fnmol.2018.00304

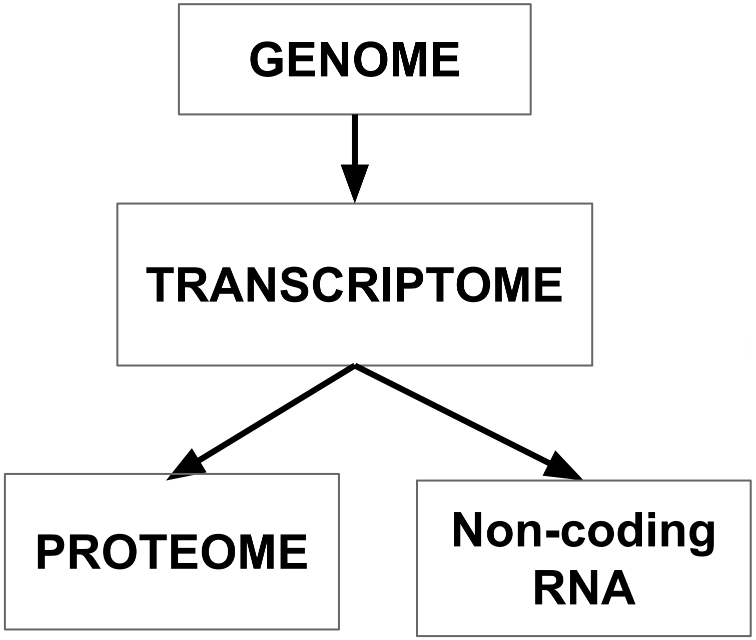
Complexity arising from differential processing



	Human[b]	Mouse[b]	Fly[c]	Worm[c]
Genome size	3,300 MB	3,300 MB	165 MB	100 MB
Protein-coding genes	22,180	22,740	13,937	20,541
Multiexonic genes (percentage with 2+ isoforms)	21,144 (88%)	19,654 (63%)	11,767 (45%)	20,008 (25%)
Isoforms (average number per gene)	215,170 (3.4)	94,929 (2.4)	29,173 (1.9)	56,820 (1.2)
Average number of unique exons per gene (median)	33 (26)	22 (15)	7.5 (4)	8.6 (6)
Average number of unique introns per multiexonic gene (median)	28 (21)	19 (12)	8.7 (5)	7.2 (5)
Average exon length (median length)	320 bp (145 bp)	323 bp (141 bp)	494 bp (272 bp)	222 bp (157 bp)
Average intron length (median length)	7,563 bp (1,964 bp)	6,063 bp (1,693 bp)	2,068 bp (642 bp)	561 bp (354 bp)
Genes (all)	63,677	39,179	15,682	46,726
Isoforms (all) (average number per gene)	215,170 (3.4)	94,929 (2.4)	29,173 (1.9)	56,820 (1.2)

Lee & Rio (2015). doi:10.1146/annurev-biochem-060614-034316

RNA composition in the cell

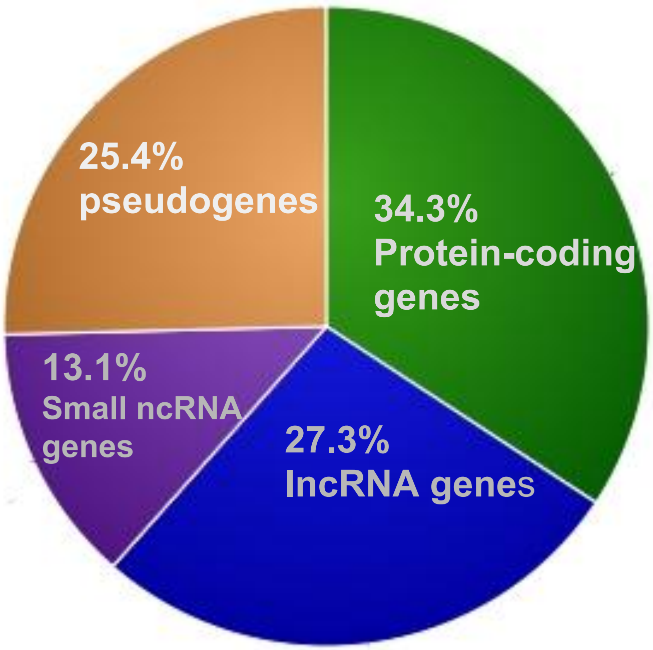


GENOME

TRANSCRIPTOME

PROTEOME

Non-coding RNA



From gencode v.26 annotation

- Only part of the human transcriptome encode proteins
- Many different type of regulatory RNAs, small <200nt and long >200nt
- lncRNAs: transcribed by RNA Polymerase II, actively processed
- Functionally important, have many signatures of mRNAs
- XIST, HOTAIR, TelRNAs

Reference gene annotation

Reference gene annotation

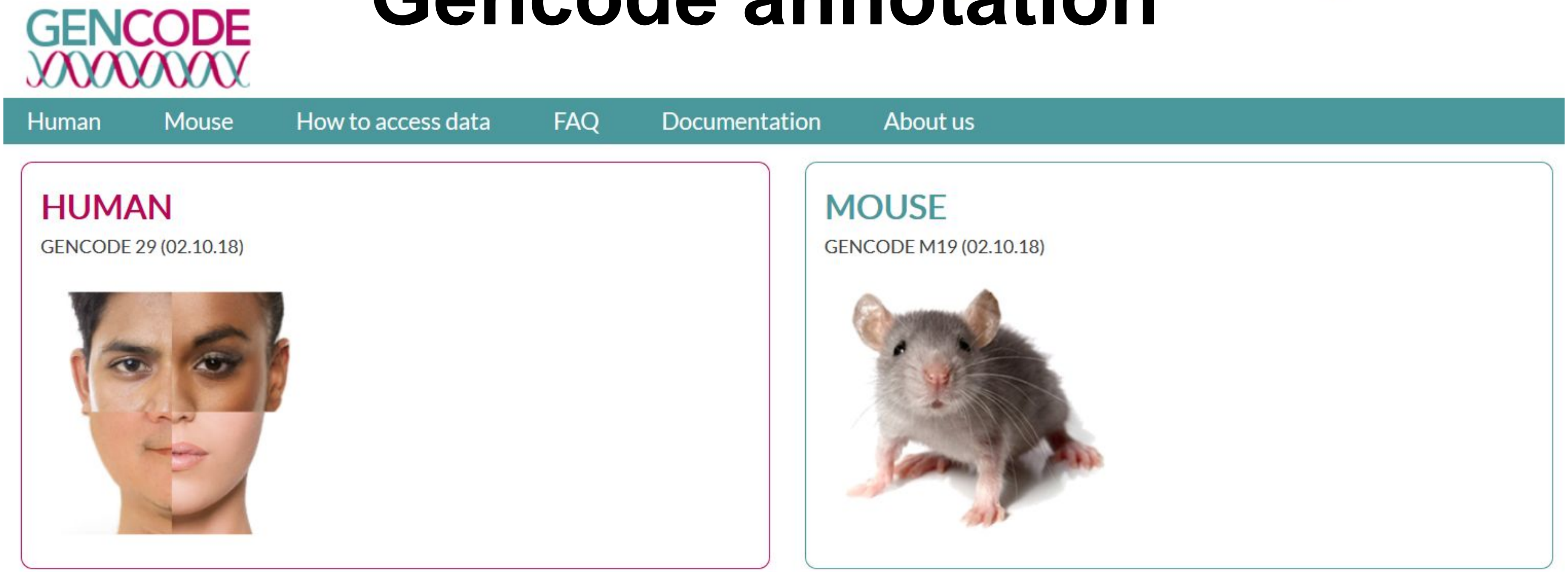
- For a given species and associated genome assembly, the reference gene annotation is the collection of all genes known for this species

- A gene annotation (like a genome assembly) can be at various completion stages depending on the species. High-quality annotations: human, mouse, *D. melanogaster*, *C. elegans* or yeast.

- It is important to choose well the reference gene annotation beforehand since it will represent the known transcriptome to which the RNA-seq transcriptome will be compared.

Always check the annotation version you're going to use.

Gencode annotation



HUMAN

GENCODE 29 (02.10.18)

MOUSE

GENCODE M19 (02.10.18)

https://www.gencodegenes.org/

- **4 broad gene categories**: protein-coding genes (~20,000), long non-coding genes, pseudogenes, small non-coding genes
- **Several features:** gene, transcript, exon, CDS, UTR
- **3 confidence levels**: automatically annotated < manually annotated < validated
- **File formats**: GTF/GFF3

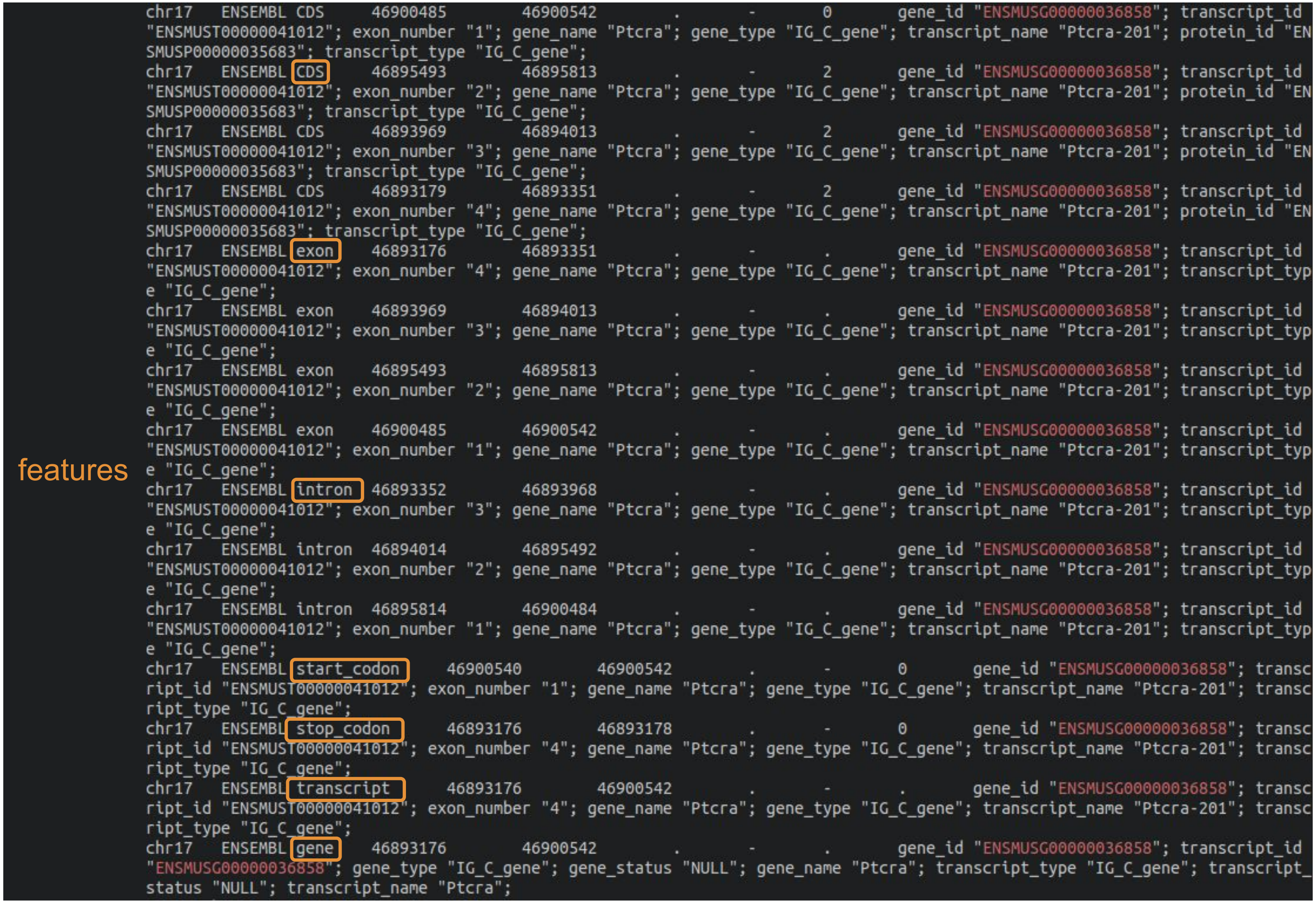
Gencode lncRNA gene annotation

- Gencode has always annotated lncRNA genes and was calling them "processed_transcript"
- Since they are more and more numerous and interesting to people, Gencode now better classifies them, partly using their location to PCGs:

3prime_overlapping_ncrna	Transcripts where ditag and/or published experimental data strongly supports the existence of long non-coding transcripts transcribed from the 3'UTR.
sense_intronic	Long non-coding transcript in introns of a coding gene that does not overlap any exons.
sense_overlapping	Long non-coding transcript that contains a coding gene in its intron on the same strand.
antisense	Transcript believed to be an antisense product used in the regulation of the gene to which it belongs.
non_coding	Transcript which is known from the literature to not be protein coding.
processed_transcript	Doesn't contain an ORF.
lincRNA	Long, intervening noncoding (linc)RNAs, that can be found in evolutionarily conserved, intergenic regions.

GTF format

a text-based format for storing features information



features

```
chr17   ENSEMBL CDS     46900485        46900542        .       -       0       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "1"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; protein_id "EN
SMUSP00000035683"; transcript_type "IG_C_gene";
chr17   ENSEMBL CDS     46895493        46895813        .       -       2       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "2"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; protein_id "EN
SMUSP00000035683"; transcript_type "IG_C_gene";
chr17   ENSEMBL CDS     46893969        46894013        .       -       2       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "3"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; protein_id "EN
SMUSP00000035683"; transcript_type "IG_C_gene";
chr17   ENSEMBL CDS     46893179        46893351        .       -       2       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "4"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; protein_id "EN
SMUSP00000035683"; transcript_type "IG_C_gene";
chr17   ENSEMBL exon    46893176        46893351        .       -       .       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "4"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transcript_typ
e "IG_C_gene";
chr17   ENSEMBL exon    46893969        46894013        .       -       .       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "3"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transcript_typ
e "IG_C_gene";
chr17   ENSEMBL exon    46895493        46895813        .       -       .       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "2"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transcript_typ
e "IG_C_gene";
chr17   ENSEMBL exon    46900485        46900542        .       -       .       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "1"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transcript_typ
e "IG_C_gene";
chr17   ENSEMBL intron  46893352        46893968        .       -       .       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "3"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transcript_typ
e "IG_C_gene";
chr17   ENSEMBL intron  46894014        46895492        .       -       .       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "2"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transcript_typ
e "IG_C_gene";
chr17   ENSEMBL intron  46895814        46900484        .       -       .       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUST00000041012"; exon_number "1"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transcript_typ
e "IG_C_gene";
chr17   ENSEMBL start_codon     46900540        46900542        .       -       0       gene_id "ENSMUSG00000036858"; transc
ript_id "ENSMUST00000041012"; exon_number "1"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transc
ript_type "IG_C_gene";
chr17   ENSEMBL stop_codon      46893176        46893178        .       -       0       gene_id "ENSMUSG00000036858"; transc
ript_id "ENSMUST00000041012"; exon_number "4"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transc
ript_type "IG_C_gene";
chr17   ENSEMBL transcript      46893176        46900542        .       -       .       gene_id "ENSMUSG00000036858"; transc
ript_id "ENSMUST00000041012"; exon_number "4"; gene_name "Ptcra"; gene_type "IG_C_gene"; transcript_name "Ptcra-201"; transc
ript_type "IG_C_gene";
chr17   ENSEMBL gene    46893176        46900542        .       -       .       gene_id "ENSMUSG00000036858"; transcript_id
"ENSMUSG00000036858"; gene_type "IG_C_gene"; gene_status "NULL"; gene_name "Ptcra"; transcript_type "IG_C_gene"; transcript_
status "NULL"; transcript_name "Ptcra";
```

Hands-on

Setup environment 1

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Next generation sequencing

NGS: Illumina sequencing

- Illumina Sequencing (short reads ~ max. 150bp)
    - *single end*
        1) Library preparation: DNA fragmentation, adapter ligation, PCR amplification
        2) Solid-phase *bridge* amplification
        3) Flowing of fluorescent reversible terminator dNTPs; incorporation of a single base per cycle. *Sequencing by synthesis*.
        4) Read identity of each base of a cluster from sequential images

    - *paired end*

        5) After completion of the first read, the templates can be regenerated *in situ* to enable a second read from the opposite end.

NGS: Third generation sequencing

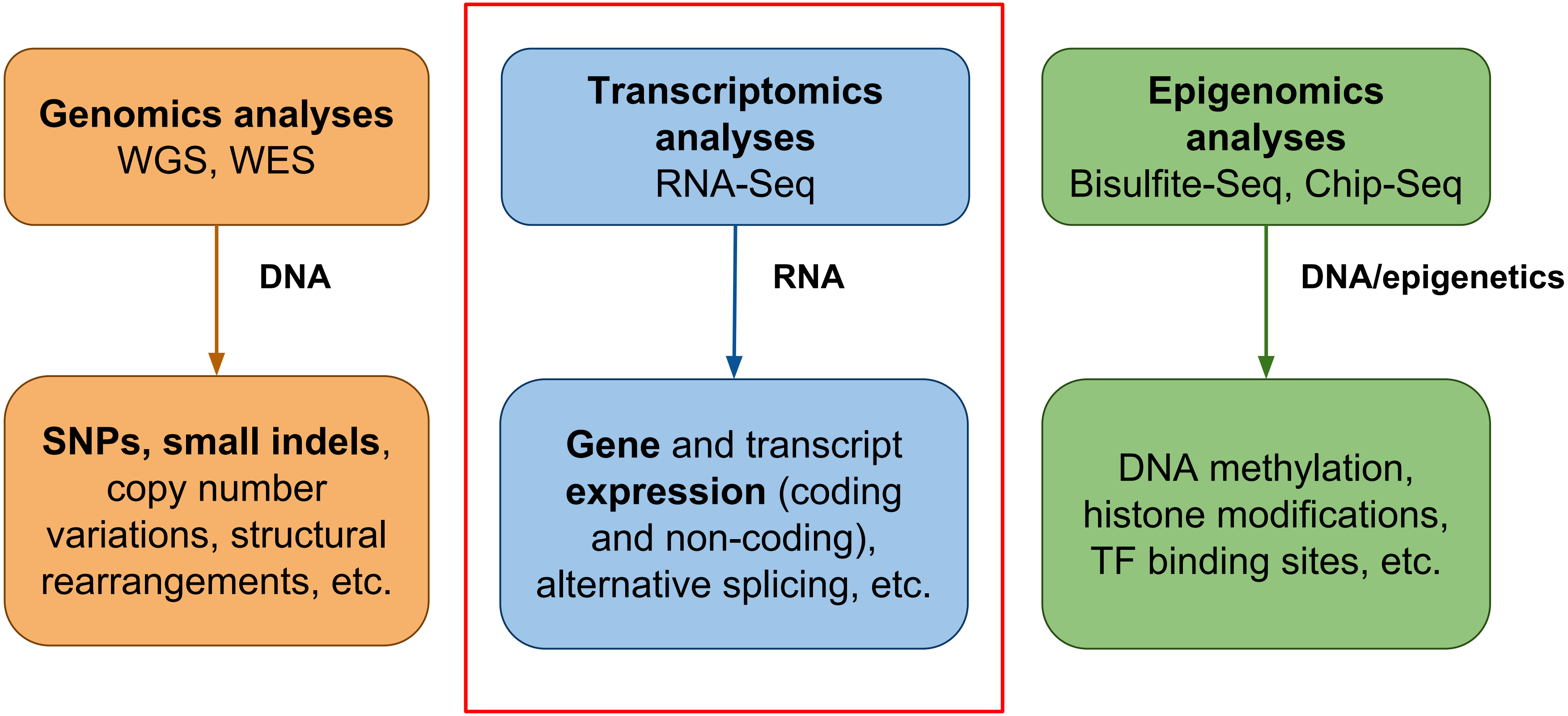
- Although Illumina is by far the most popular, there are many other sequencing technologies, such as PacBio, Ion Torrent or Oxford NanoPore that:

  - allow sequencing genomic material without neither fragmentation nor clonal amplification.

  - enable getting longer reads (tens of Kb!), but at the price of a much higher error rate than Illumina.

  - have been mostly used for genome sequencing, since those reads can span complicated repeat-rich regions which are trickier to assemble using short reads.

Which *-Seq do I need?


Genomics analyses
WGS, WES
DNA
SNPs, small indels, copy number variations, structural rearrangements, etc.
Transcriptomics analyses
RNA-Seq
RNA
Gene and transcript **expression** (coding and non-coding), alternative splicing, etc.
Epigenomics analyses
Bisulfite-Seq, Chip-Seq
DNA/epigenetics
DNA methylation, histone modifications, TF binding sites, etc.

