Primary analysis for RNAseq data: Quality control, mapping and quantification

Daniel Crespo

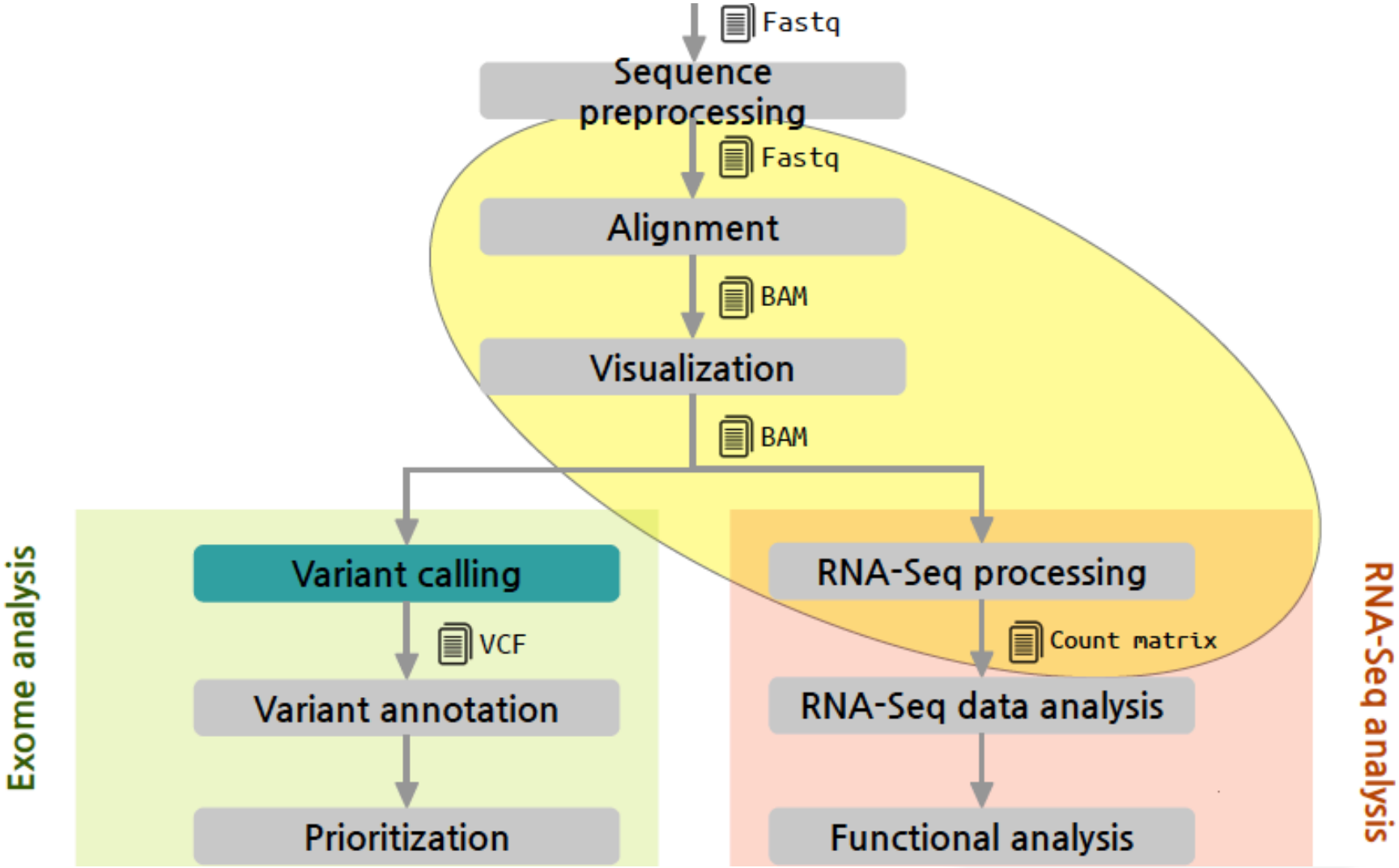
March 8, 2017

GDA
International Course on
Genomic Data Analysis

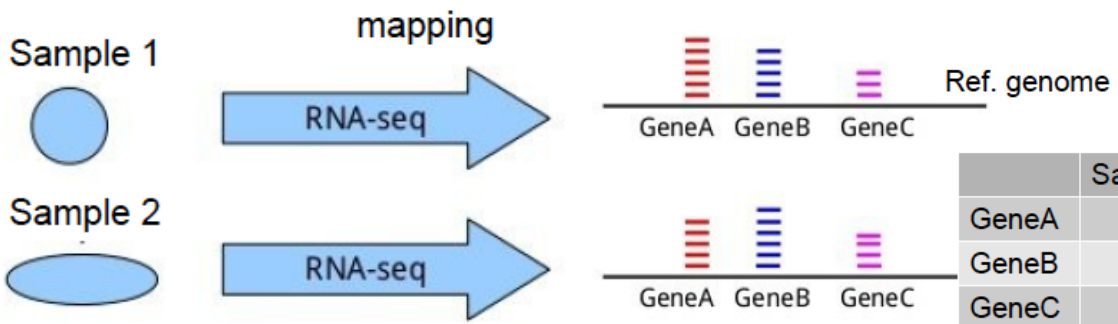
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Index

I Goal:
Where are we?



I Quantify expression of genomic features



	Sample 1	Sample 2
GeneA	6	5
GeneB	5	6
GeneC	3	4

NGS RNA-seq data
(FASTQ reads files)

Matrix of counts

Babelomics analysis suite

- Differential expression analysis
- Clustering analysis
- Predictors
- Functional enrichment analysis
- …

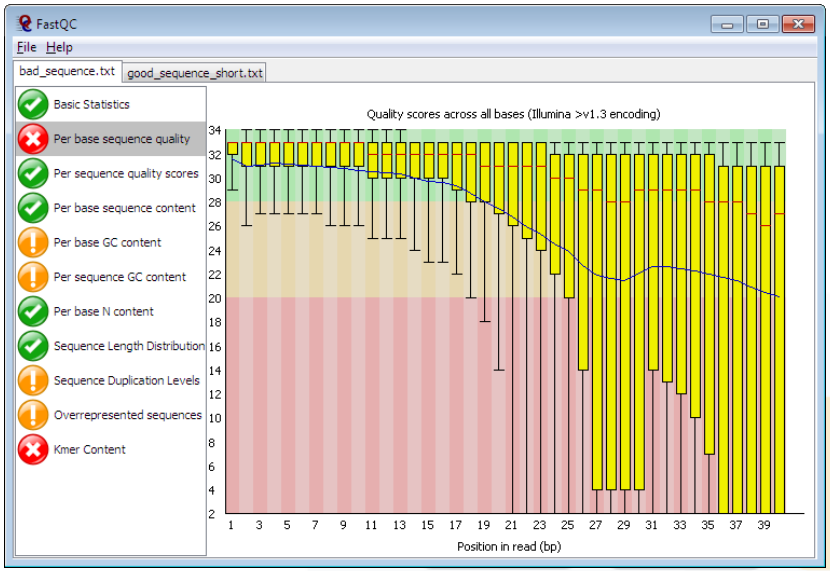
II Sequence quality evaluation:
FASTQ file format

- FASTQ file is “a fasta file with qualities”. Each read consists of four lines:
  1. Header (like a fasta header but starting with “@”
  2. Sequence of nucleotides
  3. “+” (and optionally the sequence ID)
