Studying the transcriptome using RNA-seq

Cecilia Coimbra Klein

IBUB
Institut de Biomedicina
de la Universitat de Barcelona

UNIVERSITAT DE BARCELONA

CRG
Centre
for Genomic
Regulation

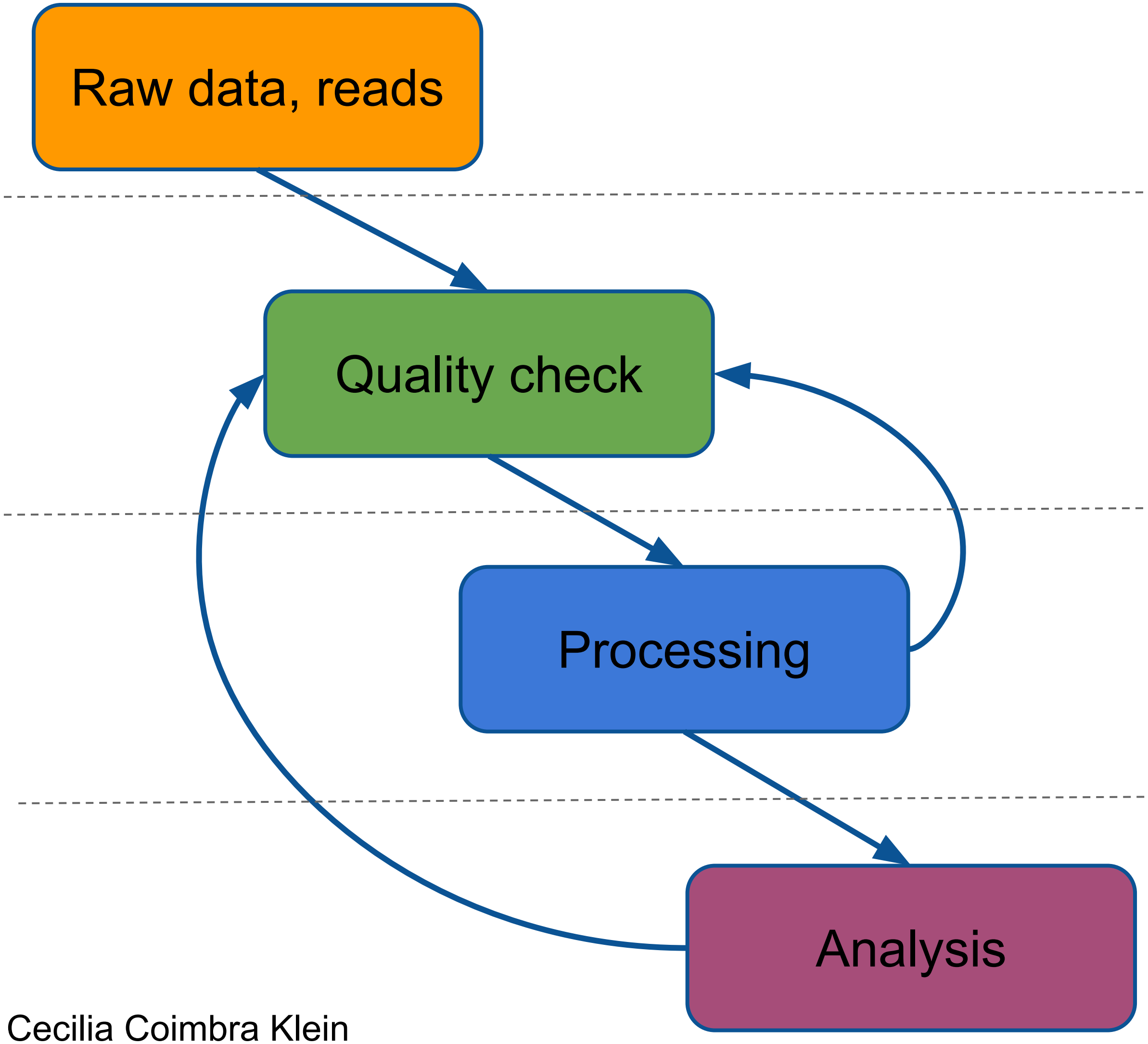
UVIC UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL DE CATALUNYA

Master in Omics
Data Analysis

Outline

1. Introduction
2. Basic concepts
3. Short-read RNA-seq data processing
    - 3.1. Quality control
    - 3.2. Read mapping
    - 3.3. Visualization of gene expression signal
    - 3.4. Gene expression quantification and normalization
4. Short-read RNA-seq data processing
5. Gene level RNA-seq data analysis
6. Isoform level RNA-seq analyses
7. Regulation of gene expression