- Learn more about your favourite *-Seq here!
- Note that we are always talking about *re-sequencing*, which is something different from *de novo sequencing* (what is done for a new genome assembly)

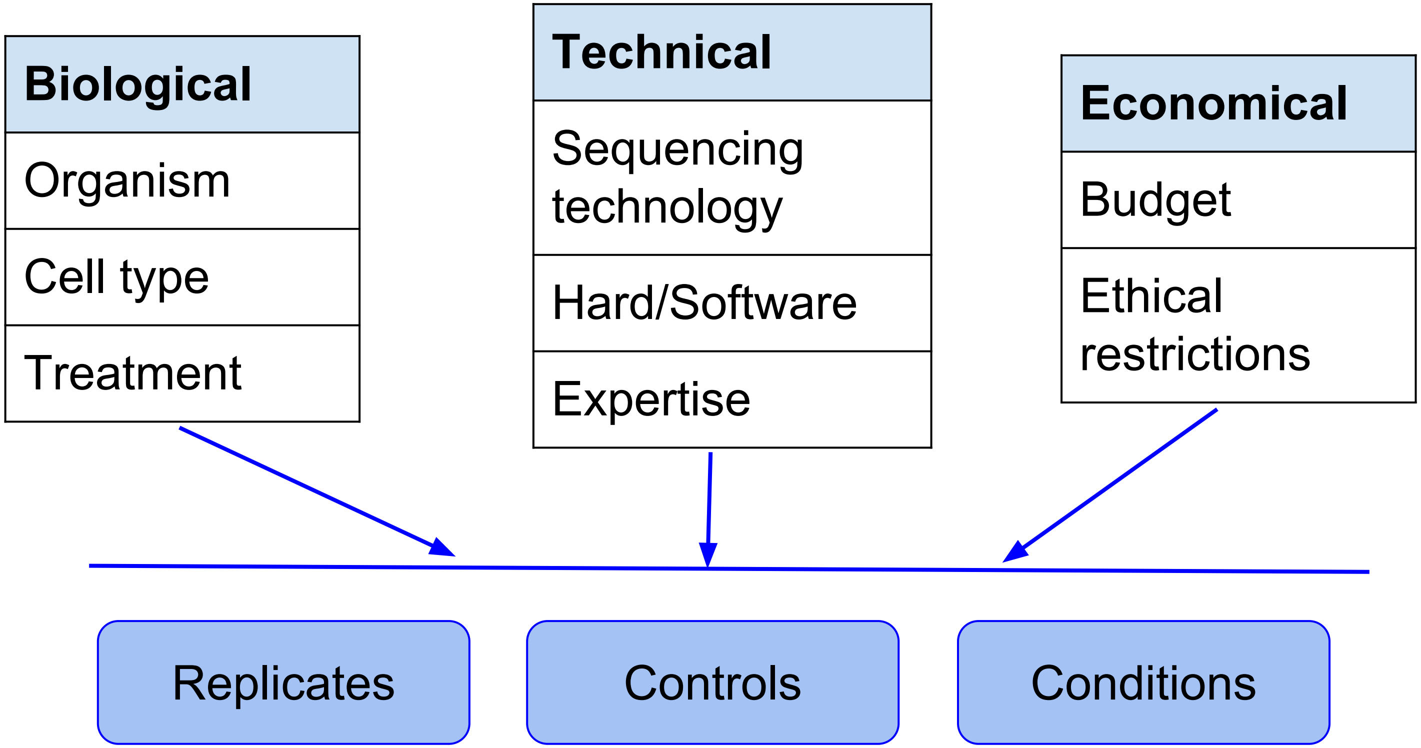
Cecilia Coimbra Klein

RNA sequencing

Why is it useful?

- Measure gene and transcript expression at different conditions, developmental stages, etc.
- Discover / annotate novel elements: genes (coding and non-coding), transcripts, exons, (chimeric) junctions, circular RNAs, etc.
- Alternative splicing, transcription start and termination (polyadenylation) sites.

Experimental design



Biological
Organism
Cell type
Treatment

Technical
Sequencing technology
Hard/Software
Expertise

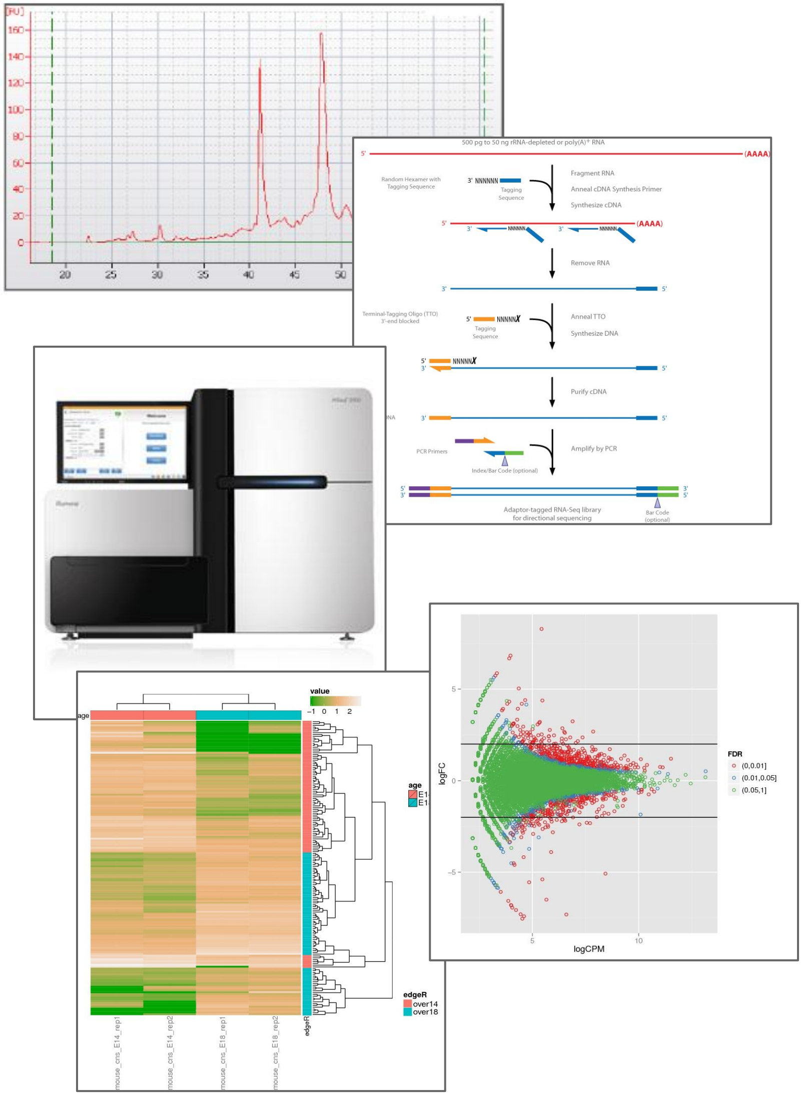
Economical
Budget
Ethical restrictions

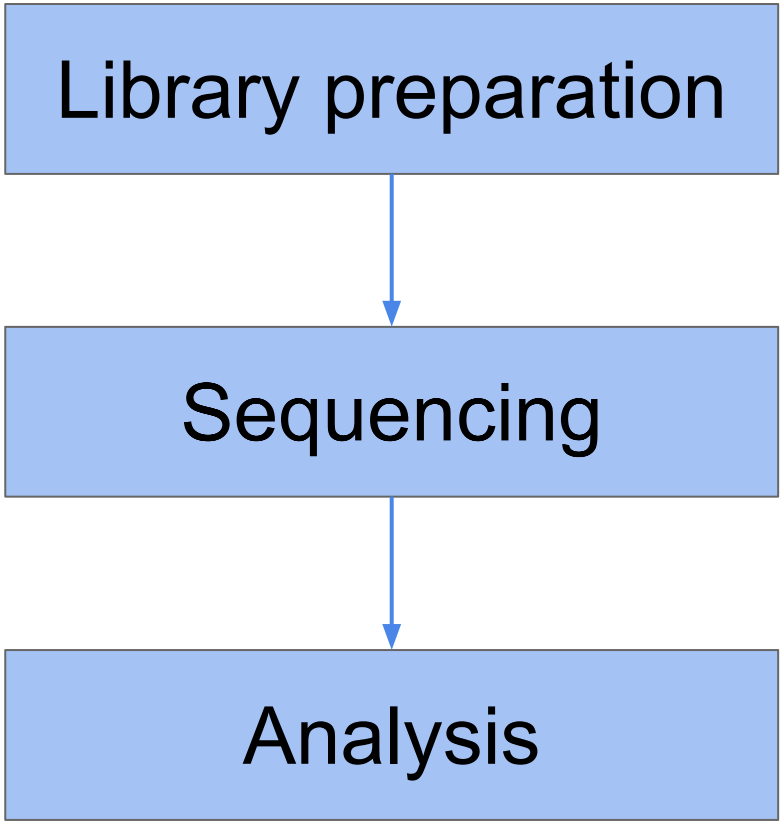
Replicates

Controls

Conditions

RNA-seq experiment





Library preparation

Sequencing

Analysis

Experimental variables of RNA-seq

Cellular localization
Whole cell
Chromatin
Exosome
Nucleus
Cytoplasm

RNA purification
Total RNA
PolyA+
PolyA-
Ribo-

Size selection
Long (>200nt)
Short (<200nt)

Preparation
Single end
Paired end

Strandness
Stranded
Unstranded

Special protocols
Single-cell RNA-seq
Nascent RNA-seq (GRO-seq/NUN-seq)
miRNA-seq

Experimental variables of RNA-seq

Cellular localization
Whole cell
Chromatin
Exosome
Nucleus
Cytoplasm

RNA purification
Total RNA
PolyA+
PolyA-
Ribo-

Size selection
Long (>200nt)
Short (<200nt)

Preparation
Single end
Paired end

Strandness
Stranded
Unstranded

Special protocols
Single-cell RNA-seq
Nascent RNA-seq (GRO-seq/NUN-seq)
miRNA-seq

Experimental variables of RNA-seq

Cellular localization
Whole cell
Chromatin
Exosome
Nucleus
Cytoplasm

RNA purification
Total RNA
PolyA+
PolyA-
Ribo-

Size selection
Long (>200nt)
Short (<200nt)

Preparation
Single end
Paired end

Strandness
Stranded
Unstranded

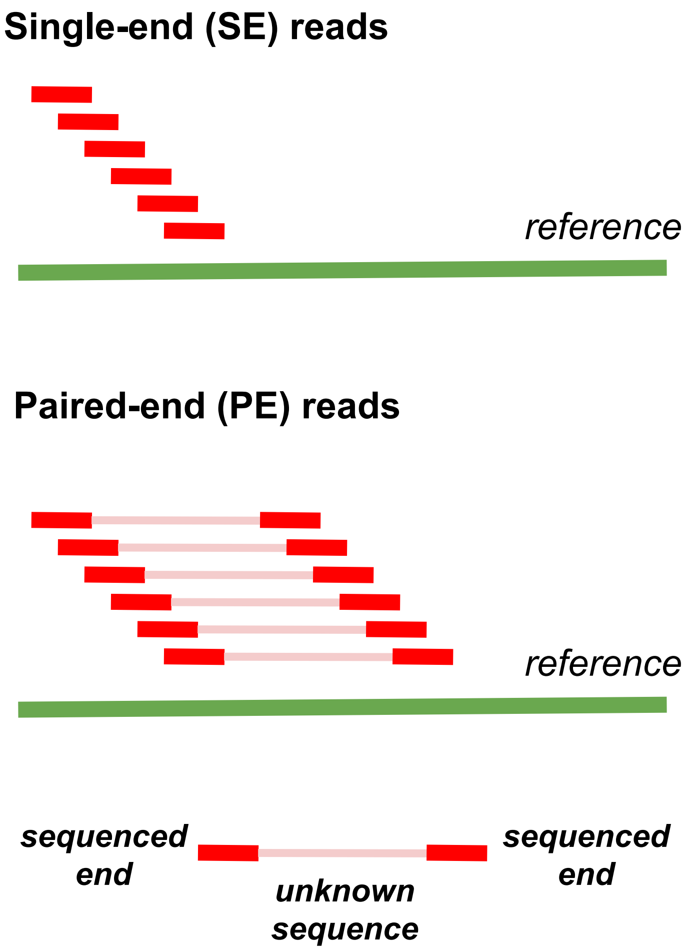
Special protocols
Single-cell RNA-seq
Nascent RNA-seq (GRO-seq/NUN-seq)
miRNA-seq

OUR HANDS-ON

RNA purification protocol

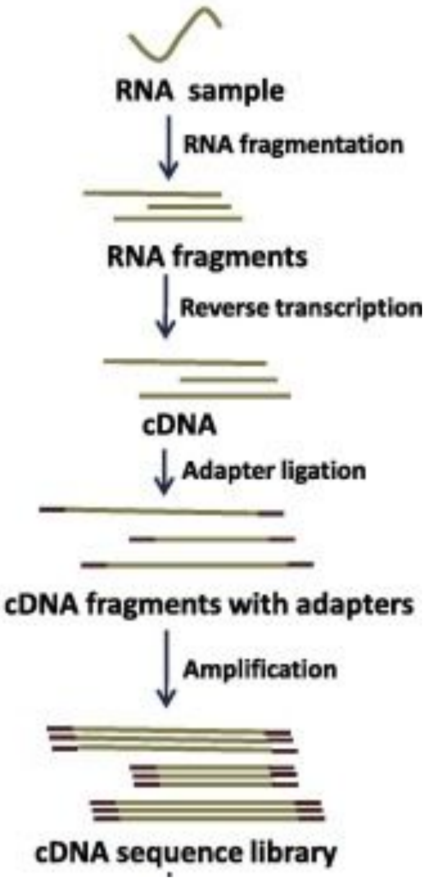
- **PolyA+** gets rid of the ribosomal RNAs and purify mature polyadenylated transcripts.
- **PolyA-** enrichs for non-mature RNAs
- **Ribo-** gets rid of the ribosomal RNAs but capture both mature and non-mature RNAs

Preparation

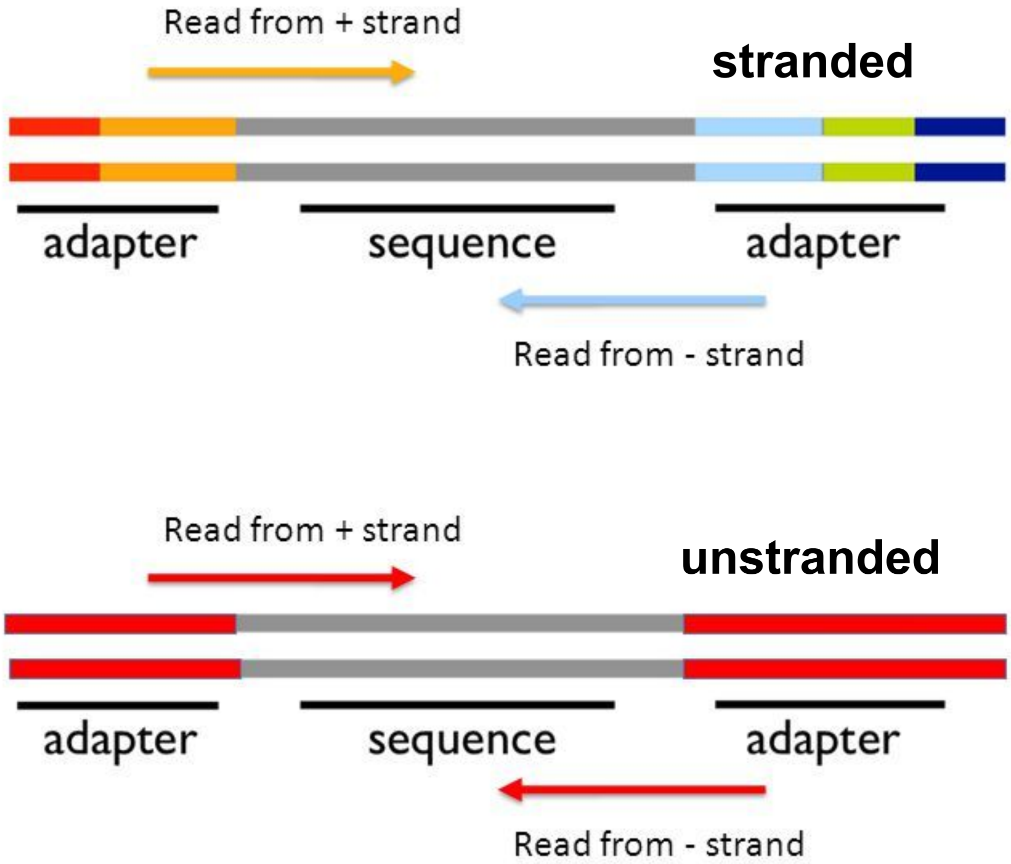


Cecilia Coimbra Klein

Library preparation



Strandness

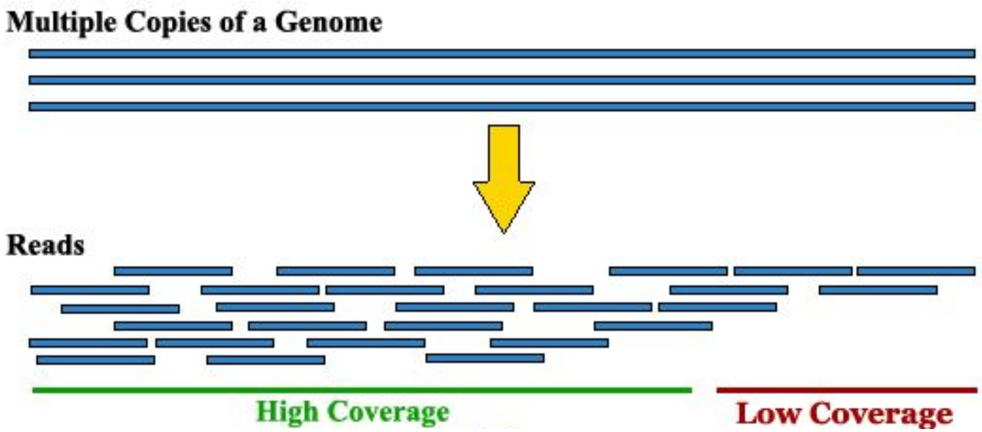


Cecilia Coimbra Klein

How much to sequence?