  4. Encoded quality of the sequence

```
@ENST00000286800_45243_44929_1_0_0_0_2:0:0_1:0:0_0/1
GTTCTGCCAATGTTTCTGTGAATGTGAATATATGCATTTTTTTTAATTGCTGAGAAGATCCATAGCTTTCATTTGACGCTCAAACACATCTAACCACAGC
+
2222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222
@ENST00000286800_35367_34985_1_0_0_0_3:0:0_0:0:0_1/1
CAAGAGGTCAGGAGATCAAGACCGTCCTGGCTAACACAGTGAAACCCCGTCTCTACTAAAAATATAAAAAAATTACCCCGGCATGGTGGCCAGCGCCTTT
+
2222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222
@ENST00000286800_38454_38101_1_0_0_0_3:1:0_5:0:0_2/1
TCAGGATTTCTAATGTACTTTTTACTTTCATTTACTGAGGCCTGATATATATGAATTCATGGTATAATGAGCTTTATTTTATACCCAGATGCATATATTT
+
2222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222
```

II Quality control tools

- Quality control tools
  - FastQC
  - Fastx-toolkit
  - NGS QC Toolkit
  - HPG FastQ
  - …



- Sequence filtering tools
  - Cutadapt
  - Trimmomatic
  - Fastx-toolkit
  - SeqTK
  - HPG FastQ
  - …