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

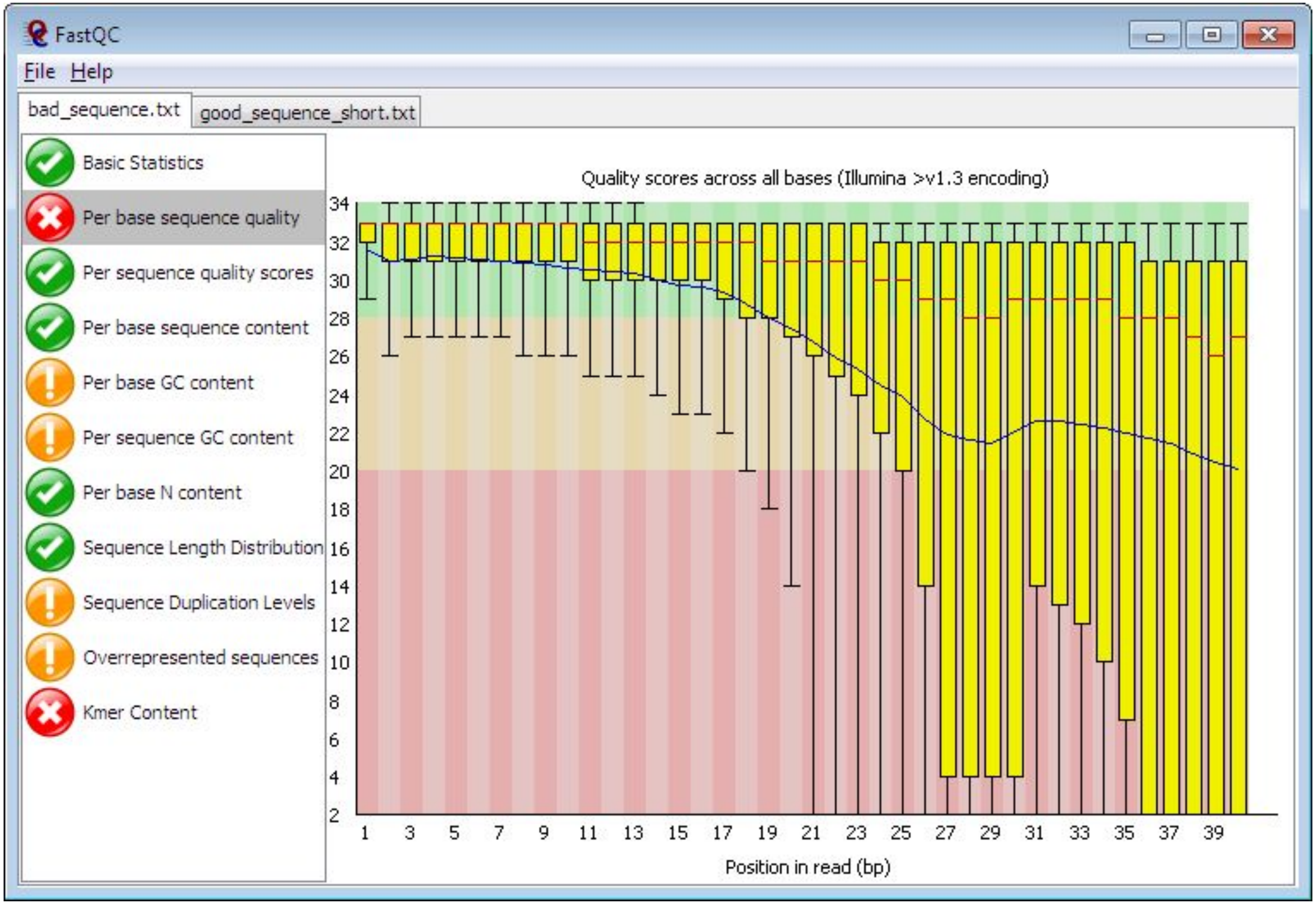
Quality metrics

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/

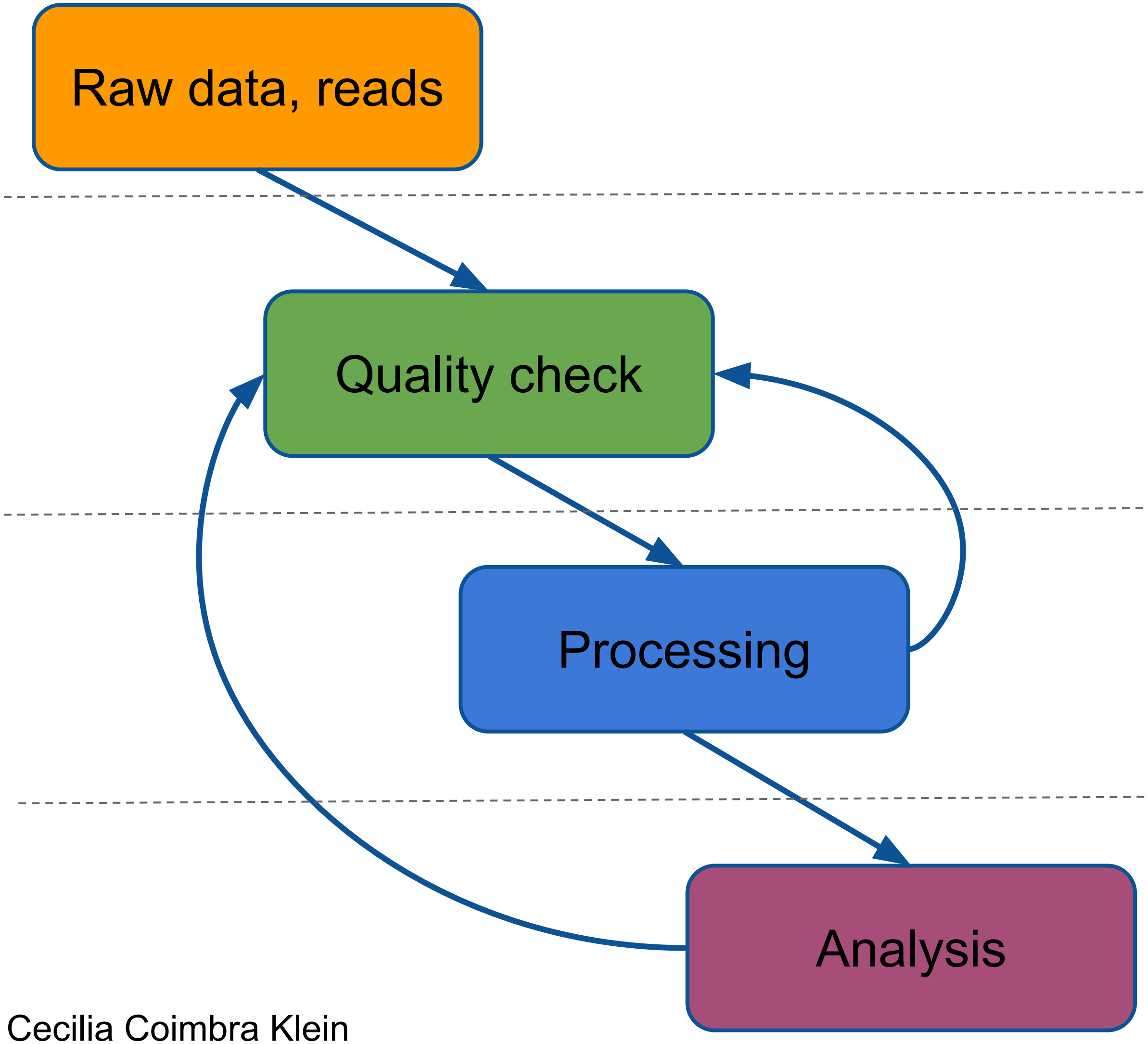
Cecilia Coimbra Klein

Hands-on

Fastq files and read QC 3.1

https://public-docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/#_fastq_files_and_read_qc