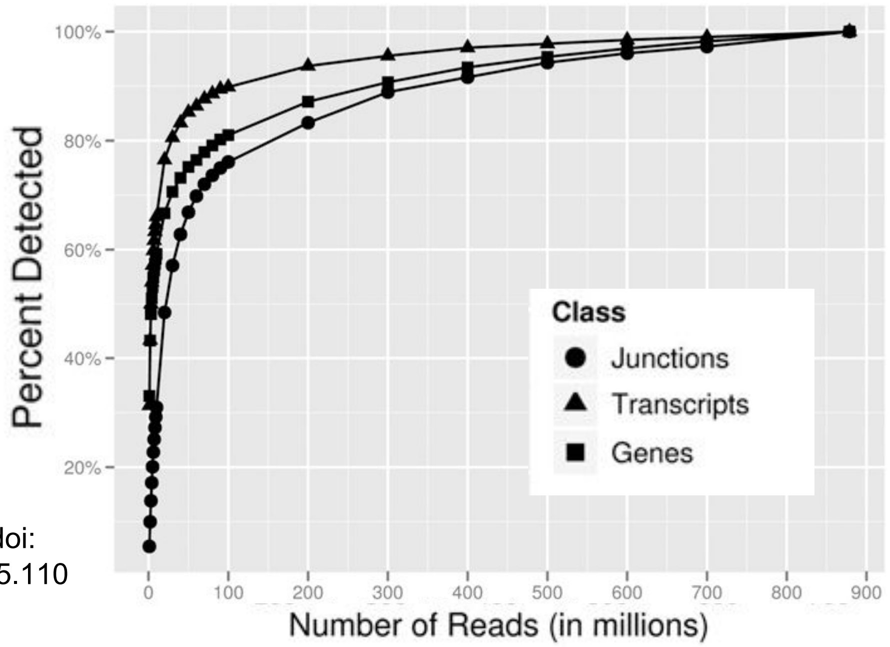
Depends on multiple factors:

- goal of experiment
- protocol
- species
- etc.



e.g. in humans:

>30M reads for simple analyses
>100M reads for novel elements discovery



Toung, J. (2011) doi: 10.1101/gr.116335.110

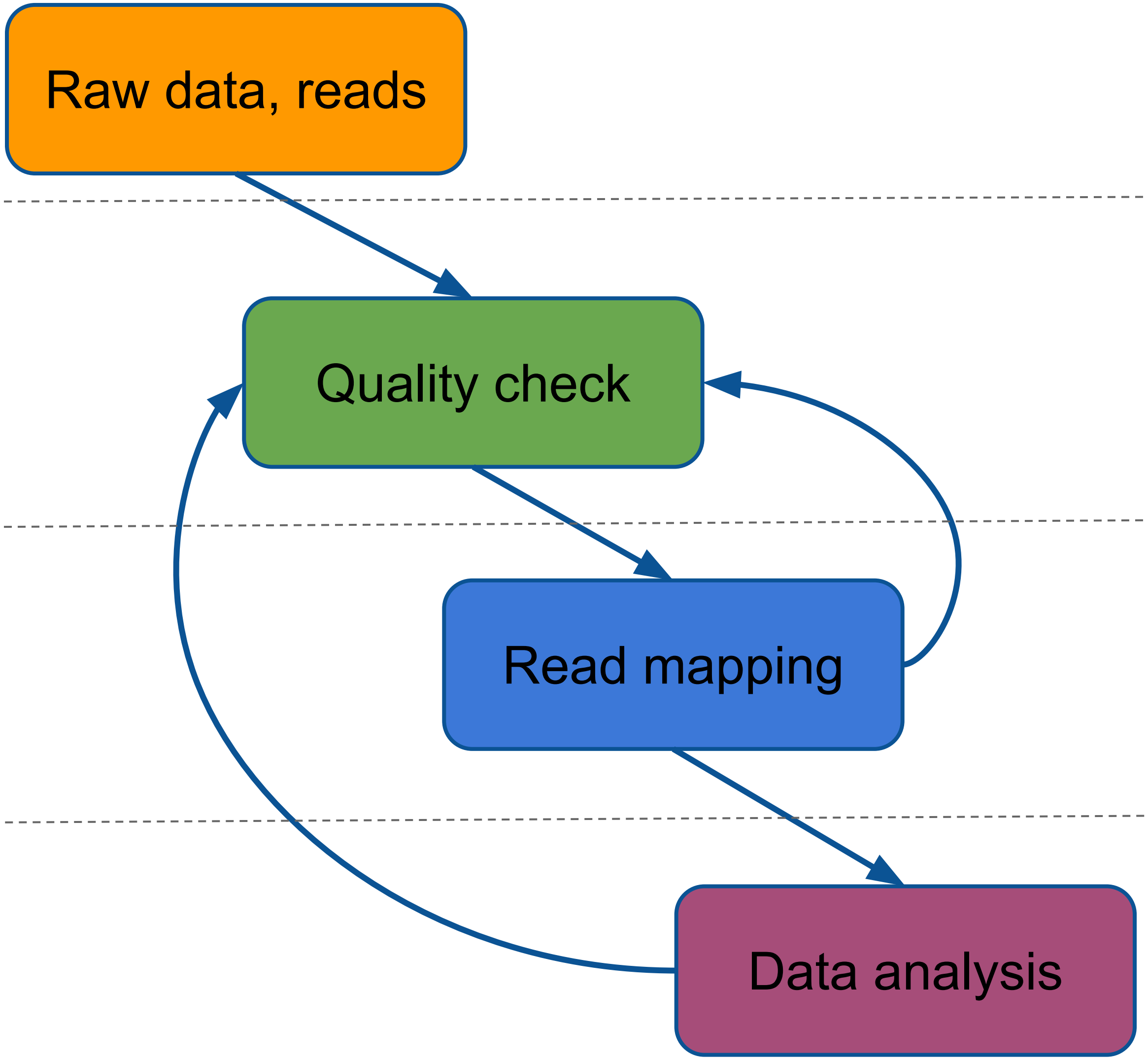
Hands-on

Setup environment 1

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Data formats

Typical pipeline



Some data formats

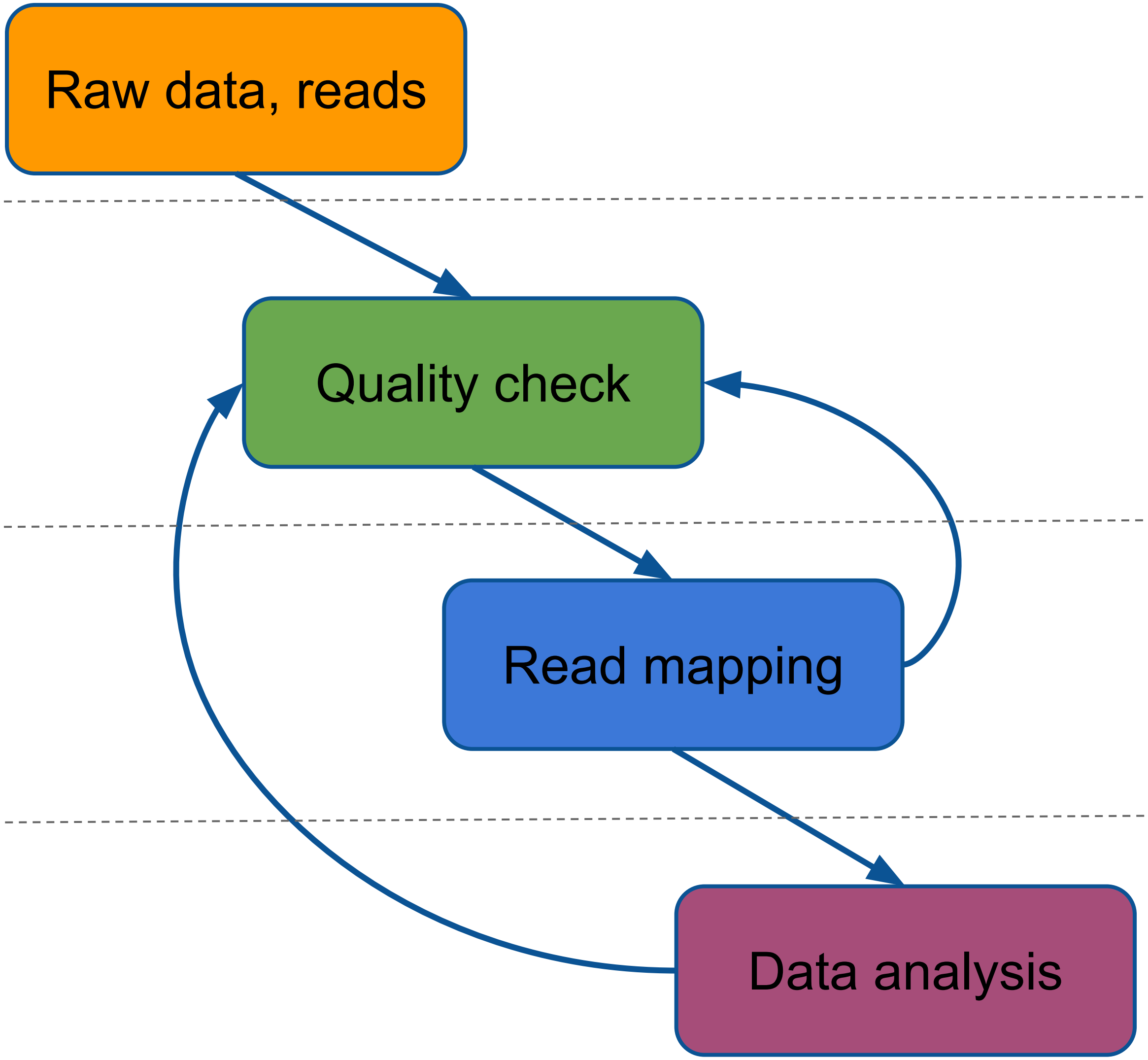
*.fastq, *.fa,
*.sff, *.sra

*.fastq
*.tsv, *.html..

*.sam, *.bam
*.bed, *.wig, *.bw
*.bedgraph
*.gtf, *.fa,..

*.vcf
*.tsv
*.ace, *.agp

Typical pipeline



Some data formats

*.fastq, *.fa,
*.sff, *.sra

*.fastq
*.tsv, *.html..

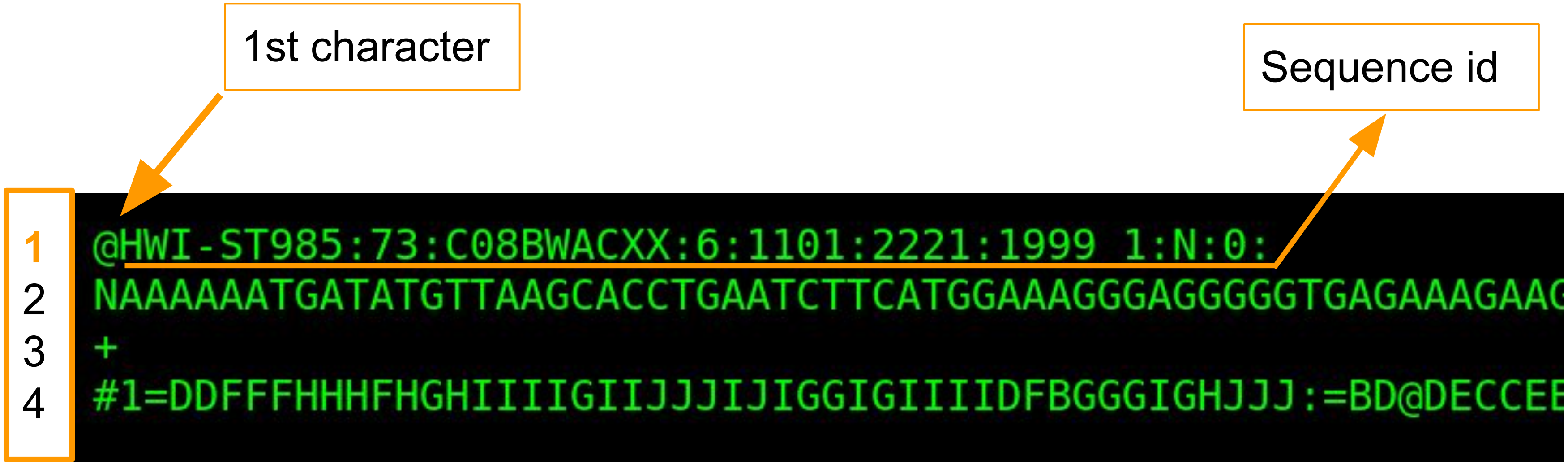
*.sam, *.bam
*.bed, *.wig, *.bw
*.bedgraph
*.gtf, *.fa,..

*.vcf
*.tsv
*.ace, *.agp

FASTQ format

FASTQ Format

a text-based format for storing biological sequences and their corresponding quality scores

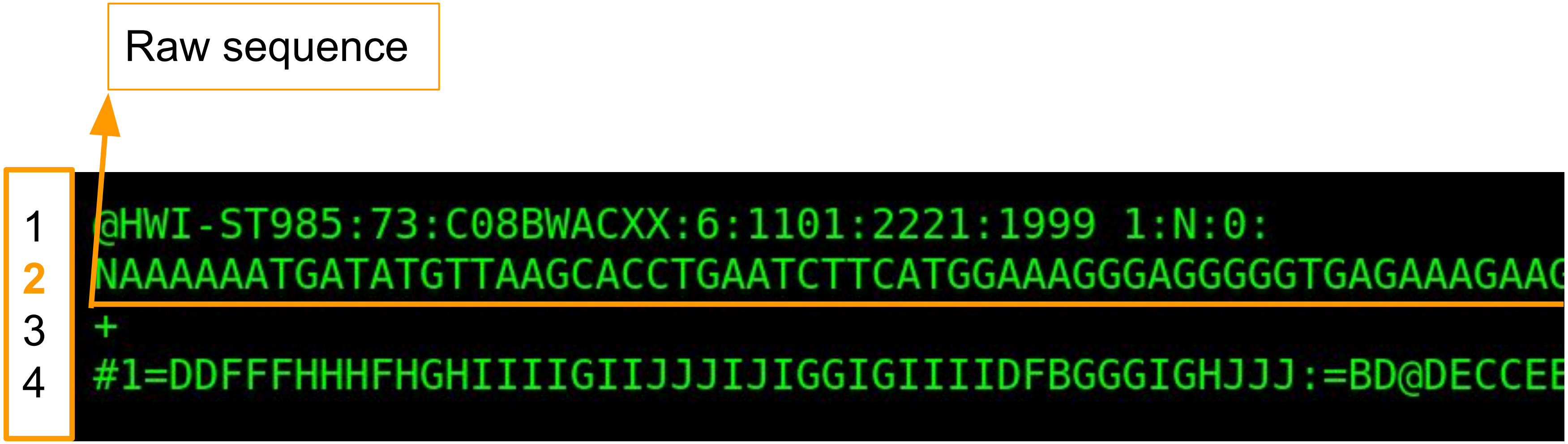


```
@HWI-ST985:73:C08BWACXX:6:1101:2221:1999 1:N:0:
NAAAAAATGATATGTTAAGCACCTGAATCTTCATGGAAAGGGAGGGGGTGAGAAAGAAG
+
#1=DDFFFHHHFHGHIIIIGIIJJJIJIGGIGIIIIDFBGGGIGHJJJ:=BD@DECCEE
```

Optionally: The sequence id can be followed by a description

FASTQ Format

a text-based format for storing biological sequences and their corresponding quality scores



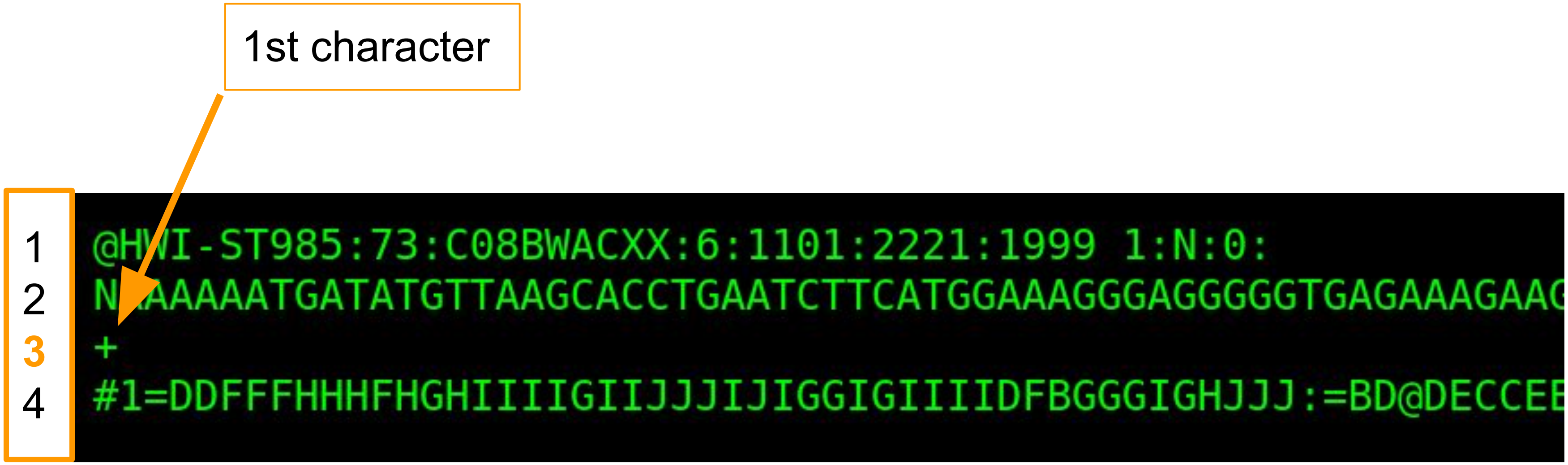
Raw sequence

```
@HWI-ST985:73:C08BWACXX:6:1101:2221:1999 1:N:0:
NAAAAAATGATATGTTAAGCACCTGAATCTTCATGGAAAGGGAGGGGGTGAGAAAGAAG
+
#1=DDFFFHHHFHGHIIIIGIIJJJIJIGGIGIIIIDFBGGGIGHJJJ:=BD@DECCEE
```

Cecilia Coimbra Klein

FASTQ Format

a text-based format for storing biological sequences and their corresponding quality scores



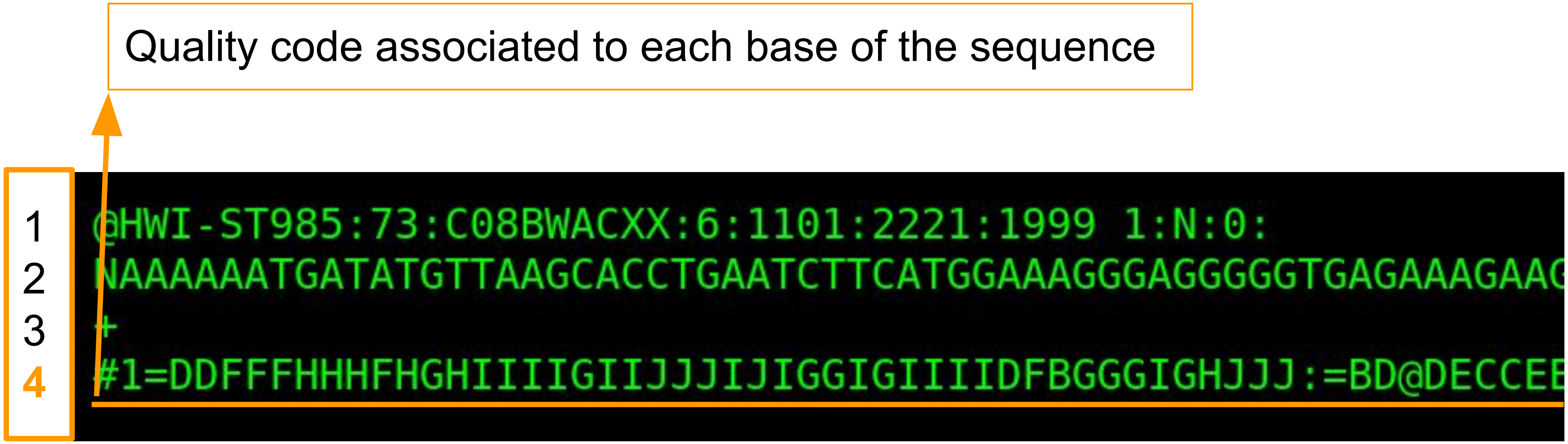
1st character

```
@HWI-ST985:73:C08BWACXX:6:1101:2221:1999 1:N:0:
NAAAAAATGATATGTTAAGCACCTGAATCTTCATGGAAAGGGAGGGGGTGAGAAAGAAG
+
#1=DDFFFHHHFHGHIIIIGIIJJJIJIGGIGIIIIDFBGGGIGHJJJ:=BD@DECCEE
```

Optionally: “+” can can be followed by the sequence id and any description

FASTQ Format

a text-based format for storing biological sequences and their corresponding quality scores



Quality code associated to each base of the sequence

```
1  @HWI-ST985:73:C08BWACXX:6:1101:2221:1999 1:N:0:
2  NAAAAAATGATATGTTAAGCACCTGAATCTTCATGGAAAGGGAGGGGGTGAGAAAGAAG
3  +
4  #1=DDFFFHHHFHGHIIIIGIIJJJIJIGGIGIIIIDFBGGGIGHJJJ:=BD@DECCEE
```

Cecilia Coimbra Klein

FASTQ Format - summary

Four lines per sequence are used in a FASTQ file:

1. begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a FASTA title line)
2. the raw sequence
3. begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description)
4. encodes the quality values for the sequence contained in line 2 (must contain the same number of symbols as the sequence)

FASTQ Format - quality offset

A quality value *Q* is an integer mapping of *p* (i.e., the probability that the corresponding base call is incorrect). The most used formula is the Phred quality score:

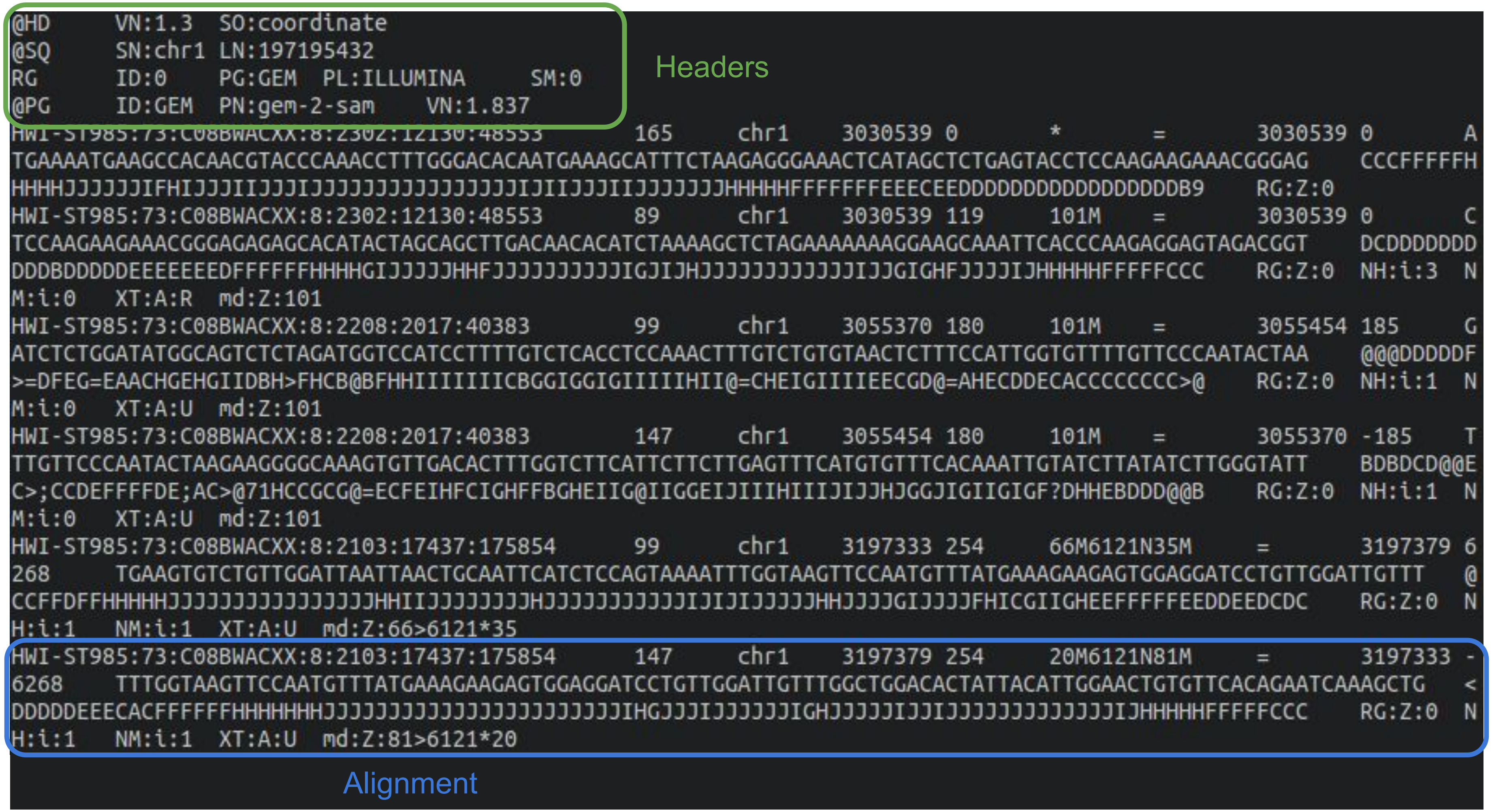
$$Q_{phred} = -10 \log_{10} p$$

offset	max Phred score range	max ASCII range	real-world Phred score range	real-world ASCII range
33	0 - 93	33 - 126	0 - 40	33 - 73
64	0 - 62	64 - 126	0 - 40	64 - 104

https://en.wikipedia.org/wiki/FASTQ_format#Encoding

SAM format

Sequence Alignment/Map

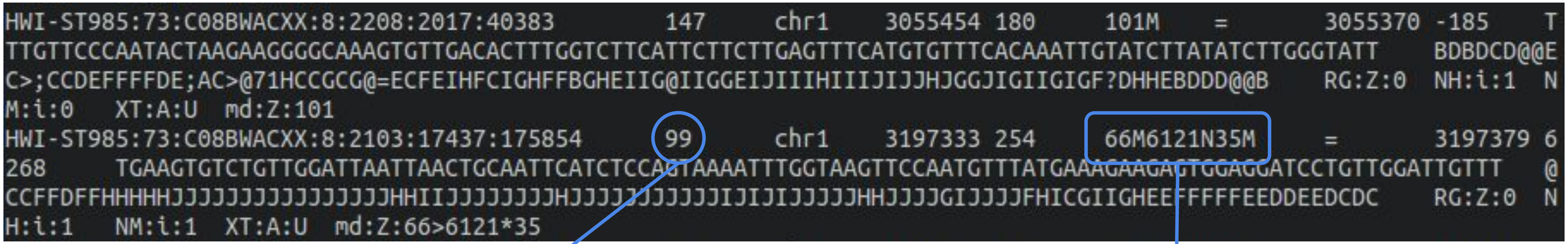


```
@HD     VN:1.3  SO:coordinate
@SQ     SN:chr1 LN:197195432
RG      ID:0    PG:GEM  PL:ILLUMINA     SM:0
@PG     ID:GEM  PN:gem-2-sam    VN:1.837
HWI-ST985:73:C08BWACXX:8:2302:12130:48553       165     chr1    3030539 0       *       =       3030539 0       A
TGAAAATGAAGCCACAACGTACCCAAACCTTTGGGACACAATGAAAGCATTTCTAAGAGGGAAACTCATAGCTCTGAGTACCTCCAAGAAGAAACGGGAG    CCCFFFFFH
HHHHJJJJJJIFHIJJJIIJJJIJJJJJJJJJJJJJJJJJJIJIIJJJIIJJJJJJJHHHHHFFFFFFFEEECEEDDDDDDDDDDDDDDDDDDB9    RG:Z:0
HWI-ST985:73:C08BWACXX:8:2302:12130:48553       89      chr1    3030539 119     101M    =       3030539 0       C
TCCAAGAAGAAACGGGAGAGAGCACATACTAGCAGCTTGACAACACATCTAAAAGCTCTAGAAAAAAAGGAAGCAAATTCACCCAAGAGGAGTAGACGGT    DCDDDDDDD
DDDBDDDDDEEEEEEEDFFFFFFHHHHGIJJJJJHHFJJJJJJJJJJJIGJIJHJJJJJJJJJJJJJJIJJGIGHFJJJJIJHHHHHFFFFFCCC    RG:Z:0  NH:i:3  N
M:i:0   XT:A:R  md:Z:101
HWI-ST985:73:C08BWACXX:8:2208:2017:40383        99      chr1    3055370 180     101M    =       3055454 185     G
ATCTCTGGATATGGCAGTCTCTAGATGGTCCATCCTTTTGTCTCACCTCCAAACTTTGTCTGTGTAACTCTTTCCATTGGTGTTTTGTTCCCAATACTAA    @@@DDDDDF
>=DFEG=EAACHGEHGIIDBH>FHCB@BFHHIIIIIIICBGGIGGIGIIIIIHII@=CHEIGIIIIEECGD@=AHECDDECACCCCCCCC>@    RG:Z:0  NH:i:1  N
M:i:0   XT:A:U  md:Z:101
HWI-ST985:73:C08BWACXX:8:2208:2017:40383        147     chr1    3055454 180     101M    =       3055370 -185    T
TTGTTCCCAATACTAAGAAGGGGCAAAGTGTTGACACTTTGGTCTTCATTCTTCTTGAGTTTCATGTGTTTCACAAATTGTATCTTATATCTTGGGTATT    BDBDCD@@E
C>;CCDEFFFFDE;AC>@71HCCGCG@=ECFEIHFCIGHFFBGHEIIG@IIGGEIJIIIHIIIJIJJHJGGJIGIIGIGF?DHHEBDDD@@B    RG:Z:0  NH:i:1  N
M:i:0   XT:A:U  md:Z:101
HWI-ST985:73:C08BWACXX:8:2103:17437:175854      99      chr1    3197333 254     66M6121N35M     =       3197379 6
268     TGAAGTGTCTGTTGGATTAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGAGTGGAGGATCCTGTTGGATTGTTT    @
CCFFDFFHHHHHJJJJJJJJJJJJJJJJJJHHIIJJJJJJJJJHJJJJJJJJJJJJJIJIJIJJJJJJHHJJJJGIJJJJFHICGIIGHEEFFFFFEEDDEEDCDC    RG:Z:0  N
H:i:1   NM:i:1  XT:A:U  md:Z:66>6121*35
HWI-ST985:73:C08BWACXX:8:2103:17437:175854      147     chr1    3197379 254     20M6121N81M     =       3197333 -
6268    TTTGGTAAGTTCCAATGTTTATGAAAGAAGAGTGGAGGATCCTGTTGGATTGTTTGGCTGGACACTATTACATTGGAACTGTGTTCACAGAATCAAAGCTG    <
DDDDDEEECACFFFFFFHHHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJIHGJJJIJJJJJJIGHJJJJJIJJIJJJJJJJJJJJJJJIJHHHHHFFFFFCCC    RG:Z:0  N
H:i:1   NM:i:1  XT:A:U  md:Z:81>6121*20
```

Cecilia Coimbra Klein

SAM format

Sequence Alignment/Map



```
HWI-ST985:73:C08BWACXX:8:2208:2017:40383        147     chr1    3055454 180     101M    =       3055370 -185    T
TTGTTCCCAATACTAAGAAGGGGCAAAGTGTTGACACTTTGGTCTTCATTCTTCTTGAGTTTCATGTGTTTCACAAATTGTATCTTATATCTTGGGTATT    BDBDCD@@E
C>;CCDEFFFFDE;AC>@71HCCGCG@=ECFEIHFCIGHFFBGHEIIG@IIGGEIJIIIHIIIJIJJHJGGJIGIIGIGF?DHHEBDDD@@B    RG:Z:0  NH:i:1  N
M:i:0   XT:A:U  md:Z:101
HWI-ST985:73:C08BWACXX:8:2103:17437:175854      99      chr1    3197333 254     66M6121N35M     =       3197379 6
268     TGAAGTGTCTGTTGGATTAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGAGTGGAGGATCCTGTTGGATTGTTT    @
CCFFDFFHHHHHJJJJJJJJJJJJJJJJJHHIIJJJJJJJJJHJJJJJJJJJJJIJIJIJJJJJJHHJJJJGIJJJJFHICGIIGHEEFFFFFEEDDEEDCDC    RG:Z:0  N
H:i:1   NM:i:1  XT:A:U  md:Z:66>6121*35
```

Flag:
https://broadinstitute.github.io/picard/explain-flags.html

CIGAR:

- N → intron
- M → match
- I → insertion
- D → deletion
- S → soft-clip

More specification on SAM format:
https://samtools.github.io/hts-specs/SAMv1.pdf

BAM format

compressed binary representation of the SAM format

- specific block compression
  - BGZF
- support random access through the **index**

  ➡ fast retrieval of alignments overlapping a specified region

! BAM file must be sorted by genomic position (chromosome name and leftmost coordinate) in order to be indexed!

CRAM format

improved compressed binary representation of SAM

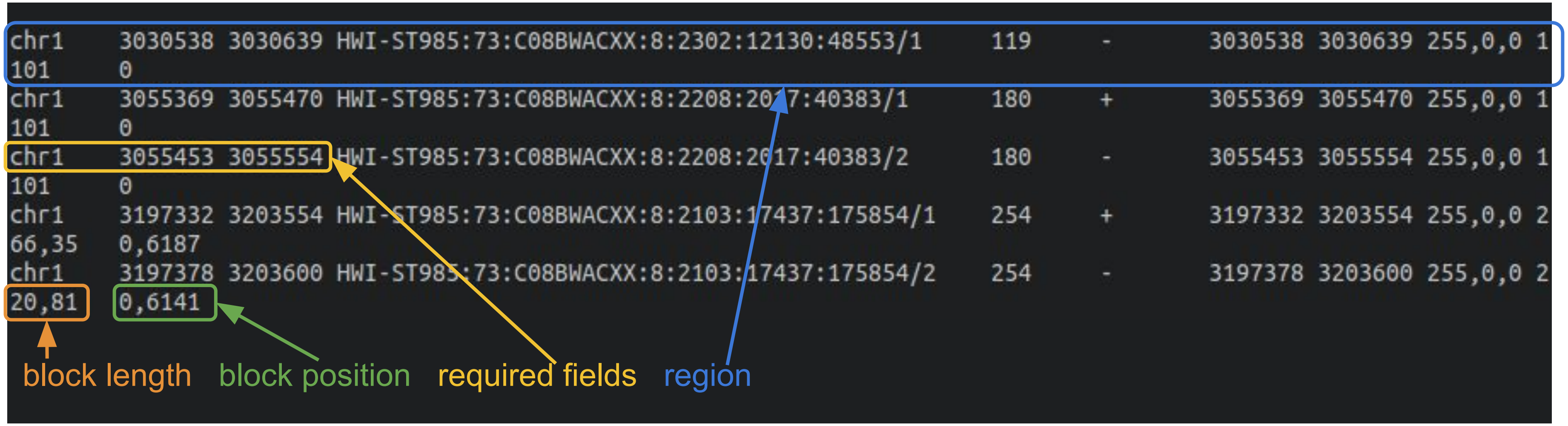
- different compression formats
    - gzip, bzip2, CRAM records
- CRAM records use different encoding strategies, e.g. bases are reference compressed by encoding base differences rather than storing the bases themselves
- random access support through the format itself (slices)

CRAM indexing is external to the file format itself and may change independently of the file format specification in the future

BED format

provides a flexible and compact way to represent genomic regions (with breaks)

- 3 required fields + additional 9 fields
- more compact than GFF ➡ **tradeoff between size and provided information**



```
chr1    3030538 3030639 HWI-ST985:73:C08BWACXX:8:2302:12130:48553/1     119     -       3030538 3030639 255,0,0 1
101     0
chr1    3055369 3055470 HWI-ST985:73:C08BWACXX:8:2208:2017:40383/1      180     +       3055369 3055470 255,0,0 1
101     0
chr1    3055453 3055554 HWI-ST985:73:C08BWACXX:8:2208:2017:40383/2      180     -       3055453 3055554 255,0,0 1
101     0
chr1    3197332 3203554 HWI-ST985:73:C08BWACXX:8:2103:17437:175854/1    254     +       3197332 3203554 255,0,0 2
66,35   0,6187
chr1    3197378 3203600 HWI-ST985:73:C08BWACXX:8:2103:17437:175854/2    254     -       3197378 3203600 255,0,0 2
20,81   0,6141
```

10) blockCount - The number of blocks (exons) in the BED line.

11) blockSizes - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.

12) blockStarts - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

https://genome.ucsc.edu/FAQ/FAQformat.html#format1

bedGraph and wig formats

bedGraph

- allows the display of continuous-valued data
- useful for probability scores and transcriptome data (CHIp-seq, RNA-seq)
- is a text file

```
track type=bedGraph name="BedGraph Format" description="BedGraph format" visibility=full color=200,100,0 altColor=0,100,200
priority=20
chr19 49302000 49302300 -1.0
chr19 49302300 49302600 -0.75
```

wig

- allows the display of continuous-valued data
- more compressed than bedGraph
- is a text file

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

bigBed, bigWig

Useful formats to display data on the UCSC genome browser

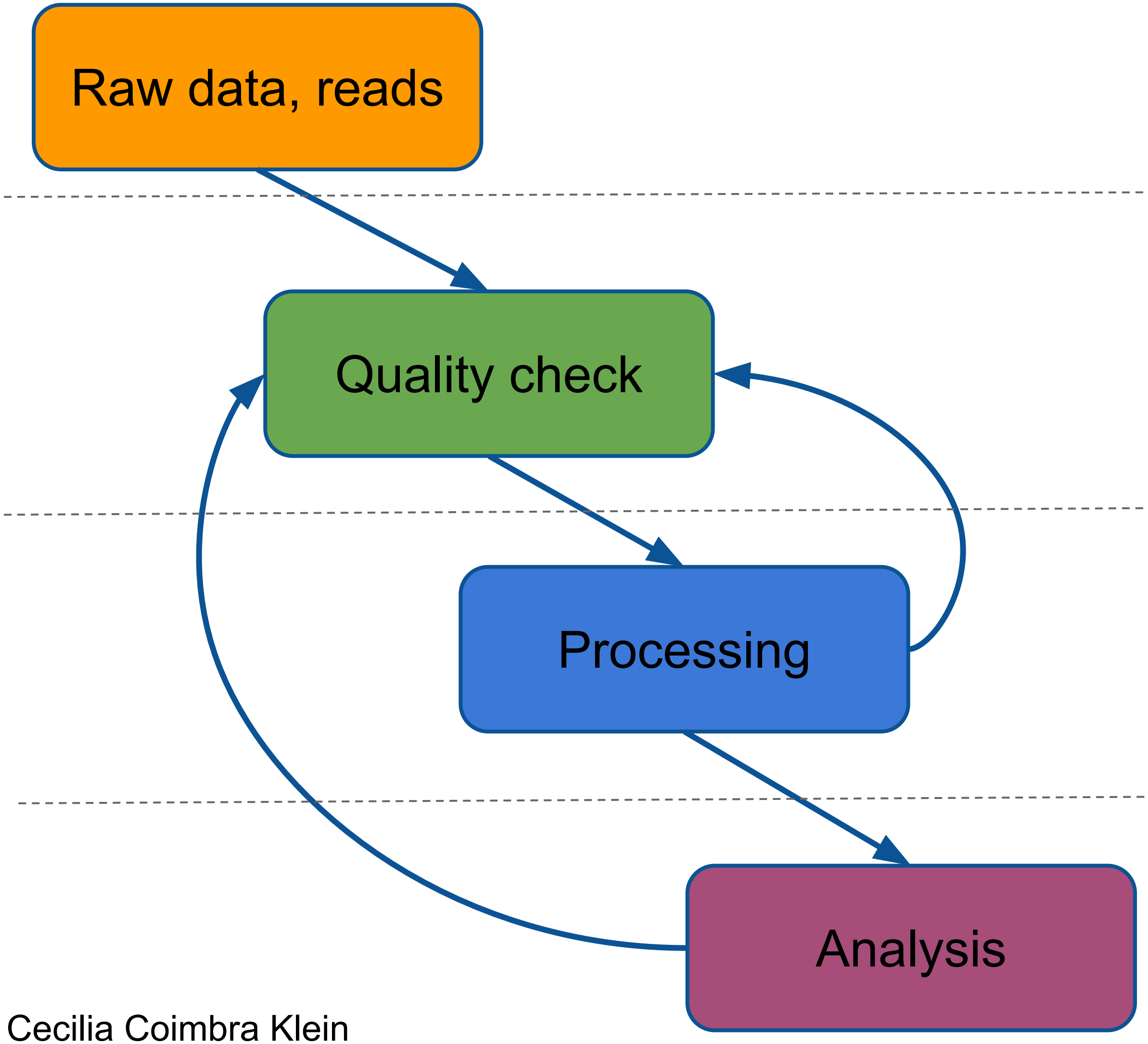
- BED, bedGraph, wig - are tab delimited text files
- bigBed, bigWig - are binary version of this files
- for each type of file there is a specific procedure to make a binary form
  - easily transferable
  - not so big
  - allows indexed access

Hands-on

Data formats 2

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Post-sequencing: usual pipeline



Some data formats

*.fastq, *.fa,
*.sff, *.sra

*.fastq
*.tsv, *.html..

*.sam, *.bam
*.bed, *.wig, *.bw
*.bedgraph
*.gtf, *.fa,..