Remove bad quality data to improve our confidence on downstream analysis

II Inputs and outputs

Reference genome (FASTA)
Sequence reads (FastQ)
Genomic features (GTF or similar)

List of counts per feature
(gene, exon, transcript…)

ENSG00000141956	68
ENSG00000141959	49
ENSG00000142149	37
ENSG00000142156	33
ENSG00000142166	30
ENSG00000142168	9
ENSG00000142173	38
ENSG00000142178	22
ENSG00000142182	11
ENSG00000142185	26

II Getting a reference genome

- A **reference genome** is a consensus sequence built up from high quality sequencing samples from different populations. It is the control *reference sequence* to compare our samples.
- **Genome Reference Consortium (GRC)** created to deliver assemblies:
  - https://www.ncbi.nlm.nih.gov/grc/human/data
- Current human assembly is **GRCh38.p10**
- Reference genomes can be downloaded from:
  - **GRC:** https://www.ncbi.nlm.nih.gov/grc
  - **Ensembl:** http://www.ensembl.org/index.html
  - **Ensembl Genomes:** http://ensemblgenomes.org
  - **UCSC:** http://hgdownload.soe.ucsc.edu/downloads.html

II Genomic features

- List of features with sequence annotations adapted to our data.

- Usually GFF/GTF format file

Sample GTF output from Ensembl data dump:

```
1 transcribed_unprocessed_pseudogene  gene        11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_sour
1 processed_transcript                transcript  11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328
```

Sample GFF output from Ensembl export:

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

- Some databases provide useful annotation files:
  - Ensembl: http://www.ensembl.org/index.html
  - MiRBase: http://www.mirbase.org

III

RNA-seq mapping:
Desirable features of a RNA-seq aligner

- **Sensitive**, we are looking for genomic variants, reads with mismatches and INDELS must be properly aligned
- **Specificity**, no wrong alignments should be provided
- Being able to perform gapped alignments, exones must be correctly located
- Good performance, efficiency matters
- Easy to use
- Open-source and maintained

III RNA: Tophat, the standard aligner

- Tophat is the standard for RNA-seq mapping
  - http://tophat.cbcb.umd.edu/
- It uses Bowtie2 to align reads, so it´s not very sensitive, usually maps 75% of reads
- Not ready for long reads (>150bp), mapping decrease to below 50%
- Poor performance, can take several hours to map
- Big memory footprint and a lot of disk used
- Mapping fall down with mismatches, INDELS and longer reads
- Written in Python and C. Not designed to take advantage of new technologies and clusters

III RNA: STAR and MapSlice

- STAR developed for the ENCODE project
  - Http://code.google.com/p/rna-star/
  - High performance, not very high sensitivity
- MapSplice
  - http://www.netlab.uky.edu/p/bioinfo/MapSplice2
  - Not bad sensitivity but very slow

IV SAM/BAM specification:
Mapping output: SAM/BAM format

Text file that stores large nucleotide sequence alignments:
SAM specification: https://samtools.github.io/hts-specs/SAMv1.pdf

- SAM file under BGZF compression format
  - Binary file
  - Save disk space (~ 80% of compression)
  - Indexing for efficient random access
  - Easy to convert to one another using SAMtools
  - Accepted by most of the available software

Header

```
@HD    VN:1.0  SO:coordinate
@SQ    SN:chr1 LN:249250621
@PG    ID:TopHat    VN:2.0.8    CL:/opt/soft/ngs/tophat/tophat-2.0.8.Linux_x86_64/tophat -p 4 -o
    /clinicfs/projects/3.ENCODE/mappings/Gm12878/Gm12878_Rp1_pair --no-coverage-search -r 300 --mate-std-dev 200 --
    library-type fr-unstranded /clinicfs/common/reference-genomes/homo_sapiens/bt2/hg19_ucsc/hg19_ucsc
    /clinicfs/projects/3.ENCODE/reads/Gm12878_Rp1_1.fastq /clinicfs/projects/3.ENCODE/reads/Gm12878_Rp1_2.fastq
```