Post-sequencing: usual pipeline



Raw data, reads

Quality check

Processing

Analysis

Some data formats

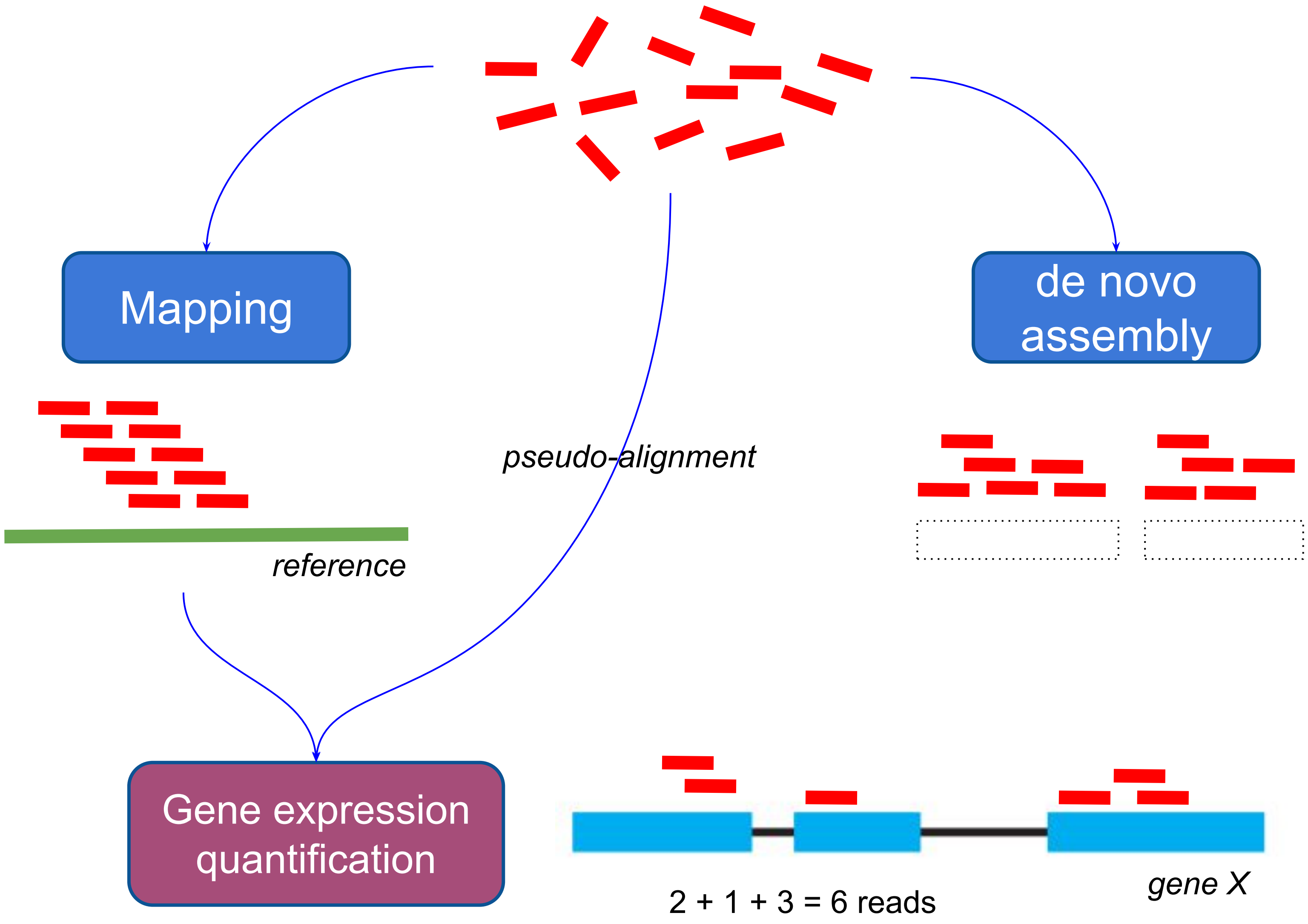
*.fastq, *.fa,
*.sff, *.sra

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

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

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

Processing



Mapping

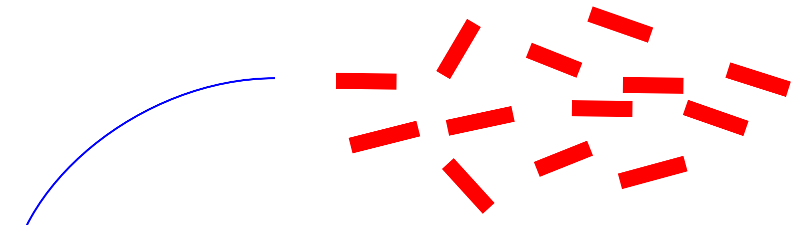
de novo assembly

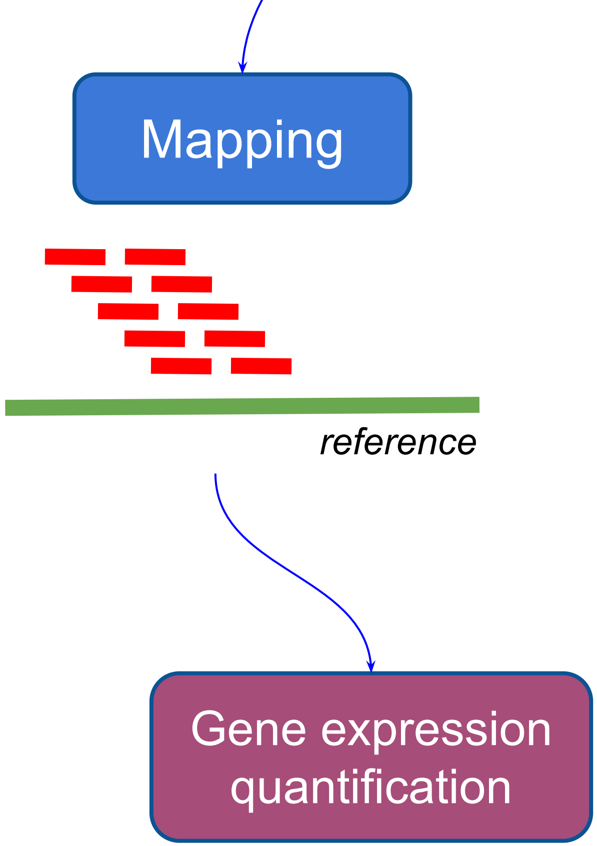
pseudo-alignment

Gene expression quantification

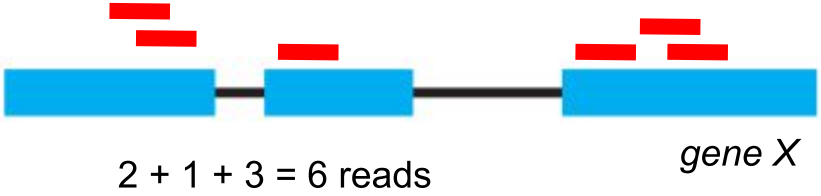