*.vcf
*.tsv
*.ace, *.agp

Quality check

Quality check

- RNA-seq library preparation/sequencing QC:
    - RNA Integrity Number (RIN), library size distribution

- Pre-mapping QC, raw reads:
    - Sequence quality
    - GC content
    - K-mers overrepresentation
    - Possible contaminants

- Post-mapping QC:
    - Mapping statistics - % reads mapped, % of multimappings, duplicated reads, detected elements, overall gene/transcript coverage, strand specificity...
    - rRNA content
    - Expression profile efficiency
    - Replicates correlation
    - Sample clustering

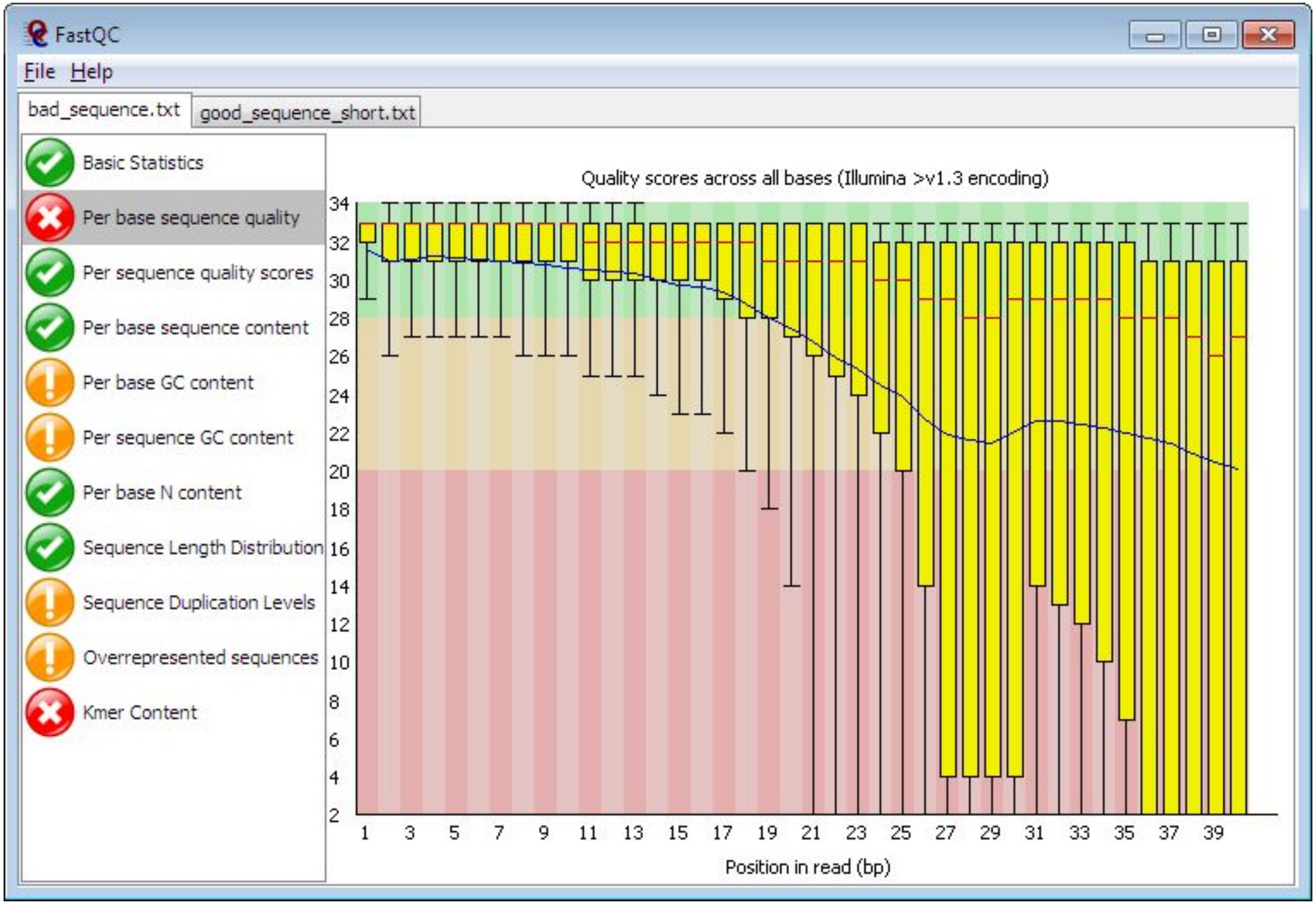
Quality metrics

ENCODE 3 standards for long RNA-seq data:

- Two or more replicates
- Read length >50bp
- >30M uniquely mapped reads
- Spearman correlation >0.8 between replicates
- Metadata control

https://www.encodeproject.org/rna-seq/long-rnas/

FastQC



FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Cecilia Coimbra Klein

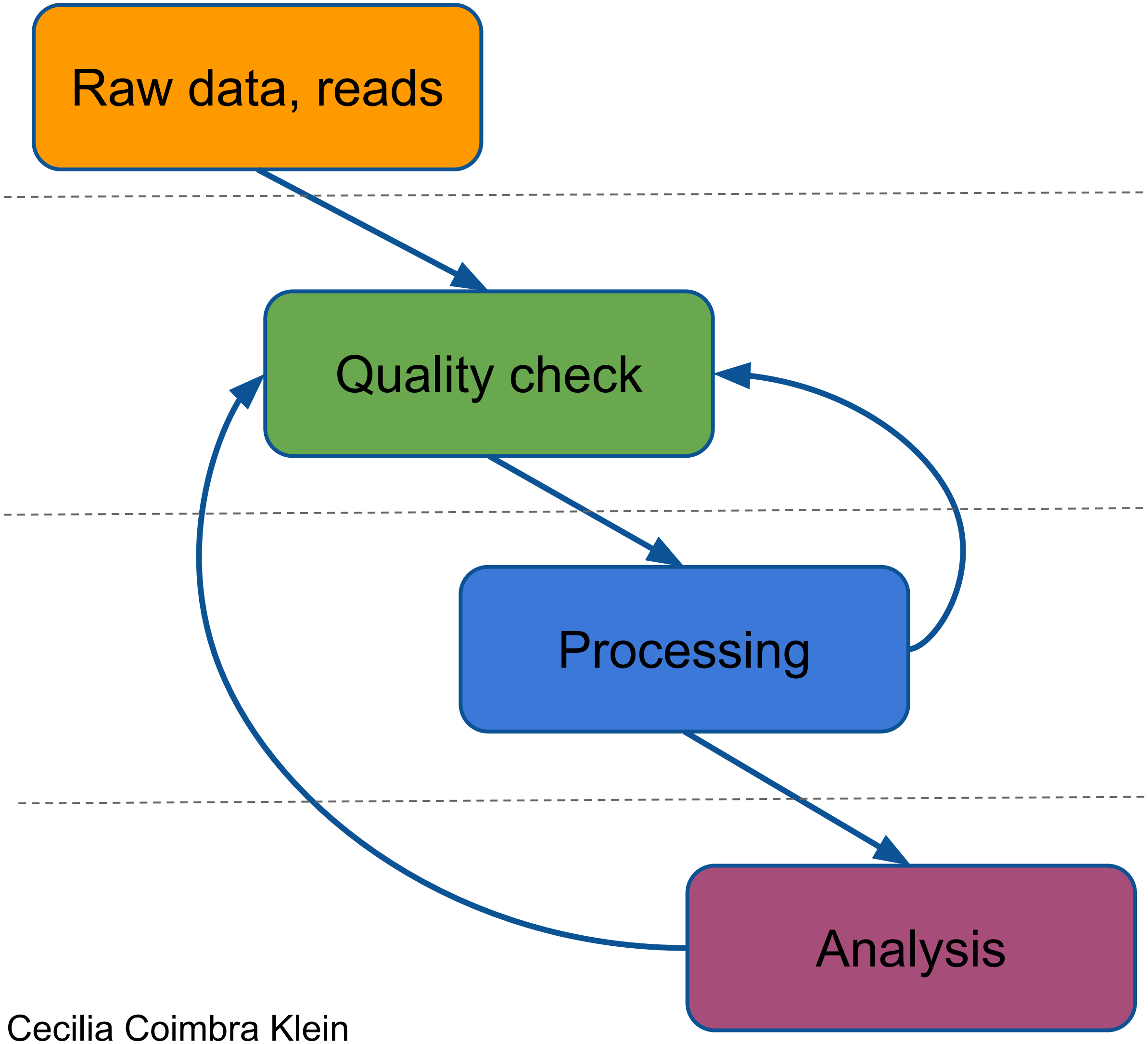
Hands-on

Reference annotation 3.1

Fastq files and read QC 3.2

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Post-sequencing: usual pipeline



Some data formats

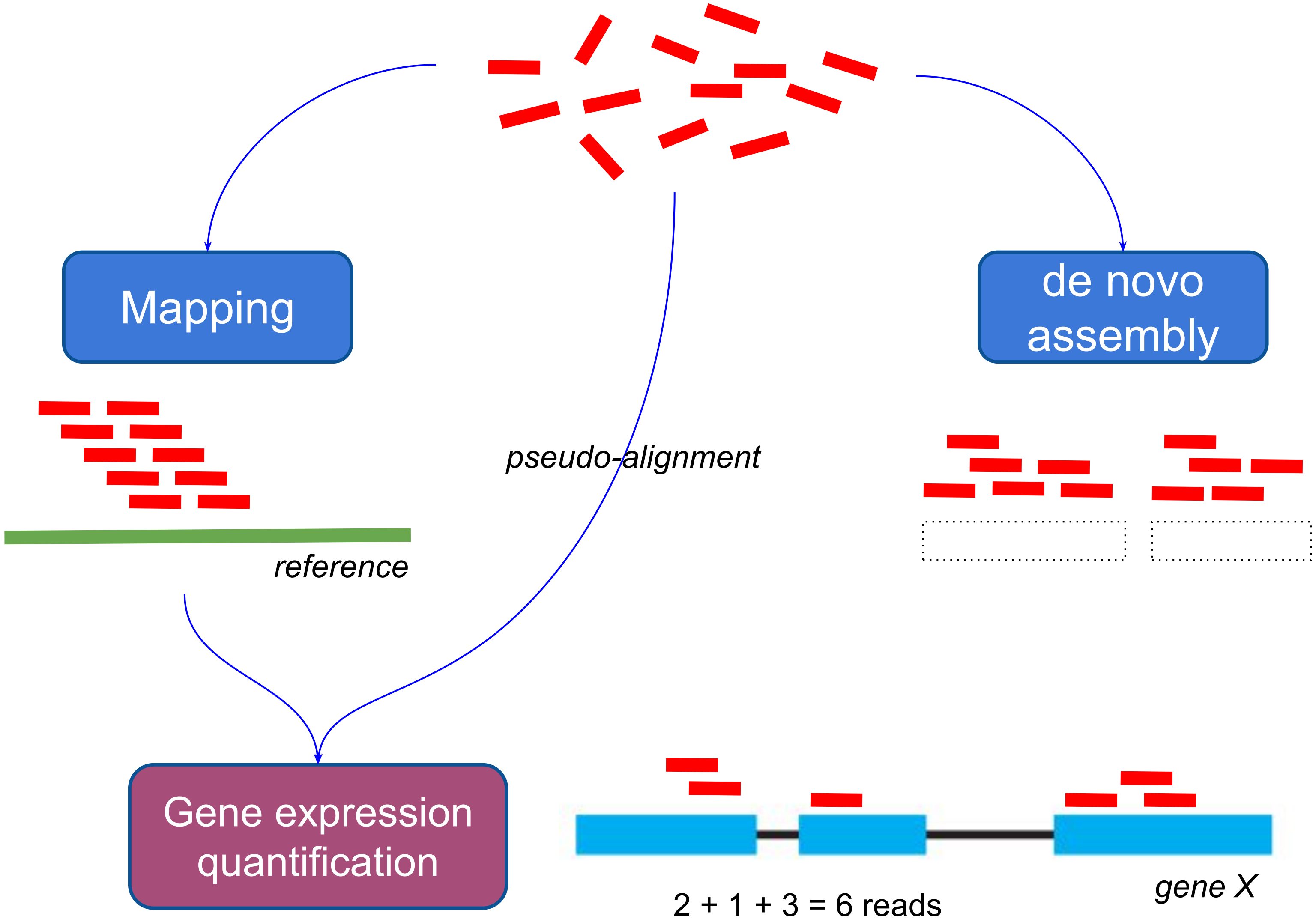
*.fastq, *.fa,
*.sff, *.sra

*.fastq
*.tsv, *.html..

*.sam, *.bam
*.bed, *.wig, *.bw
*.bedgraph
*.gtf, *.fa,..

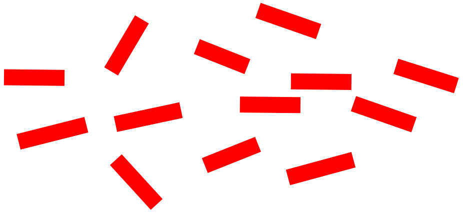
*.vcf
*.tsv
*.ace, *.agp

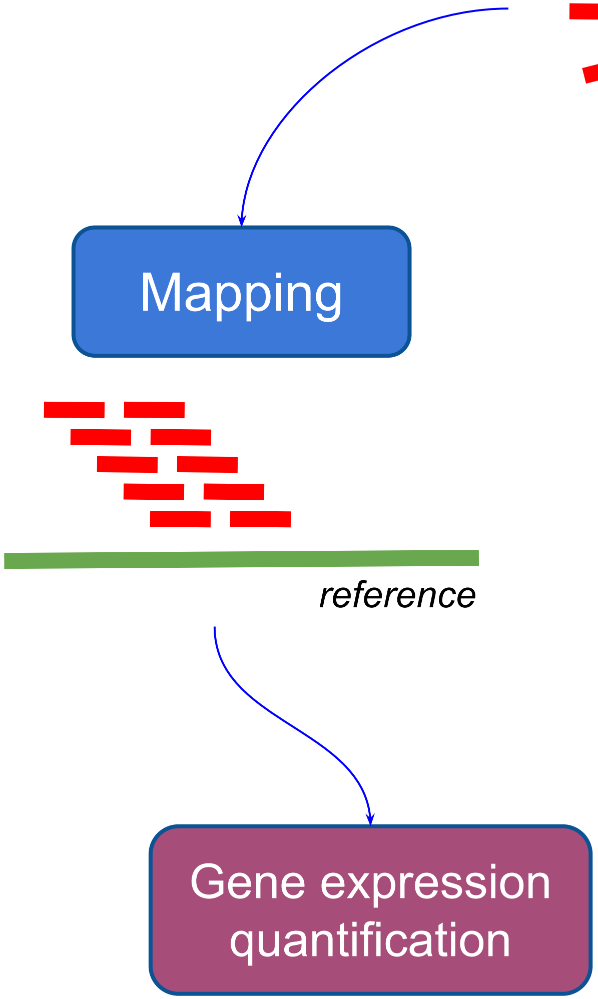
Processing


Mapping
de novo assembly
pseudo-alignment
reference
Gene expression quantification
2 + 1 + 3 = 6 reads
gene X

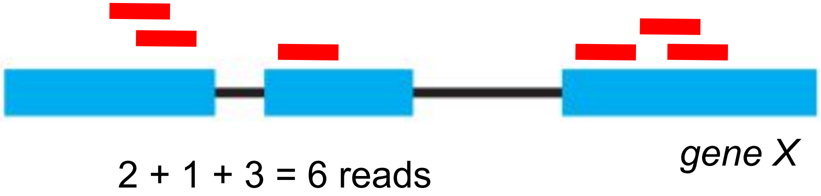
Mapping strategy

Mapping





Find a correspondence between the query sequences (RNA-seq reads) and our prior knowledge (reference genome sequence, reference gene annotation).



Alignment

A common technique for mapping is alignment:

```
Reference:  CATGGAACTTATCTCACAGCCTTT
     Read:      GAACTT–TCGCA
```

Not always easy:

- Reads are short with respect to the genome (~100 bp)
- Human genome is ~3G bp long and rather repetitive
- Reference genome is different from sample genome (SNPs, indels, structural variants)
- Reads are prone to errors (if lucky 1/1000 base calls are wrong)

Alignment - basic concepts

- online vs indexed
- global vs local
- sequence similarity
  - mismatches as base substitutions (A→T)
  - insertions/deletions or gaps
  - block transpositions or rearrangements
- multimaps
- heuristic vs exhaustive

  Given a metric distance (eg. mismatches) and a threshold (eg. 96% homology) the alignment is exhaustive if it contains all possible matches in the reference for that distance and threshold

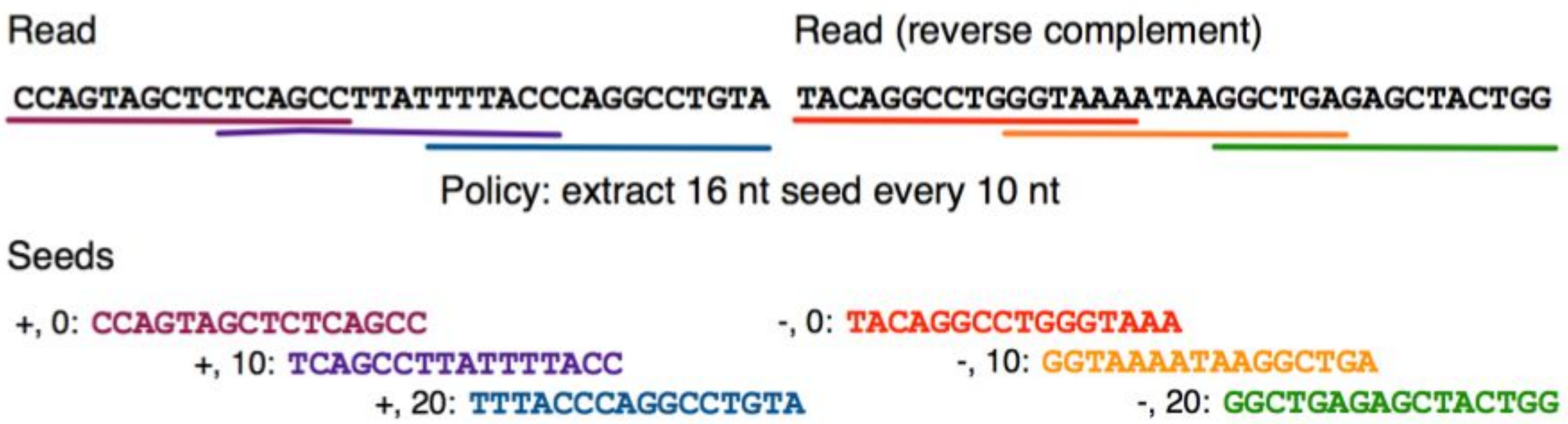
Indices

Pre-compute the reference text into an index providing fast sorted access to substrings of the reference

- indexing the **reference** (most common choice):
  - each read is mapped individually
  - references usually have big size but are fixed
  - read/sample size unknown and variable
- indexing the **reads:**
  - reference is scanned to perform the mapping
  - makes sense with small references (e.g. Yeast)
- indexing **both** the reference and the reads:
  - high memory consumption - keeps both indices

Mapping algorithms - seed-and-extend

i. extract seeds (usually exact)
ii. lookup each of them into the index
iii. “extend” the search to validate the alignments



Read

CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA

Read (reverse complement)

TACAGGCCTGGGTAAAATAAGGCTGAGAGCTACTGG

Policy: extract 16 nt seed every 10 nt

Seeds

+, 0: CCAGTAGCTCTCAGCC
+, 10: TCAGCCTTATTTTACC
+, 20: TTTACCCAGGCCTGTA

-, 0: TACAGGCCTGGGTAAA
-, 10: GGTAAAATAAGGCTGA
-, 20: GGCTGAGAGCTACTGG

sensitivity depends on seed length and overlap

⟶ poor choice of seed might lead to unmapped reads

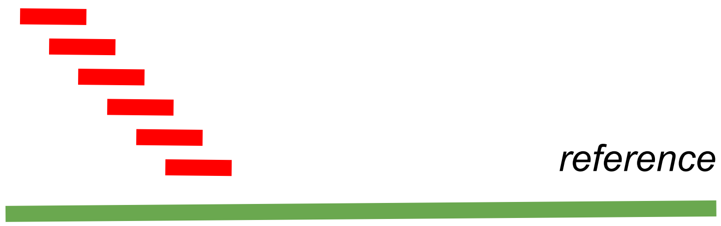
⟶ not exhaustive

Paired-end alignment

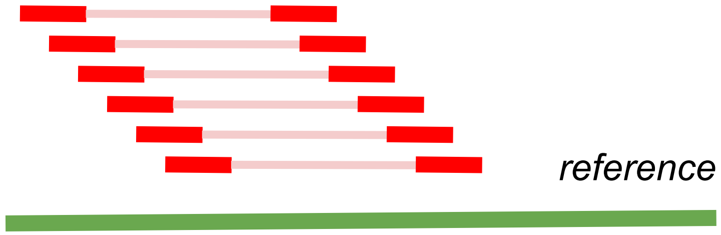
Both ends of the fragments are sequenced→paired-end reads

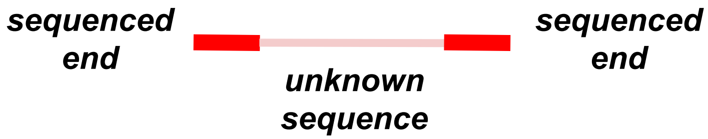
- connectivity information
- insert size and read length are known in advance (from library preparation)
- insert size distribution can be used to solve ambiguities (or even enhance the mapping process)