Alignments

```
61PKHAAXX_HWUSI-EAS627_0007.68122391   337   chr1   10536  1    76M   =    173766  163
    TACCACCGAAATCTGTGCAGAGGAGAACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGC
    E5@>>>?B?A:BCB@??BAEBBCEC?@EDE@FEEFFEC=:BFFFFFAEE=EEDFFFFFDFFDFFFFEGGGDFEFFF   AS:i:-5 XN:i:0  XM:i:1
    XO:i:0  XG:i:0  NM:i:1  MD:Z:24C51    YT:Z:UU NH:i:3  CC:Z:=  CP:i:10536    HI:i:0
61PKHAAXX_HWUSI-EAS627_0007.68122391   113   chr1   10536  1    76M   chr16  90195094    0
    TACCACCGAAATCTGTGCAGAGGAGAACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGC
    E5@>>>?B?A:BCB@??BAEBBCEC?@EDE@FEEFFEC=:BFFFFFAEE=EEDFFFFFDFFDFFFFEGGGDFEFFF   AS:i:-5 XN:i:0  XM:i:1
    XO:i:0  XG:i:0  NM:i:1  MD:Z:24C51    YT:Z:UU NH:i:3  CC:Z:=  CP:i:10536    HI:i:1
```

IV Mapping output, mandatory fields

First columns are mandatory

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

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

Flag

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse com
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

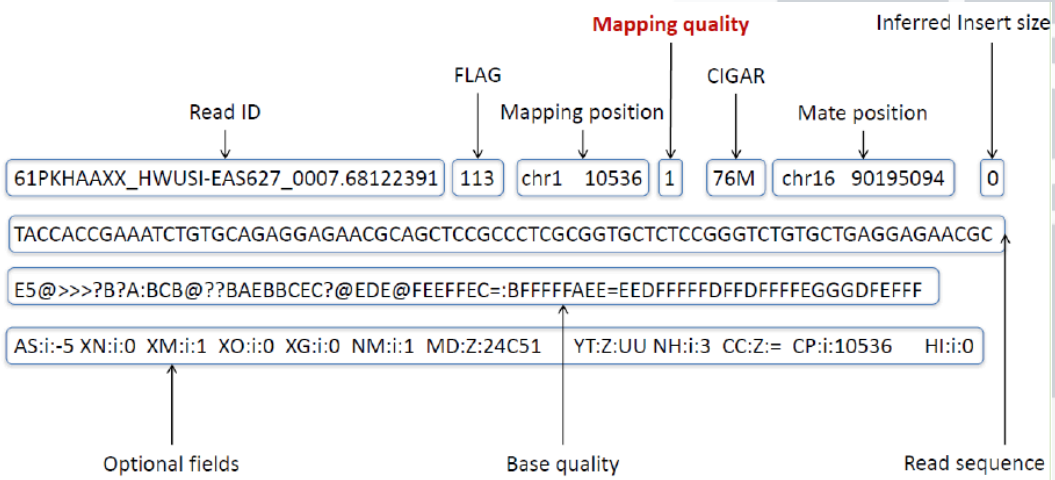
Cigar

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

IV Mapping output, optional fields

- Some optionals fields in the aligner section
- SAM specification is part of SAMtools package. More info at: http://samtools.sourceforge.net/
- A binary SAMtools is freely distributed to:
  - SAM ↔ BAM
  - Depth
  - Merge
  - Sort
  - …

Type	Regexp matching VALUE	Description
A	`[!-~]`	Printable character
i	`[-+]?[0-9]+`	Signed integer[6]
f	`[-+]?[0-9]*\.?[0-9]+([eE][-+]?[0-9]+)?`	Single-precision floating number
Z	`[ !-~]*`	Printable string, including space
H	`([0-9A-F][0-9A-F])*`	Byte array in the Hex format[7]
B	`[cCsSiIf](,[-+]?[0-9]*\.?[0-9]+([eE][-+]?[0-9]+)?)+`	Integer or numeric array