Mapping strategy

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

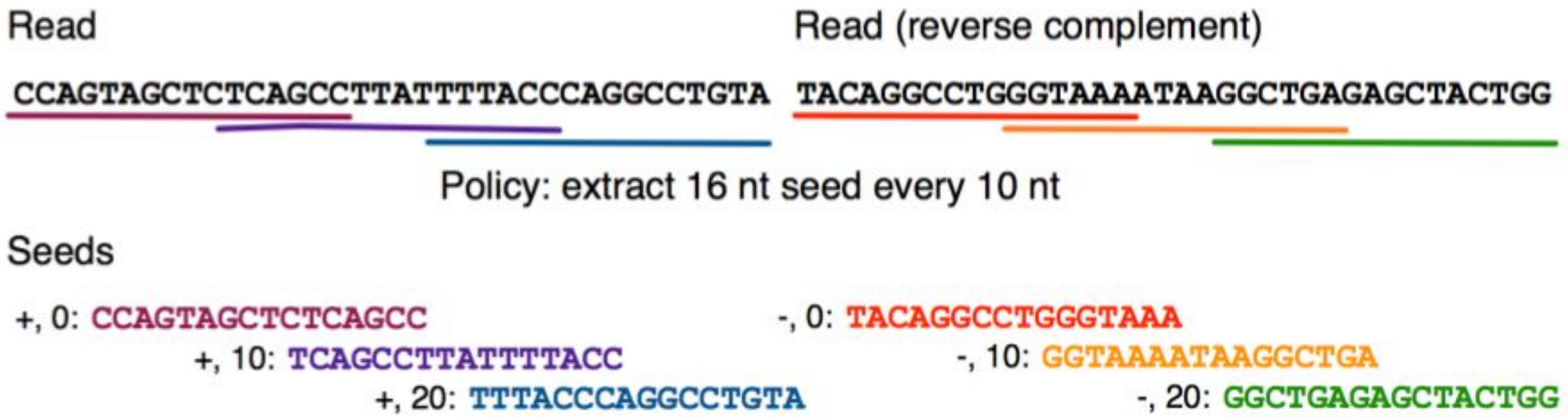
Indices

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

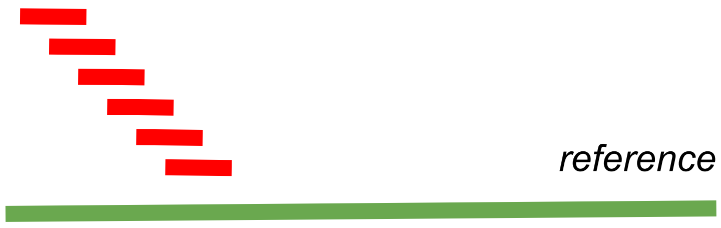
⟶ not exhaustive

Paired-end alignment

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

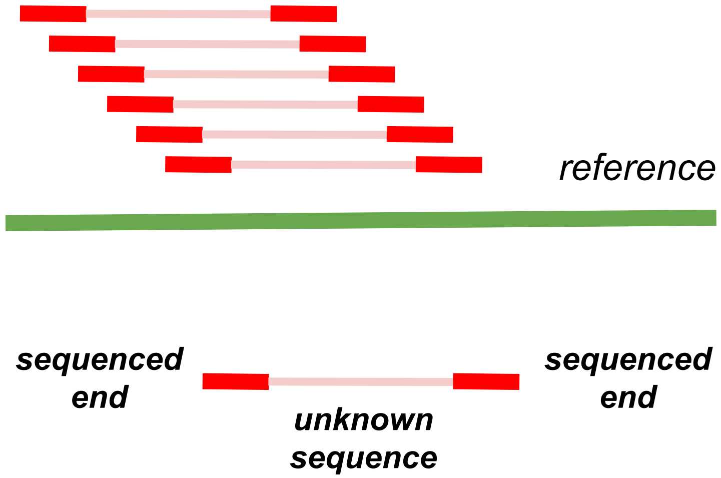
- connectivity information
- insert size and read length are known in advance (from library preparation)
- insert size distribution can be used to solve ambiguities (or even enhance the mapping process)