Single-end (SE) reads



Paired-end (PE) reads



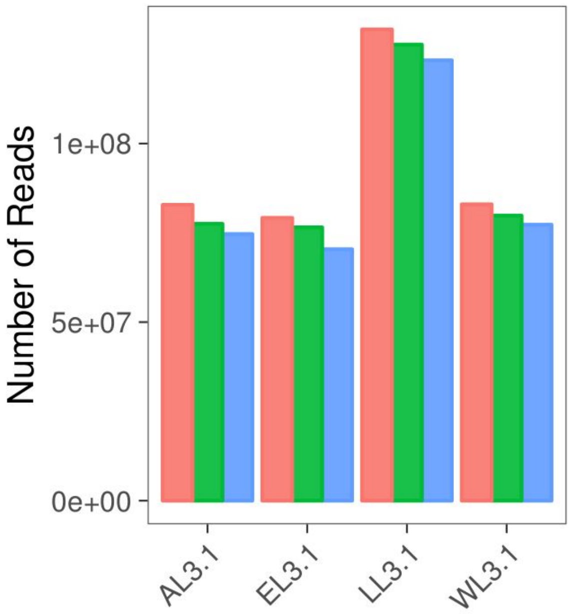


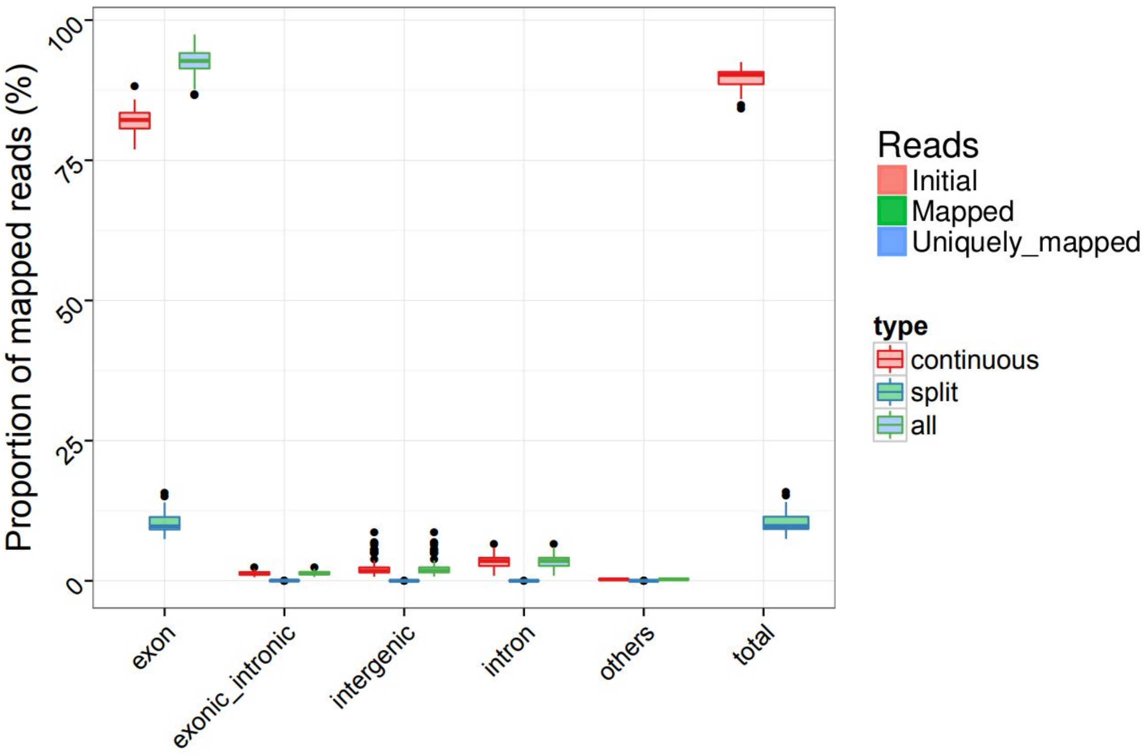
RNA-seq mapping

Specific variables to consider when mapping RNA-seq data

- intron size
- overhang
    - number of bases from each side of the junction that should be covered by the read
- splice site consensus
    - donor/acceptor splice site consensus sequences
- junction *"filtering"*:
    - chromosome/strand
    - block order
    - min/max distance

Mapping statistics





- total reads
- mapped reads (number and %)
- uniquely mapped reads (number and %)
- mappings (including multimaps)
- genomic regions (number and %)

Cecilia Coimbra Klein

Hands-on

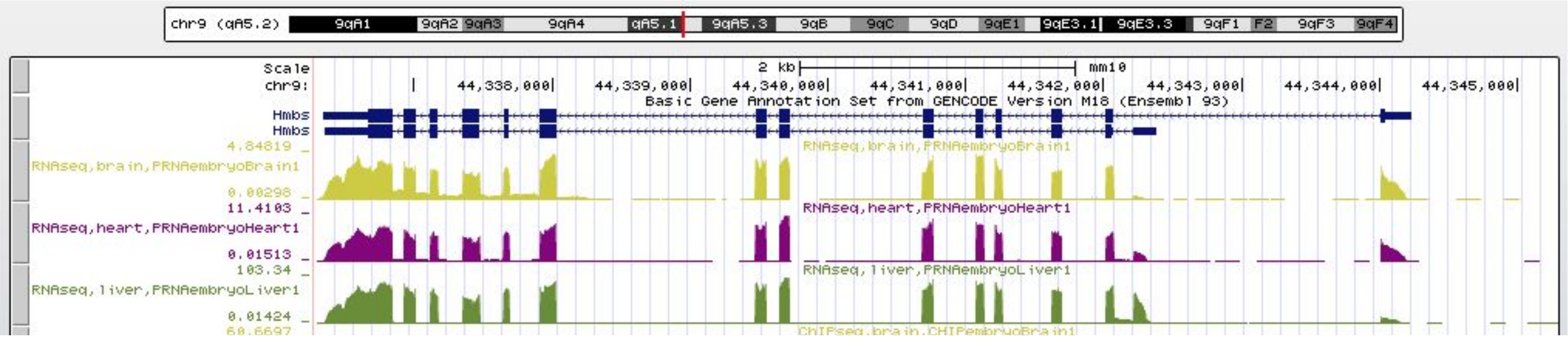
Mapping 3.3

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

RNA-seq signal

RNA-seq signal

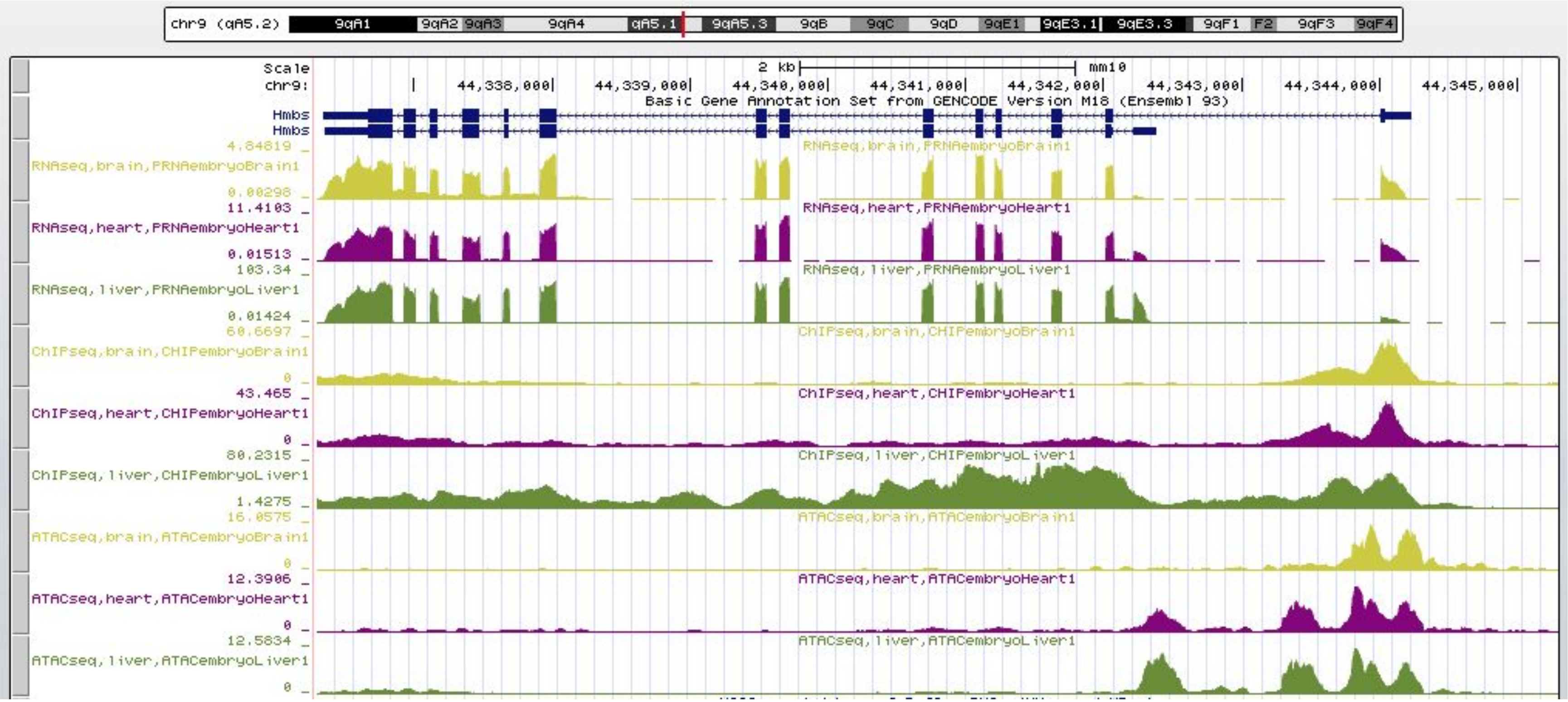
genome-euro.ucsc.edu



- expected read depth at each position in the genome
- can be normalized (e.g. RPM, reads per million reads)

UCSC: signal files

genome-euro.ucsc.edu


chr9 (qA5.2)
9qA1
9qA2
9qA3
9qA4
qA5.1
9qA5.3
9qB
9qC
9qD
9qE1
9qE3.1
9qE3.3
9qF1
F2
9qF3
9qF4
Scale
chr9:
2 kb
mm10
44,338,000|
44,339,000|
44,340,000|
44,341,000|
44,342,000|
44,343,000|
44,344,000|
44,345,000|
Basic Gene Annotation Set from GENCODE Version M18 (Ensembl 93)
Hmbs
Hmbs
4.84819 _
RNAseq,brain,PRNAembryoBrain1
0.00298 _
11.4103 _
RNAseq,heart,PRNAembryoHeart1
0.01513 _
103.34 _
RNAseq,liver,PRNAembryoLiver1
0.01424 _
60.6697 _
ChIPseq,brain,CHIPembryoBrain1
0 _
43.465 _
ChIPseq,heart,CHIPembryoHeart1
0 _
80.2315 _
ChIPseq,liver,CHIPembryoLiver1
1.4275 _
16.0575 _
ATACseq,brain,ATACembryoBrain1
0 _
12.3906 _
ATACseq,heart,ATACembryoHeart1
0 _
12.5834 _
ATACseq,liver,ATACembryoLiver1
0 _

Hands-on

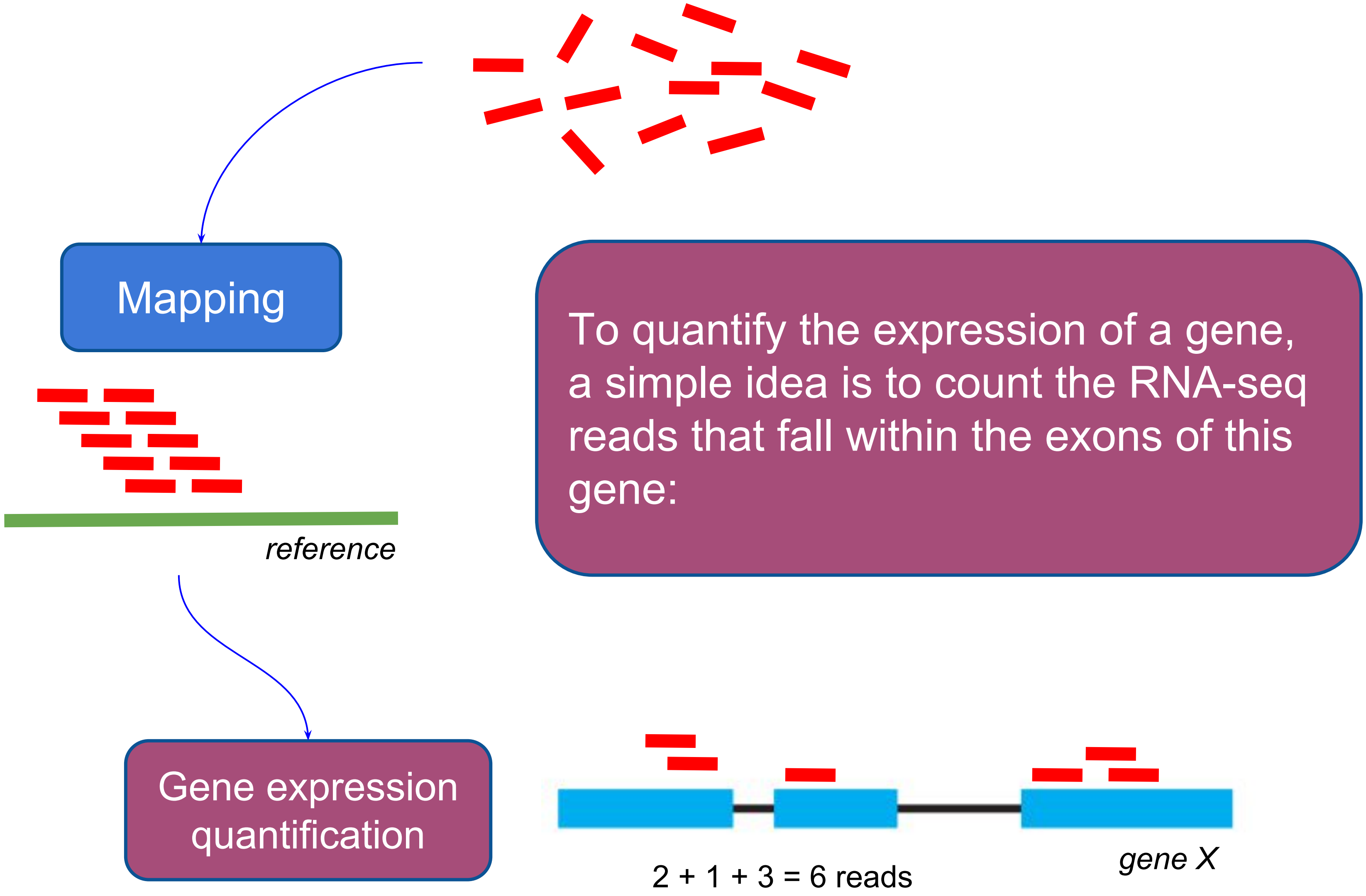
RNA-seq signal files 3.4

UCSC genome browser 3.5

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Gene expression quantification

Gene expression quantification

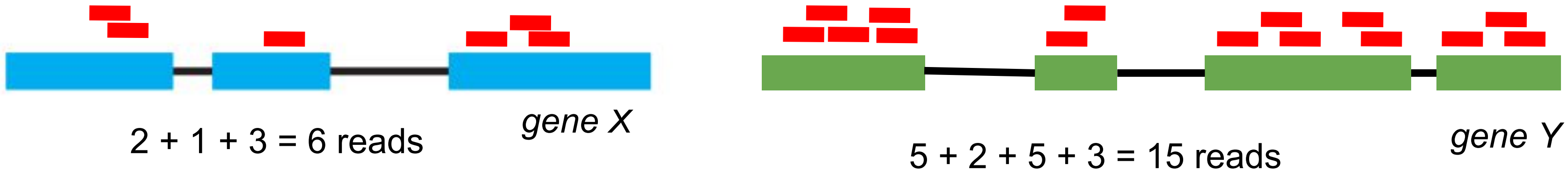


To quantify the expression of a gene, a simple idea is to count the RNA-seq reads that fall within the exons of this gene:

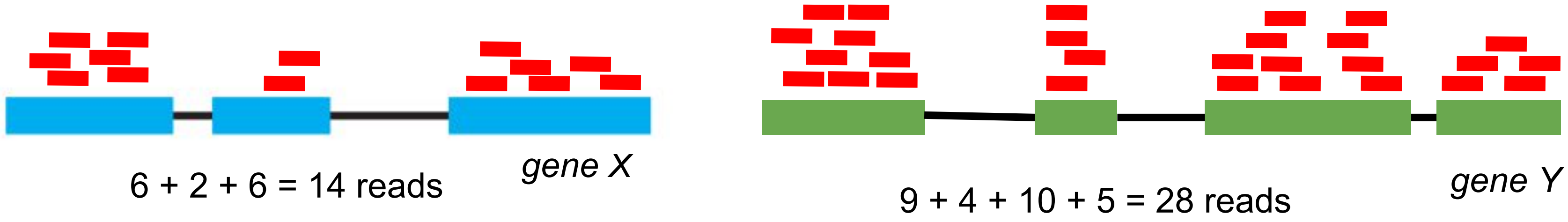
Cecilia Coimbra Klein

Gene expression quantification

- In *experiment A*, long genes (in terms of exon length) will get more reads than small genes



- In *experiment B* with a high number of mapped reads, a gene will get more reads than in an experiment with a small number of mapped reads



Gene expression quantification

- Mortazavi et al. (2008) introduced RPKM = Read Per Kilobase of exon model per Million mapped reads, which normalizes the read count of a gene in an experiment by both:
  - the length of the gene
  - the number of mapped reads in the experiment

$$RPKM = \frac{mapped\ reads * 10^9}{Tot\ mapped\ reads * Length}$$

- FPKM = Fragments Per Kilobase of exon model per Million mapped reads

Paired-end RNA-Seq experiments produce two reads per fragment (not necessarily both reads will be mappable). To avoid double-count some fragments but not others, FPKM is calculated by counting fragments, not reads.

Gene expression quantification

- RPKM is now widely used for assessing gene expression, however it assumes that the absolute amount of total RNA in each cell is similar across different cell types or experimental perturbations, which is not always the case (Loven, 2012)

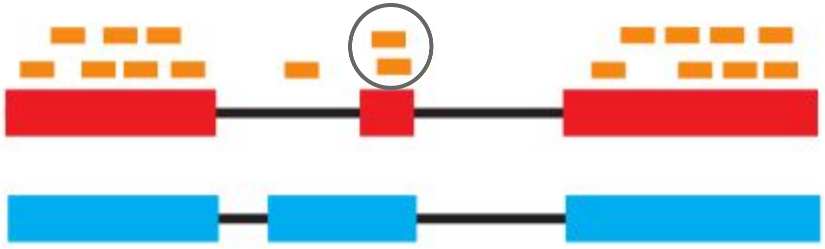
- For example, Mortazavi et al. (2008) estimates that 3 RPKM corresponds to ~ 1 transcript per cell in mouse liver, while Klish et al. (2011) say that 1 RPKM corresponds to between 0.3 and 1 transcript per cell...

$$TPM_g = \frac{RPKM_g}{\sum_g RPKM_g}$$

Li, Ruotti, Stewart, Thomson, Dewey, “RNA-seq gene expression estimation with read mapping uncertainty”, *Bioinformatics*, 26(4), 2010, 493-500.

Individual transcript expression

- Gene expression is quite easy to compute, however estimating the expression of individual transcripts of each gene is a difficult problem:



⇨ Do the two circled reads come from the red or from the blue transcript?