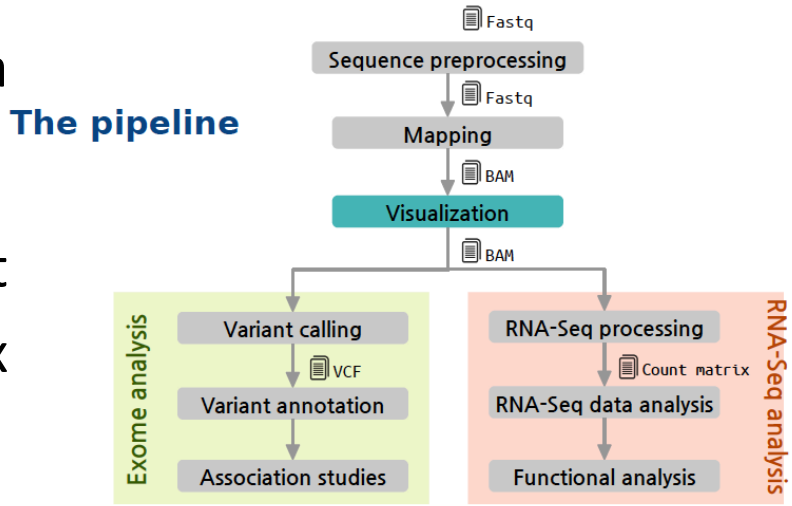
V Alignment quality evaluation

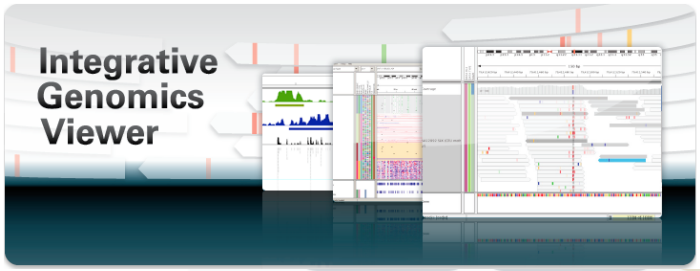
- We need to know how well the alignment process went
- Hundred of million of mapped reads
- Some biases can occur
- Some useful information
  - % reads mapped
  - Mean average error
  - Error distribution
  - Length distribution
  - Coverage…

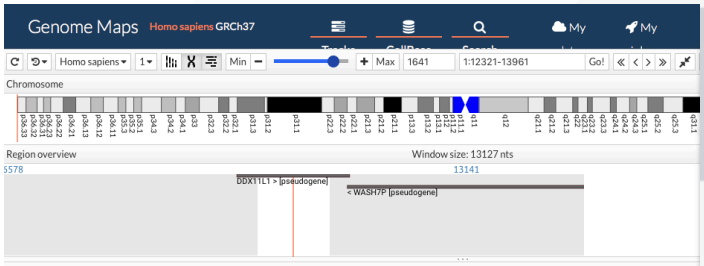
- QC and filtering tools:
  - Samtools
  - QualiMap
  - BamQC
  - SAMstat
  - NOISeq
  - BamTools
  - PicardTools…

VI Visualization:
Why visualization?

- Large quantities of genomic data (NGS, array based methods…)
- Human interpretation and judgment using visualization can gelp complex biological relationships
- Two Genomic Viewers:
  - Integrative Genomic Viewer (IGV)
  - Genome Maps (http://genomemaps.org/)







VI Integrative Genomic Viewer (IGV)

- **Integrate** different data types simultaneously
- View **large datasets** easily
- Faster navigation ow browsing
- Runs **locally** on your desktop
- Used by large-scale projects
- Open source and **freely available**

Chromosome	Source	Feature	Start	End	Score	Strand
chr1	hg19_ensGene	Exon	100	250	0	+
chr1	hg19_ensGene	Start_codon	100	102	0	+
chr1	hg19_ensGene	CDS	100	250	0	+

Chromosome	Start	End	Feature_ID	Score	Strand
chr1	941	942	Peak_1	12,67	+
chr1	2276	2277	Peak_2	14,55	+
chr1	2718	2719	Peak_3	36,44	+

Mandatory fields ← → Optional fields

- Any data related to **genome coordinates**
- **Sample annotations or attributes**
- **Genome** annotations

SOURCE DATA	RECOMMENDED FILE FORMAT
Sequence alignment data	SAM / **BAM** (must be indexed)
Genome annotations	GFF / GFF3 or BED format
Variant data	VCF
Any numeric data	IGV format, TAB or WIG format
Gene expression data	GCT or RES format

Index BAM file: “ samtools index data1.bam “

VI IGV interface

VI Genome Maps

Genome Maps Homo sapiens GRCh37 Tracks CellBase Search My data My jobs Settings Login Sign up

Homo sapiens 1 Min Max 3185 1:11549-14733 Go! gene

Chromosome

Region overview Window size: 25479 nts

402 13141

DDX11L1 > [pseudogene]

< WASH7P [pseudogene]

Detailed information Window size: 3185 nts

11549 13141

Sequence

Gene

DDX11L1 > [pseudogene]

DDX11L1-002 [processed_transcript]

DDX11L1-201 [transcribed_unprocessed_pseudogene]