Single-end (SE) reads



reference

Paired-end (PE) reads



reference

sequenced end

unknown sequence

sequenced end

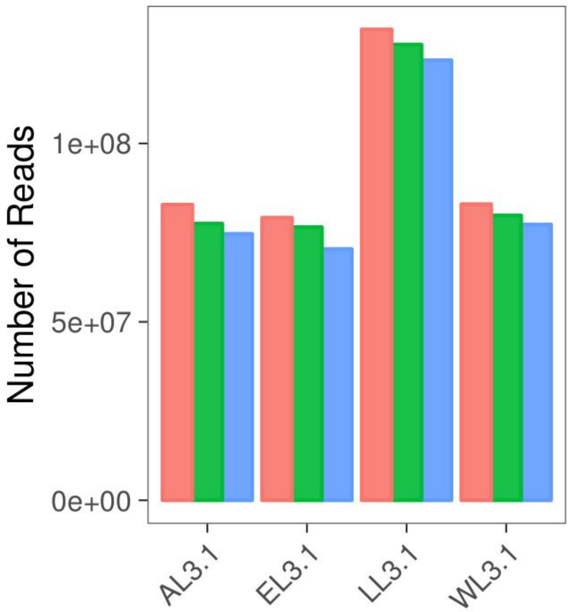
RNA-seq mapping

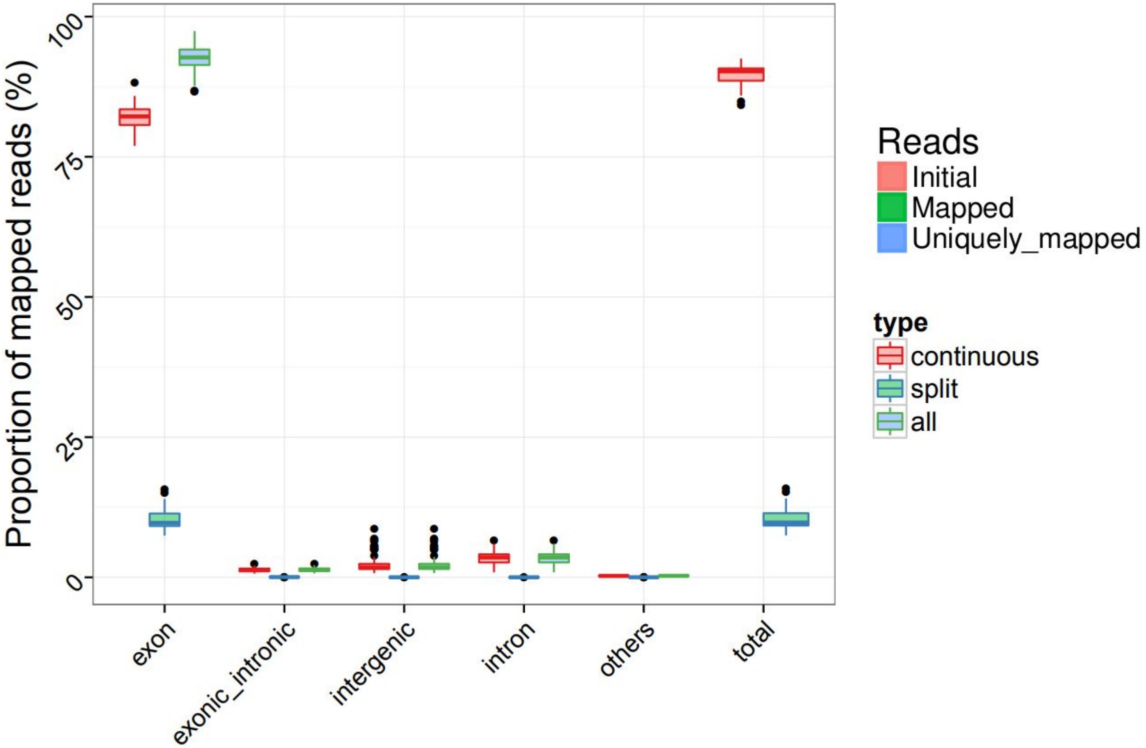
Specific variables to consider when mapping RNA-seq data

- 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

Cecilia Coimbra Klein

Mapping statistics





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

Hands-on

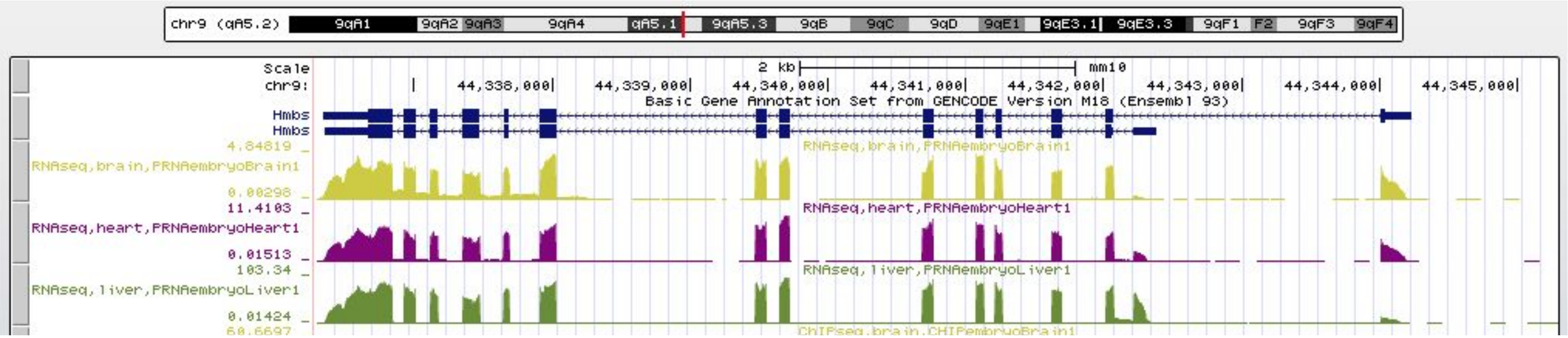
Mapping 3.2

https://public-docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/#_mapping