- Read deconvolution or transcript isoform quantification

- There are 2 categories of transcript isoform quantifiers :

    - read-centric (Cufflinks, IsoEM, RSEM, Sailfish, eXpress, Kallisto)
    - exon-centric (Poisson model, linear regression approaches like rQuant, IsoLasso, SLIDE, flux capacitor)

Hands-on

Transcript and gene expression quantification 3.6

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Summary

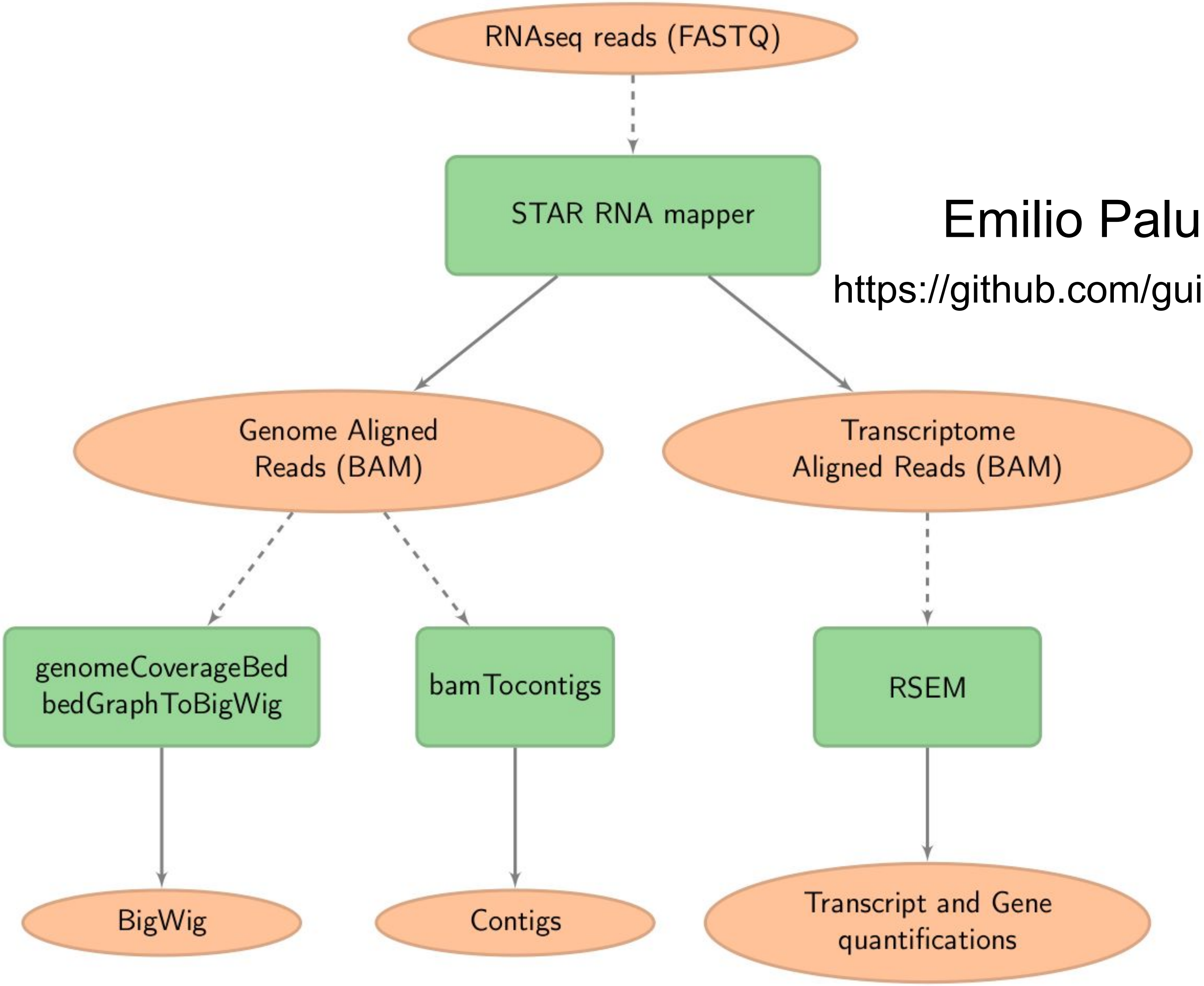
Outline

- Basic concepts
- Reference gene annotation
- Next generation sequencing
- RNA-seq experimental protocols
- Short-read RNA-seq data processing
  - mapping
  - visualisation of gene expression signal
  - gene expression quantification
- RNA-seq data analysis
  - sample clustering based on gene expression
  - differential gene expression
  - gene ontology (GO) term enrichment
  - differential splicing analysis

Outline

- ChIP-seq data processing
  - mapping
  - peak calling
  - visualisation of signal
- ChIP-seq data analysis
  - genomic locations
  - differential peaks per tissue
  - BED files in UCSC browser
- Integrative data analysis
  - promoter regions of differentially expressed genes
  - ATAC-seq signal in the UCSC genome browser
  - promoter regions of differentially spliced genes
  - omics portals

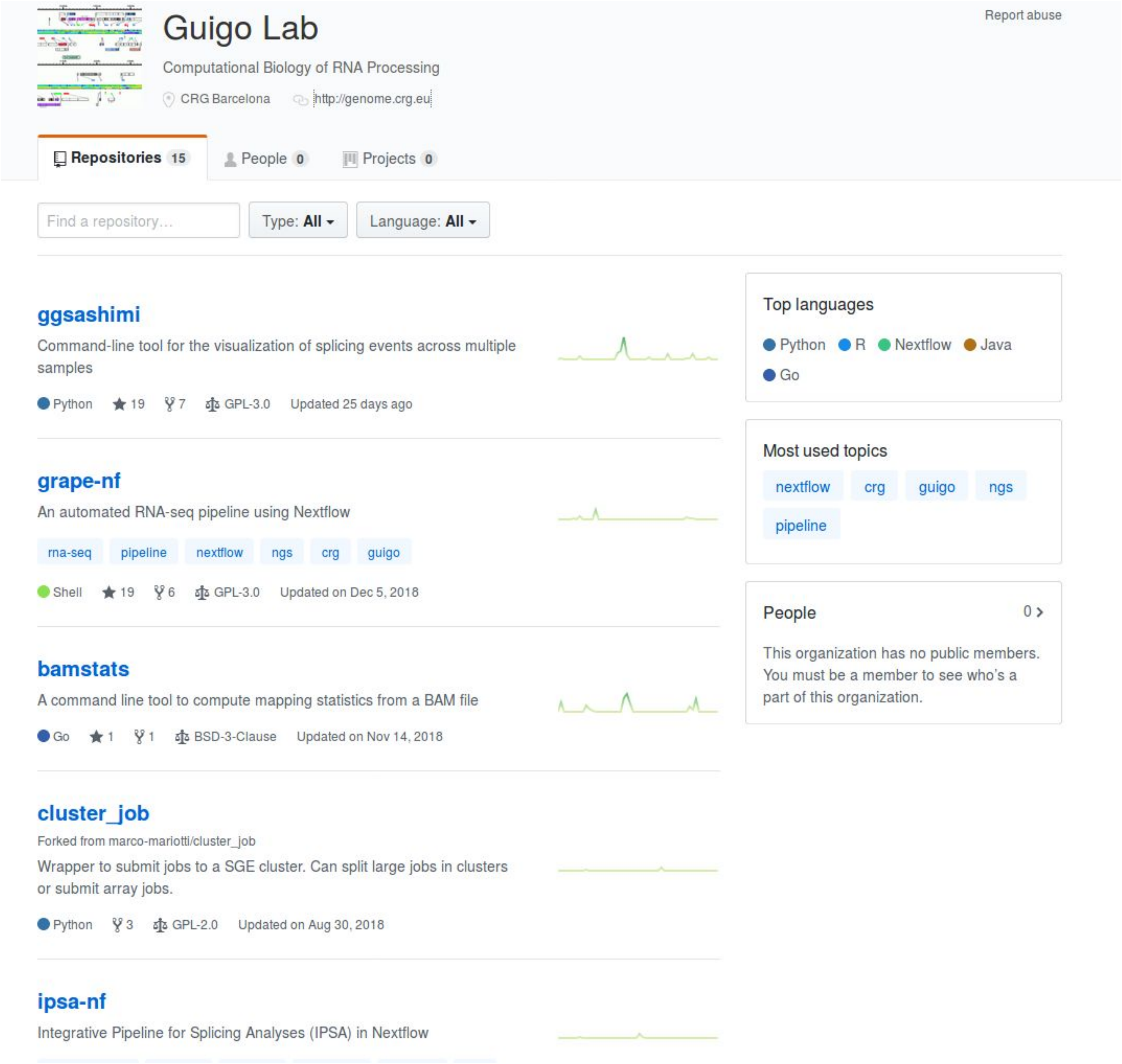
Grape pipeline


RNAseq reads (FASTQ)
STAR RNA mapper
Genome Aligned Reads (BAM)
Transcriptome Aligned Reads (BAM)
genomeCoverageBed bedGraphToBigWig
bamTocontigs
RSEM
BigWig
Contigs
Transcript and Gene quantifications

Emilio Palumbo, CRG

https://github.com/guigolab/grape-nf

Github Guigo Lab



Guigo Lab

Computational Biology of RNA Processing

CRG Barcelona http://genome.crg.eu

Report abuse

Repositories 15 People 0 Projects 0

Find a repository… Type: All ▾ Language: All ▾

ggsashimi

Command-line tool for the visualization of splicing events across multiple samples

● Python ★ 19 7 GPL-3.0 Updated 25 days ago

grape-nf

An automated RNA-seq pipeline using Nextflow

rna-seq pipeline nextflow ngs crg guigo

● Shell ★ 19 6 GPL-3.0 Updated on Dec 5, 2018

bamstats

A command line tool to compute mapping statistics from a BAM file

● Go ★ 1 1 BSD-3-Clause Updated on Nov 14, 2018

cluster_job

Forked from marco-mariotti/cluster_job

Wrapper to submit jobs to a SGE cluster. Can split large jobs in clusters or submit array jobs.

● Python 3 GPL-2.0 Updated on Aug 30, 2018

ipsa-nf

Integrative Pipeline for Splicing Analyses (IPSA) in Nextflow

Top languages

● Python ● R ● Nextflow ● Java ● Go

Most used topics

nextflow crg guigo ngs pipeline

People 0 >

This organization has no public members. You must be a member to see who's a part of this organization.

With RNA-seq you can do..

❑ Study of annotated gene and transcript expression

❑ Assemble novel transcripts with and without reference genome

❑ Novel genome annotation

❑ Splicing analysis

❑ Chimeric-transcript analysis

❑ Variation detection, including genome variation

❑ Allele-specific analysis

❑ Study of post-translational modification, i.e RNA editing

❑ QTL mapping

http://www.rna-seqblog.com

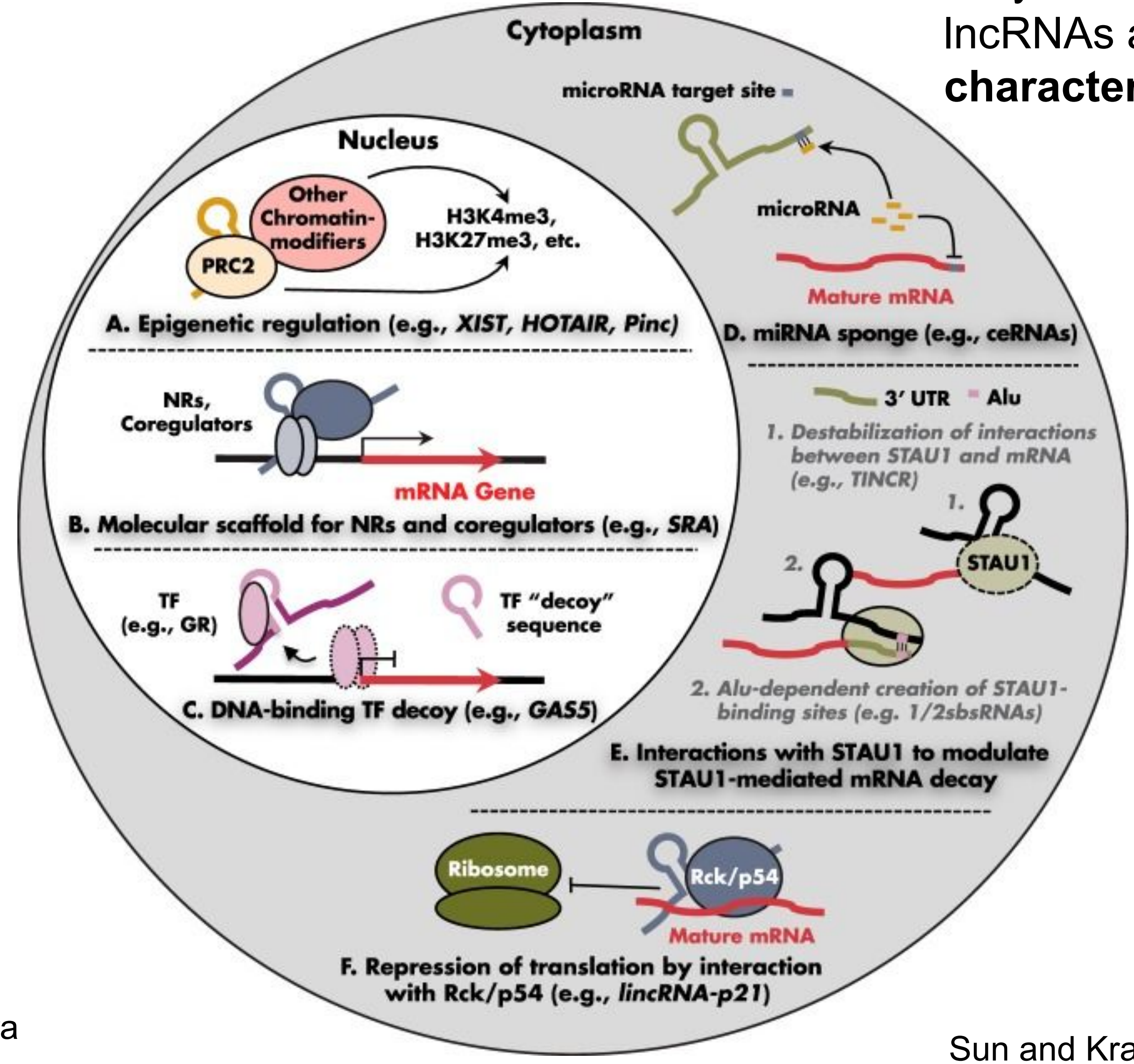
Some references

1. Ensembl: Curwen,..., Clamp, The Ensembl automatic gene annotation system, Genome Res, 2004
2. Flicek,...,Searle, Ensembl 2013. Nucleic Acids Res, 2013 / http://www.ensembl.org/index.html
3. UCSC: Hsu,..., Haussler, The UCSC Known Genes, Bioinformatics, 2006 / http://genome.ucsc.edu/
4. Gencode: Harrow,...,Hubbard, GENCODE: the reference human genome annotation for The ENCODE Project, Genome Res, 2012
5. Metzker, Sequencing technologies - the next generation, Nat Rev Genet, 2010
6. Ruffalo,..., Koyutürk, Comparative analysis of algorithms for next-generation sequencing read alignment, Bioinformatics, 2011.
7. SEQC project: NATURE BIOTECHNOLOGY, Volume 32, Number 9, Sept. 2014
8. RPKM definition: Mortazavi,..., Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat Methods, 2008.
9. Choi et al., Increasing gene discovery and coverage using RNA-seq of globin RNA reduced porcine blood samples, BMC Genomics, 2014
10. Au KF, et al. Characterization of the human ESC transcriptome by hybrid sequencing. PNAS 2013, doi: 10.1073/pnas.1320101110
11. Bolisetti et al., Determining exon connectivity in complex mRNAs by nanopore sequencing, 2015
12. Tarazona et al., Differential expression in RNA-seq:a matter of depth, Genome Res., 2011
13. https://en.wikipedia.org/wiki/FASTQ_format#Encoding
14. Haas BJ, Zody MC. Advancing RNA-Seq analysis. Nat Biotechnol. 2010 May;28(5):421-3. doi: 10.1038/nbt0510-421.
15. Robinson, Mark D., and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data." Genome Biol 11.3 (2010): R25.
16. Lovén J, et al. Revisiting global gene expression analysis. Cell. 2012 Oct 26;151(3):476-82.
17. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000Res. 2015 Oct 14;4:1070.

Additional slides

Examples of ncRNA functions

Only **~2%** of human lncRNAs are **functionally characterized**

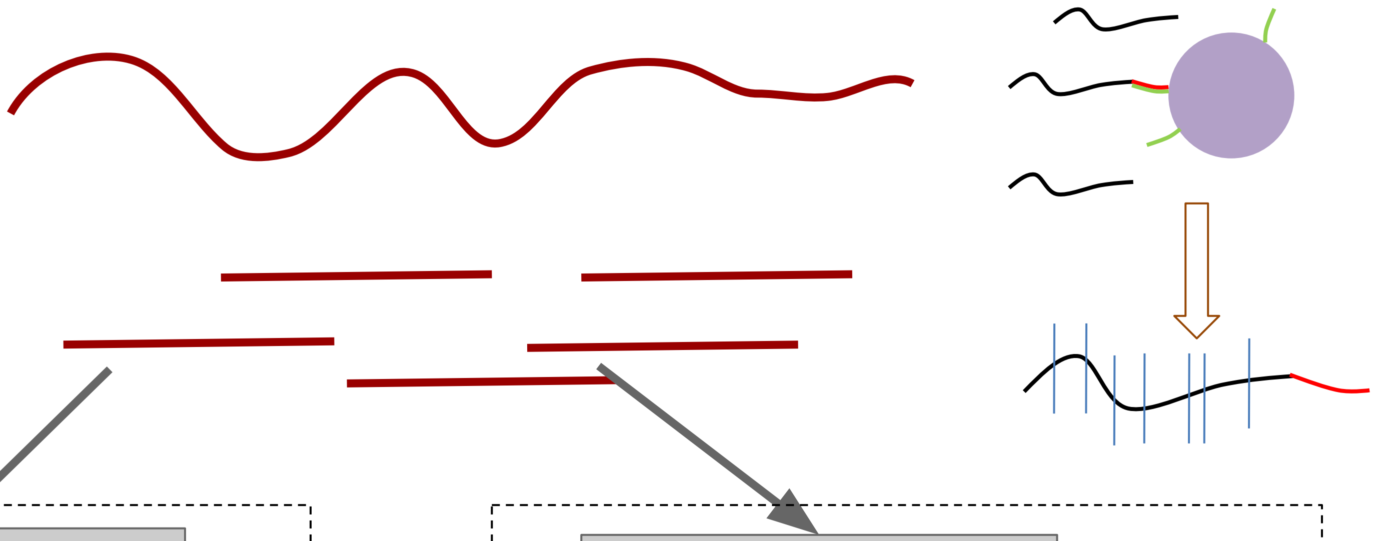

Cytoplasm
Nucleus
Other Chromatin-modifiers
PRC2
H3K4me3, H3K27me3, etc.
A. Epigenetic regulation (e.g., *XIST, HOTAIR, Pinc*)
NRs, Coregulators
mRNA Gene
B. Molecular scaffold for NRs and coregulators (e.g., *SRA*)
TF (e.g., GR)
TF "decoy" sequence
C. DNA-binding TF decoy (e.g., *GAS5*)
microRNA target site
microRNA
Mature mRNA
D. miRNA sponge (e.g., ceRNAs)
3′ UTR
Alu
1. Destabilization of interactions between STAU1 and mRNA (e.g., TINCR)
1.
2.
STAU1
2. Alu-dependent creation of STAU1-binding sites (e.g. 1/2sbsRNAs)
E. Interactions with STAU1 to modulate STAU1-mediated mRNA decay
Ribosome
Rck/p54
Mature mRNA
F. Repression of translation by interaction with Rck/p54 (e.g., *lincRNA-p21*)

Sun and Kraus 2015

Library preparation

RNA purification

Fragmentation



Non-stranded protocol

- Reverse transcription (1st and 2nd strand)
- Adenylation of 3’ ends
- Adapter ligation
- PCR amplification

Stranded protocol

- Reverse transcription (1st only)
- Using dUTP instead of dTTP for the second strand cDNA synthesis
- Adenylation of 3’ ends
- Adapter ligation
- Degradation of the second strand
- PCR amplification

Library preparation, stranded

Note:
Elimination of the second strand may be different between protocols. Some protocols (used by ENCODE and Blueprint) digest the second strand by using a UDGase (enzyme that digests the Uracil strand). More recent protocols use a DNA polymerase that is not able to amplify the Uracil strand and, thus, only enriches the 1st strand

Index structures

- Hash based
  - Simple Idea -> Store k-mers/seeds/samples using some hash function H(·)
  - Usually requires a lot of space (several times the reference size)

- SuffixArrays
  - Sort suffixes of the text, storing the sorted positions in an array

- FM-Index (BWT Based)
  - Same logic as SuffixArrays
  - Based on a compression scheme (BZIP)
    - Space efficient (sizes the same as the reference)

STAR: Suffix arrays

1	2	3	4	5	6	7	8	9	10	11	12
a	b	r	a	c	a	d	a	b	r	a	$

Index in text	Suffix
1	abracadabra$
2	bracadabra$
3	racadabra$
4	acadabra$
5	cadabra$
6	adabra$
7	dabra$
8	abra$
9	bra$
10	ra$
11	a$
12	$

Suffix Array	Sorted Suffix
12	$
11	a$
8	abra$
1	abracadabra$
4	acadabra$
6	adabra$
9	bra$
2	bracadabra$
5	cadabra$
7	dabra$
10	ra$
3	racadabra$

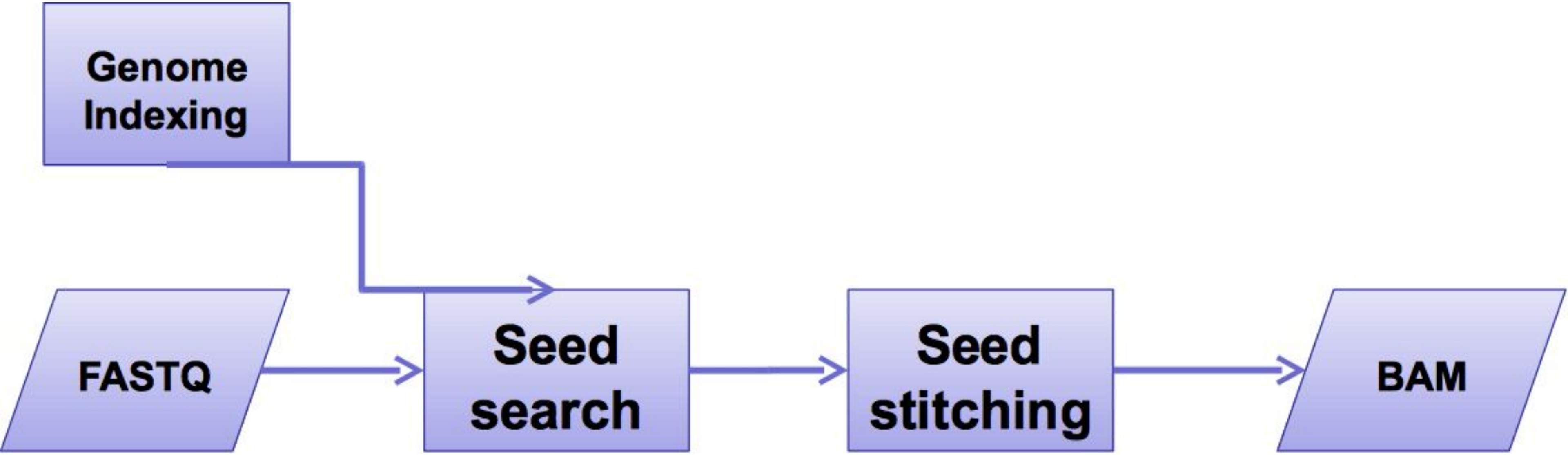
The STAR software

- STAR: Spliced Transcripts Alignment to a Reference

- fast, *de novo* detection of canonical junctions and can discover non-canonical splice and chimeric transcripts; but truncate reads and produces some FP junctions