DDX11L1-202 [transcribed_unprocessed_pseudogene]

DDX11L1-001 [transcribed_unprocessed_pseudogene]

SNP

Active Tracks

- Sequence
- Gene
- SNP

Add CellBase Tracks

- Core
  - Sequence
  - Gene/Transcript
- Variation
  - SNP
  - EMBL-EBI EVA SNVs
  - Clinical Variants
- Regulatory
  - TFBS
  - miRNA Targets
  - Others
- Conservation
  - PhastCons
  - PhyloP

Advanced Cellbase Search

Search by any external reference :

Result type :

Genes

Search

Created by Computational Genomics Department @ Principe Felipe Research Center, 2015 1:13152 SNP legend Gene legend v3.2.0

VII Extracting RNA-seq counts:
Inputs and outputs & Quantification programs

Alignments (BAM/SAM files)
Genomic features (GTF or similar)

↓

Quantification and normalization

List of counts per feature
(gene, exon, transcript…)

- Many programs:
  - HTSeq
  - Cufflinks
  - Qualimap
  - ERANGE
  - NEUMA
  - MISO
  - RSEM
  - NOISeq
  - CummeRbund
  - …

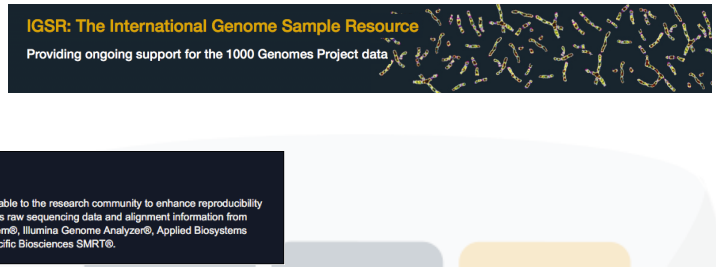
VII Count normalization

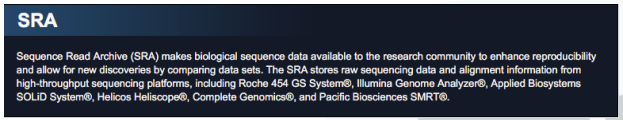
- **Gene / transcript length** influences the count
- **Library size** influences the count: more reads in the library -> more reads per gene/transcript
- Many other biases:
  - Differences on the read count distribution among samples
  - GC content of the gene affects the detection of that gene (Illumina)
  - Sequence-specific bias is introduced during the library preparation

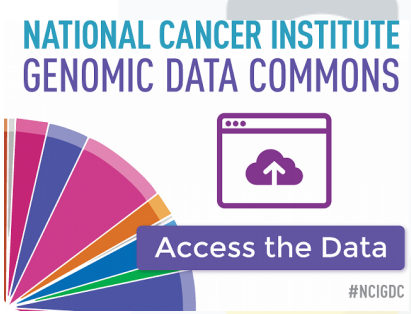
- **RPKM:** Reads per kilobase of the transcript per million mapped reads. *Mortazavi et al. 2008*
  - $RPKM = 10^9 \frac{C\ (No.\ reads\ mapped\ to\ the\ exons)}{N\ (Total\ mappable\ reads)\ L\ (Length\ of\ exons)}$
- Many other count corrections:
  - **FPKM** (Trapnell et al., 2010)
  - **TC:** Gene counts are divided by the sequencing depth associated to that sample and multiplied by the average of the total counts across all the samples. Gene counts are divided by the gene length (kb) times the total number of millions of mapped reads.
  - **TMM** (Robinson & Oshlack, 2010)
  - **Upper-quartile** (Bullard et al., 2010)
  - **Median** (Bullard et al., 2010)
  - **Quantile** (Irizarry et al., 2003)
  - **edgeR** (Robinson et al., 2009)
  - …

VIII Data repositories

- GEO, Gene Expression Omnibus
  - https://www.ncbi.nlm.nih.gov/geo/
- IGSR: The international Genome Sample Resource
  - http://www.internationalgenome.org
- SRA, Short Read Archive
  - http://www.ncbi.nlm.nih.gov/sra
- EGA, European Genome Phenome Archive
  - https://www.ebi.ac.uk.ega
- GDC, Genomic Data Commons Data Portal
  - https://gdc-portal.nci.nih.gov
- … and many others







Hands on!

Thank you very much for your attention!

Systems genomics team

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

GDA