RNA-seq signal

RNA-seq signal

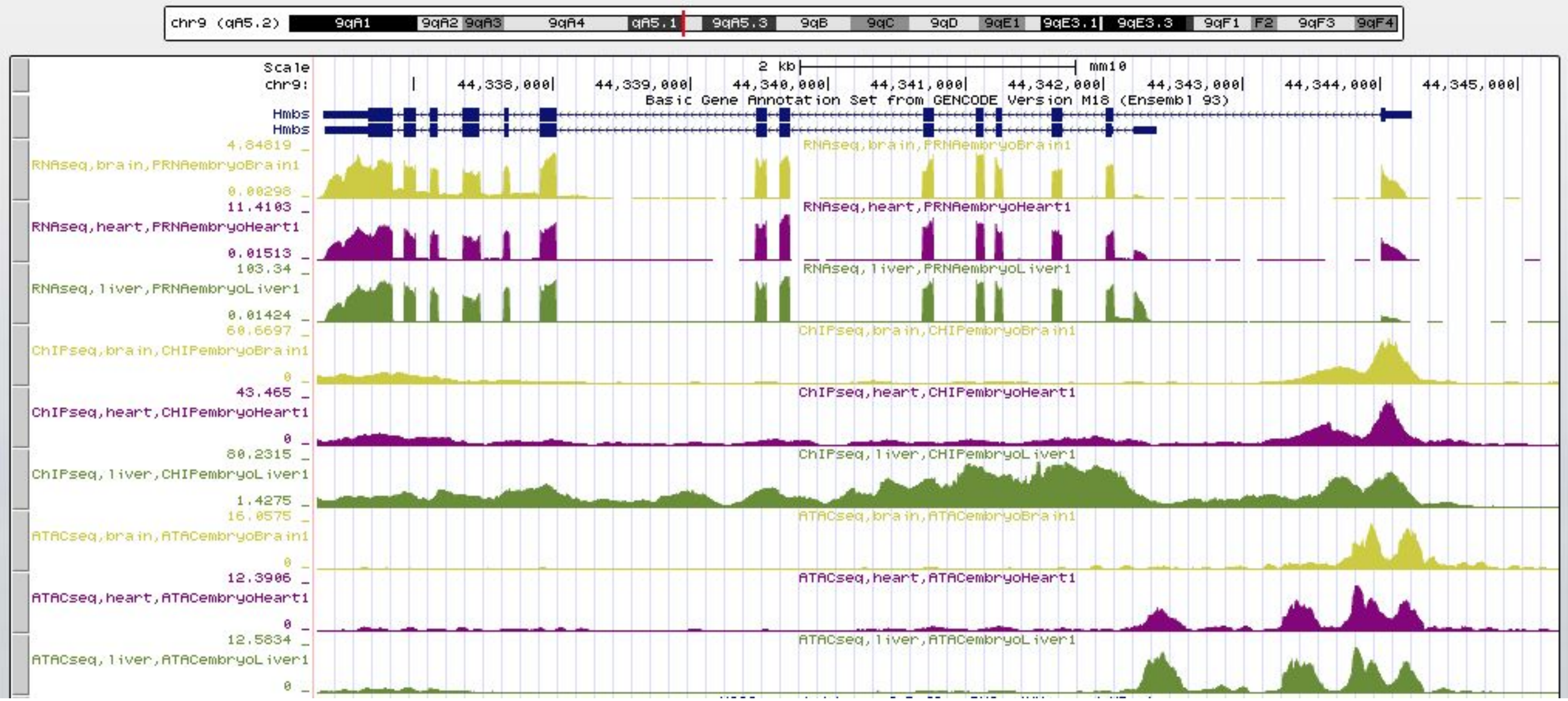
genome-euro.ucsc.edu



- 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
2 kb
mm10
chr9:
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

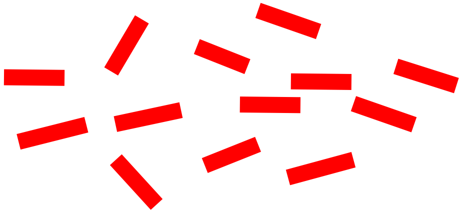
RNA-seq signal files 3.3

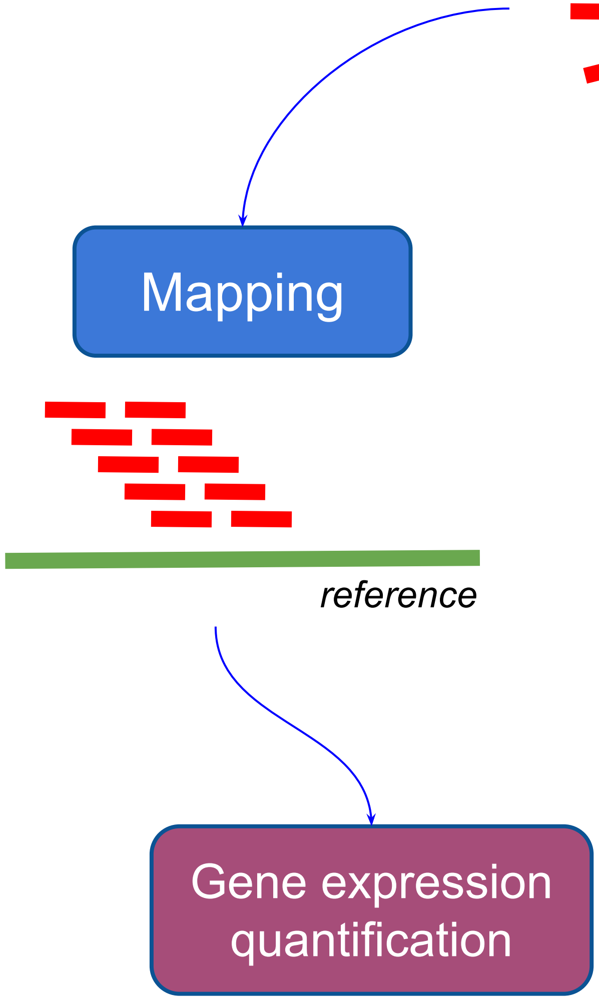
UCSC genome browser 3.4

https://public-docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/#_rna_seq_signal_files