- has a potential for accurately align long (several kilobases) reads that are emerging from the third-generation sequencing technologies

STAR workflow

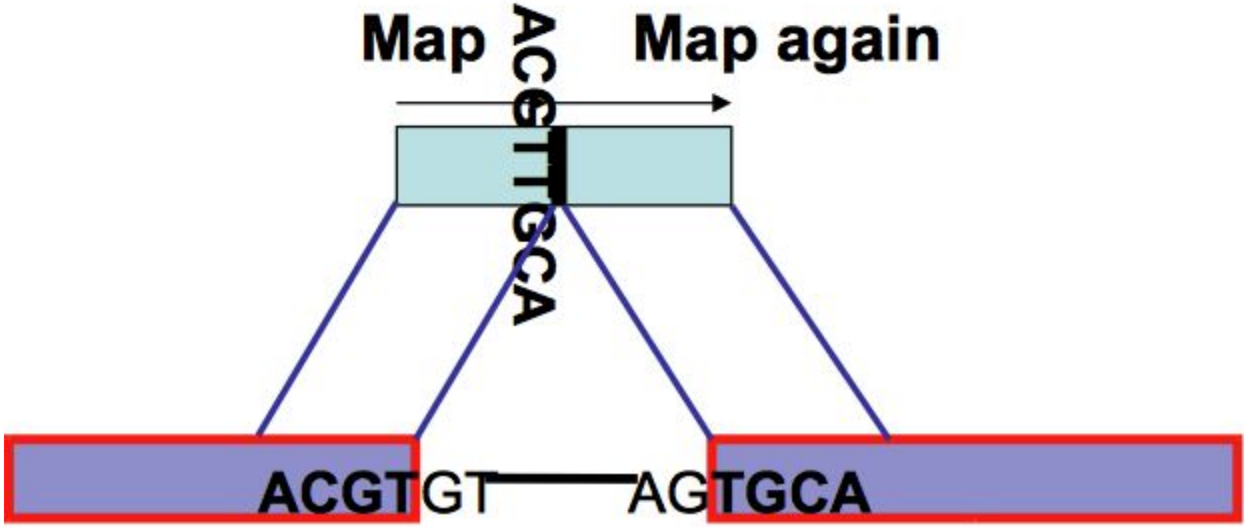


Alex Dobin,CSHL

Seed search: basic idea

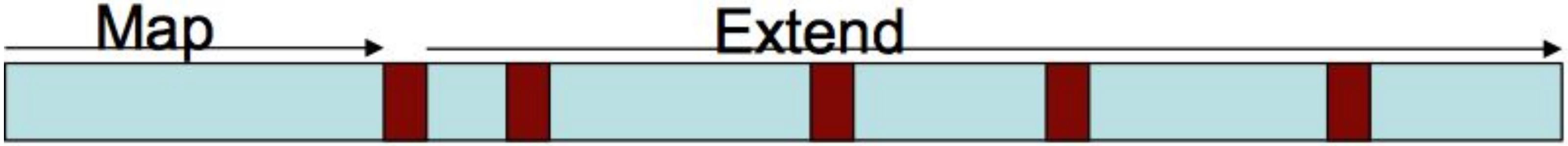
- “Consecutive maximal exact prefix search”

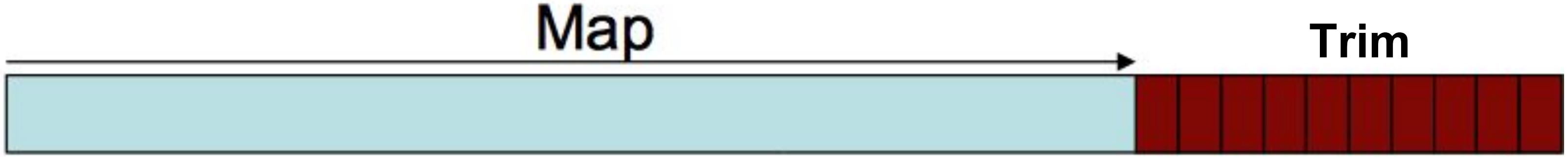
- MEM, Maximal Exact Match: Mummer, MAUVE
- BWA-MEM, Cushaw2, GEM



Alex Dobin, CSHL

Mismatches and tails





A-tail, or adapter,
or poor quality tail

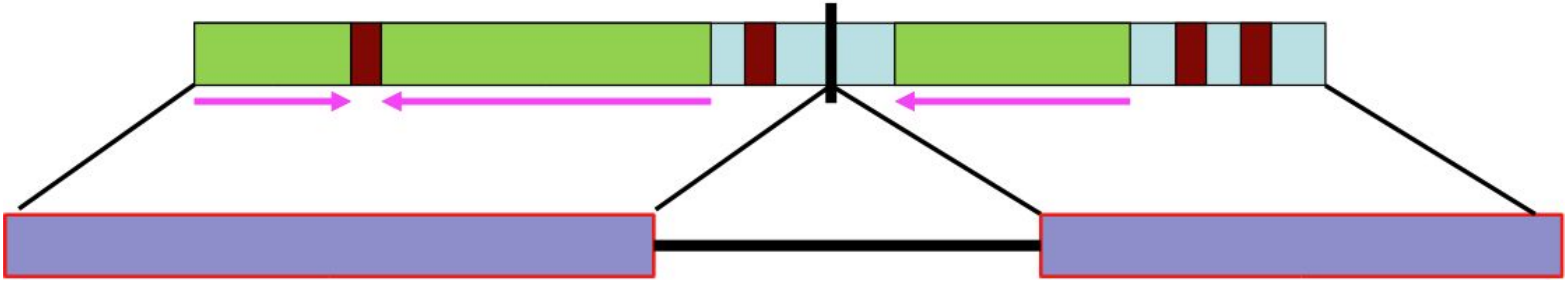
Alex Dobin, CSHL

Seed stitching strategy

- Most DNA aligners use seed-extend paradigm

- STAR uses “seed stitching” strategy:

build the best local alignment out of all seeds

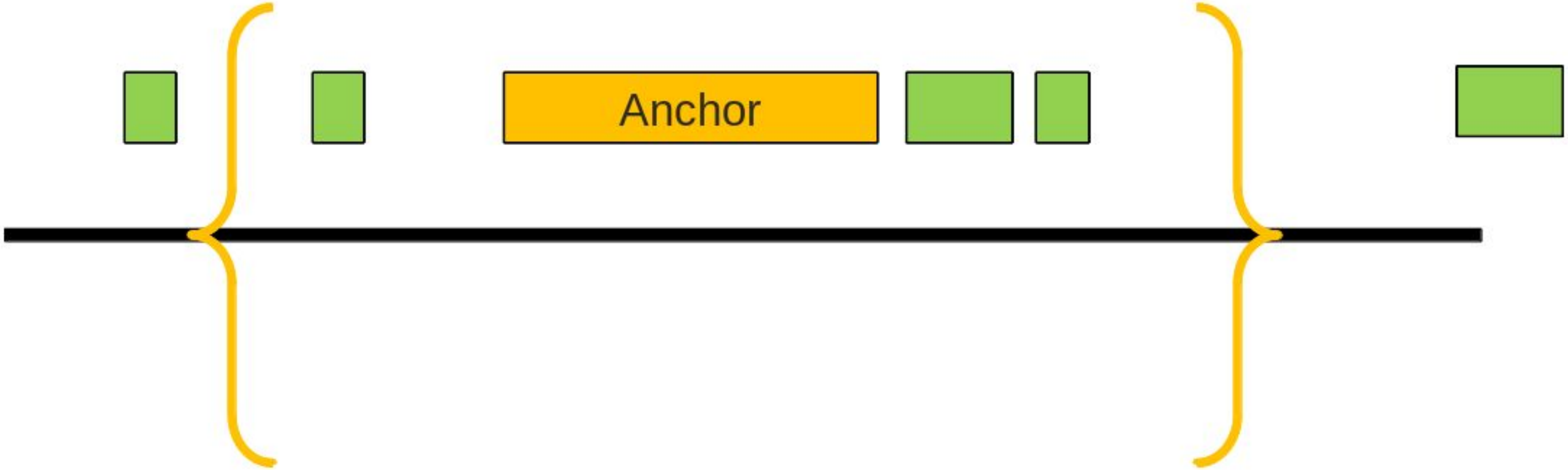


Alex Dobin, CSHL

Seed stitching strategy

- first, seeds are **clustered** together based on proximity to a selected set of "anchor" seeds (seeds that map <50 times)
- all seeds that map within user-defined genomic windows are **stitched** together
- “Alignment windows”: genome regions around anchors

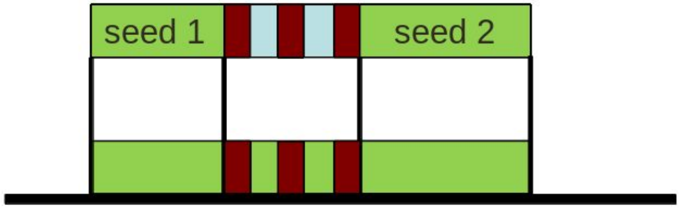
Size of the window ~ maximum intron size, ~1Mb for human

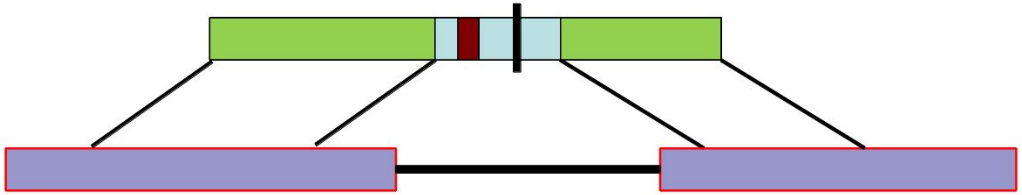


Alex Dobin, CSHL

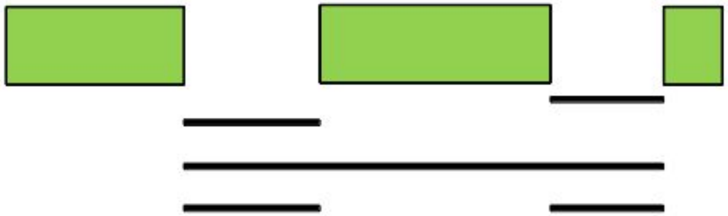
Seed stitching strategy

- dynamic programming algorithm stitches each pair of seeds, allowing for any number of mismatches but only one insertion or deletion (gap)





- local alignment **scoring** scheme
- N seeds: 2^N combinations - only works for shorter reads <200b
- longer reads: each seed is stitched to all the preceding seeds within a window
- **highest score stitched combination -> the best alignment of the read**



Alex Dobin, CSHL

Individual transcript expression

- There are two categories of transcript isoform quantifiers:
  - read-centric (Cufflinks, IsoEM, RSEM, Sailfish,eXpress, Kallisto): assign probability for each transcript fragment (paired-end read) to one transcript by maximizing the joint likelihood of read alignments based on the distribution of transcript fragment
  - exon-centric (Poisson model, linear regression approaches like rQuant, IsoLasso, SLIDE, flux capacitor): considers the read abundance on an exonic segment as the cumulative abundance of all transcript isoforms. The transcript is represented as a combination of exons and aims at estimating individual transcript abundance from the observed read counts at each exon
- The RPKM of a gene can then be obtained by summing the RPKM of its constituent transcripts (assuming that reads were assigned to transcripts in a mutually exclusive way)

Individual transcript expression

- An increasing number of programs (mostly read-centric such as RSEM, Sailfish, Kallisto) only use a mapping of the reads to the transcriptome (reference annotation) as input

- Although this can work well for well-annotated species, this will fail for species for which the annotation is not so good, since it will likely wrongly overestimate the quantifications

The RSEM software

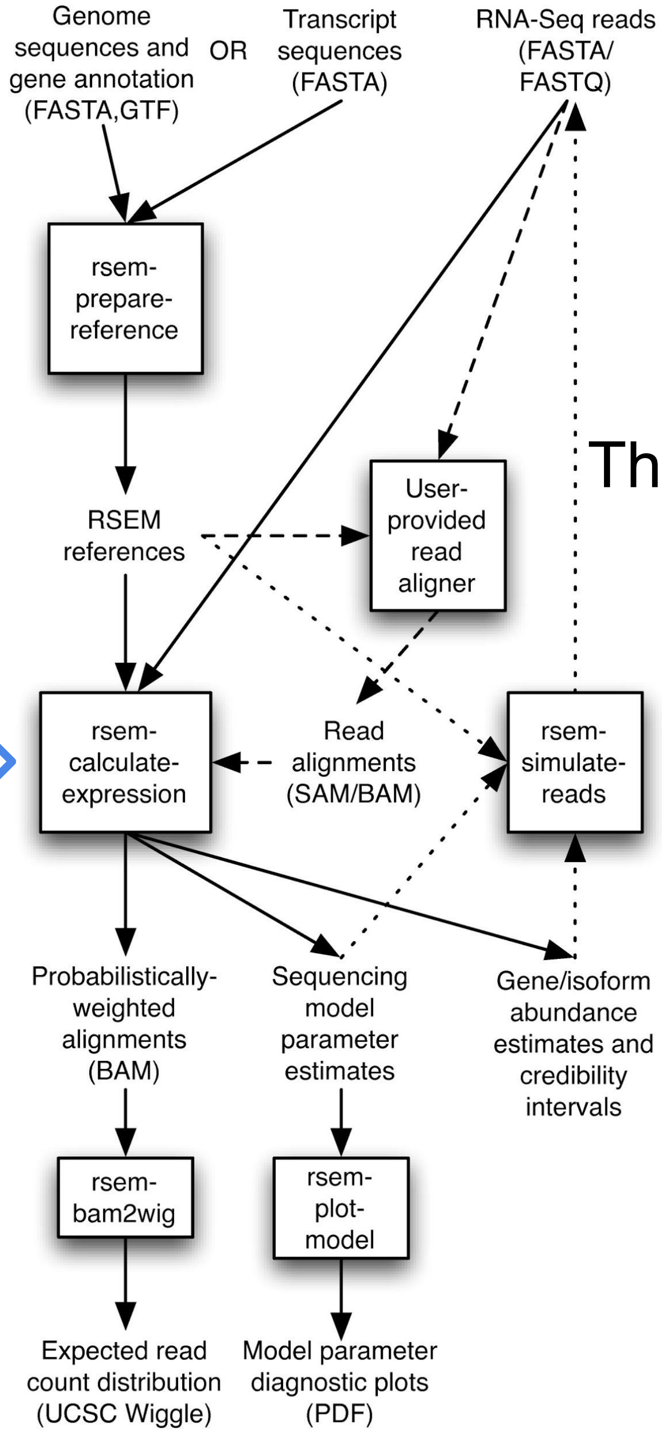
- RSEM: RNA-Seq by Expectation Maximization
  - Parameters = transcript abundances
  - Hidden variable = alignment

- Transcript-level alignment
- No need of a reference genome, requires a set of reference transcripts (eg. de novo transcriptome assembler, EST database...)

- Computes ML abundance estimates using the EM algorithm for its statistical method
- Good handling of multimaps leading to accurate quantifications

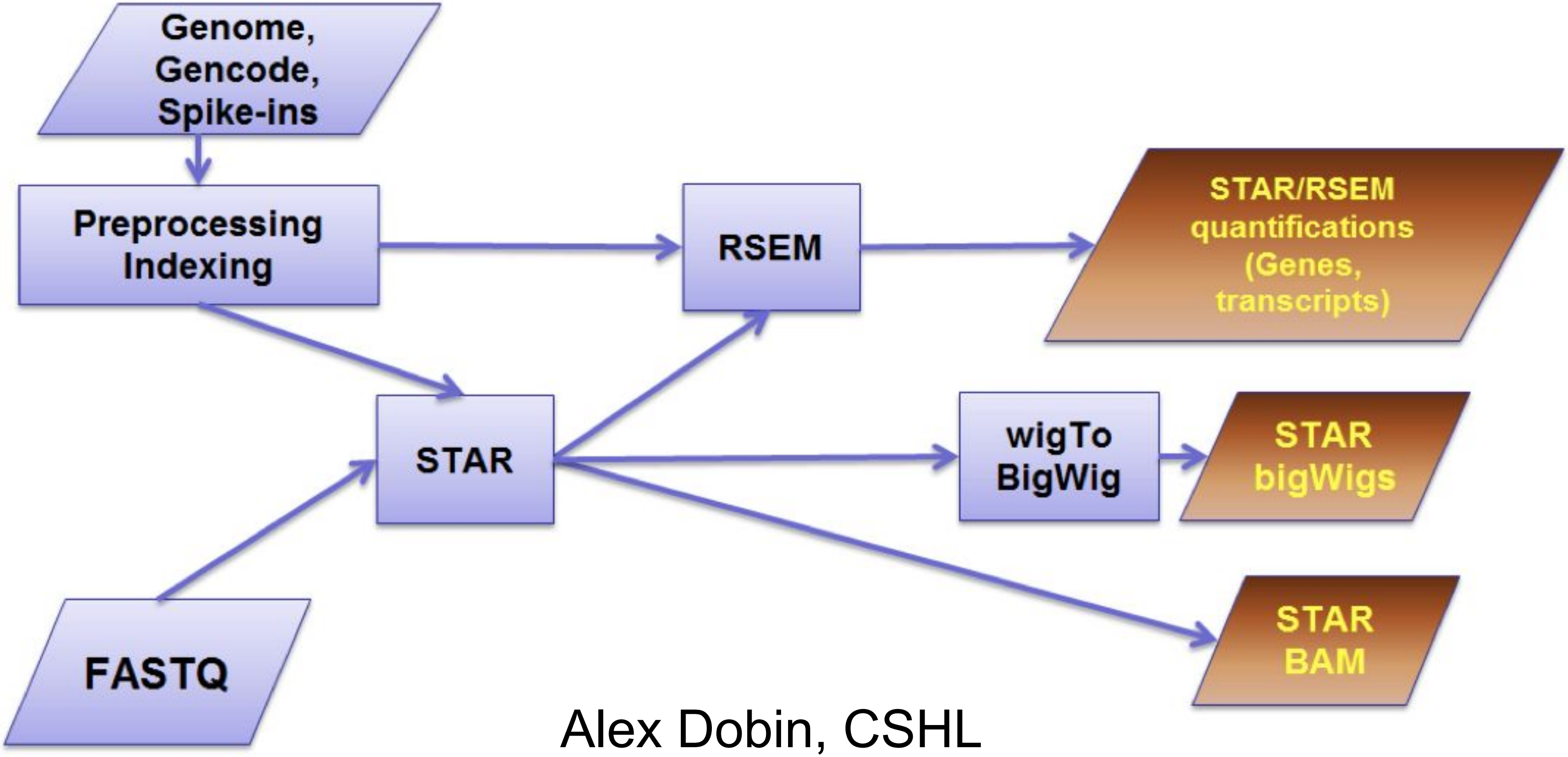
The RSEM software workflow

- Alignments of reads against reference transcript sequences (Bowtie)
- Calculate the relative abundances



Genome sequences and gene annotation (FASTA,GTF)
OR
Transcript sequences (FASTA)
RNA-Seq reads (FASTA/ FASTQ)
rsem-prepare-reference
RSEM references
User-provided read aligner
rsem-calculate-expression
Read alignments (SAM/BAM)
rsem-simulate-reads
Probabilistically-weighted alignments (BAM)
Sequencing model parameter estimates
Gene/isoform abundance estimates and credibility intervals
rsem-bam2wig
rsem-plot-model
Expected read count distribution (UCSC Wiggle)
Model parameter diagnostic plots (PDF)

STAR-RSEM pipeline



Alex Dobin, CSHL

Reads are mapped to the genome with STAR which then internally converts genome mappings to transcriptome mappings (from genome to transcriptome coordinates). RSEM takes a transcriptome mapping as input.