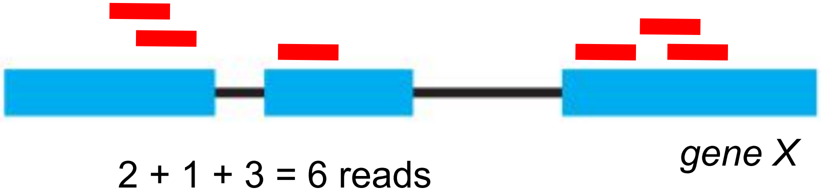
Gene expression quantification

Gene expression quantification



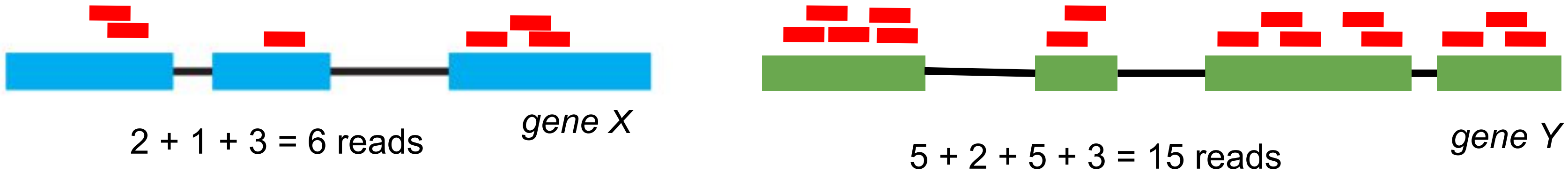


To quantify the expression of a gene, a simple idea is to count the RNA-seq reads that fall within the exons of this gene:

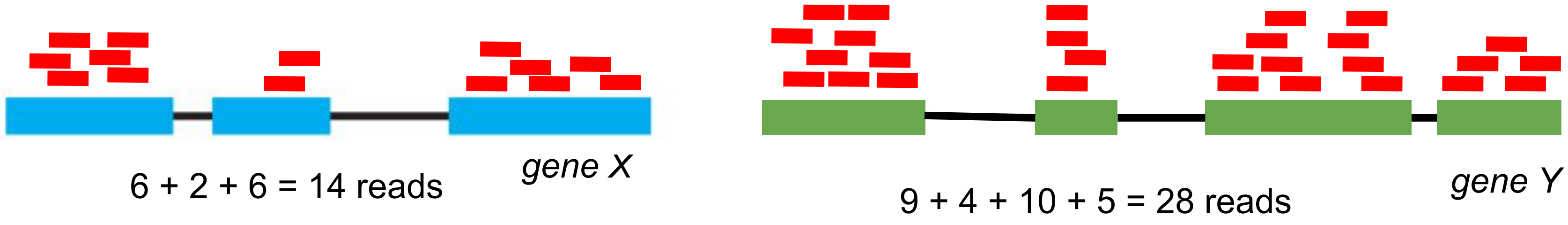


Gene expression quantification

- 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)

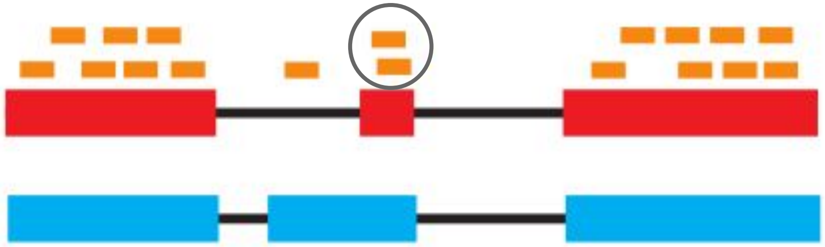
- For example, Mortazavi et al. (2008) estimates that 3 RPKM corresponds to ~ 1 transcript per cell in mouse liver, while Klish et al. (2011) say that 1 RPKM corresponds to between 0.3 and 1 transcript per cell...

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

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)

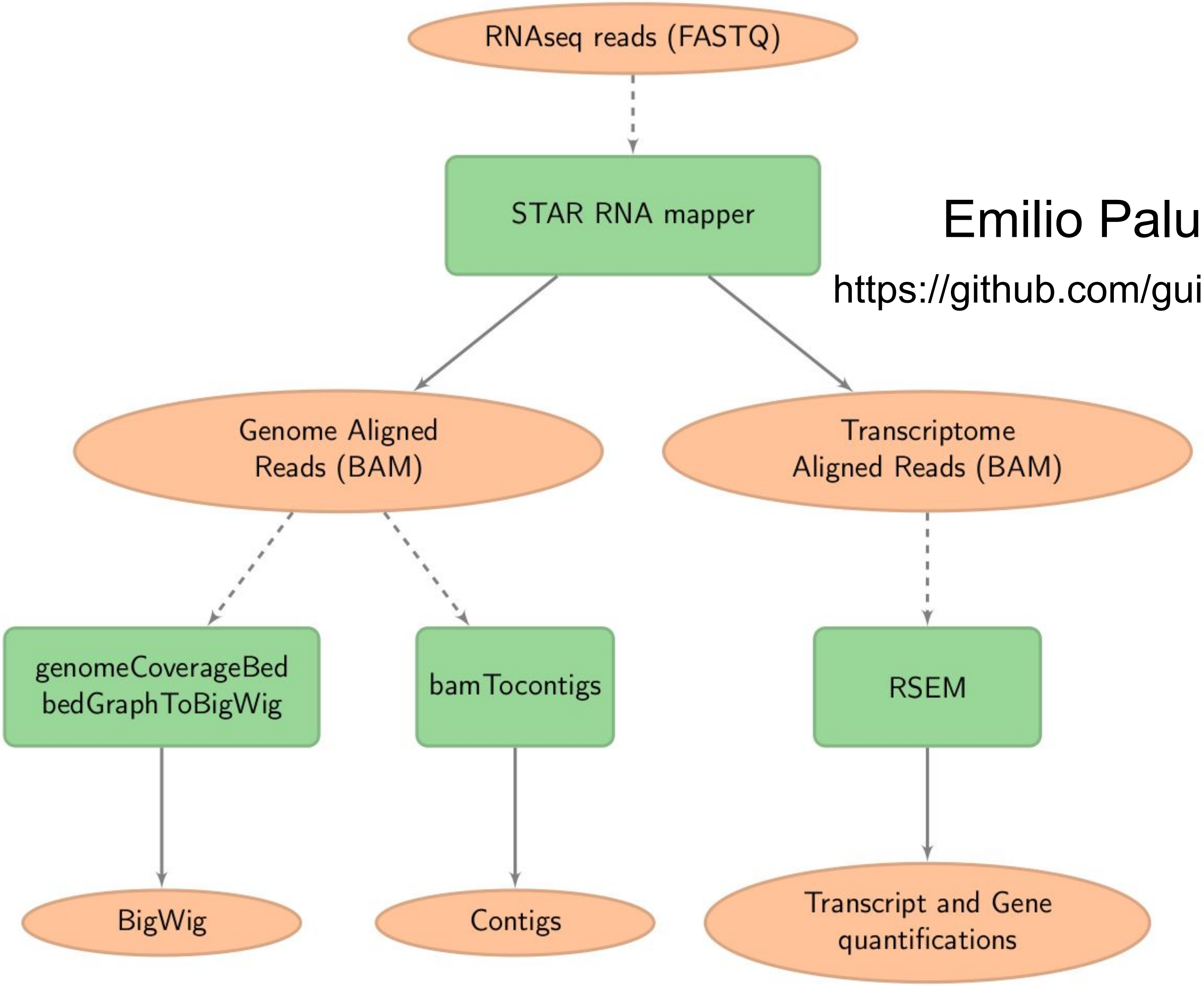
Hands-on

Transcript and gene expression quantification 3.5

https://public-docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/#_transcript_and_gene_expression_quantification

Pipeline

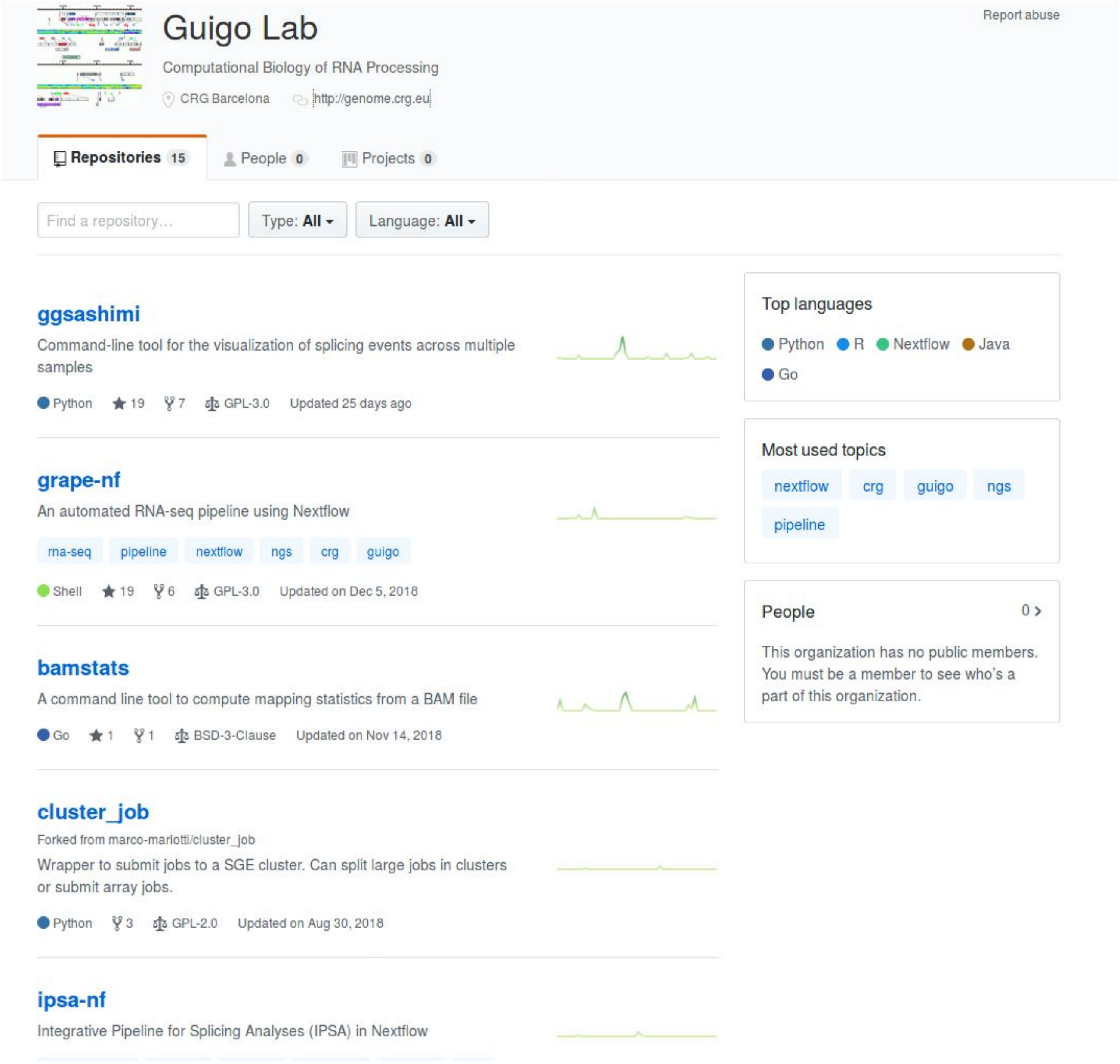
Grape pipeline


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



Report abuse

Guigo Lab

Computational Biology of RNA Processing

CRG Barcelona http://genome.crg.eu

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.

Applications

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.