From Fastq Files to Differentially Expressed Genes

Introduction to RNA-Seq Pipelines and Computational Analysis Mini Workshop

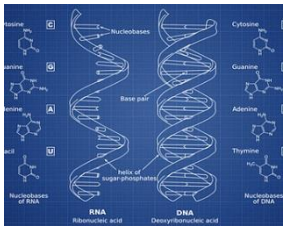
12 January 2018

Álvaro Cortés C.

GENOMICS CORELEUVEN

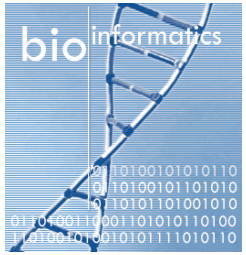
RNA-Seq Project Workflow

Experimental Design



Library Preparation

Sequencing



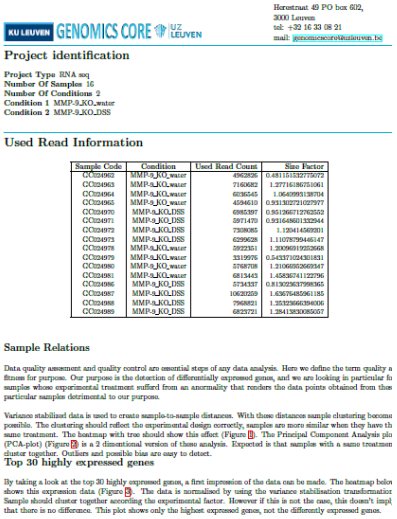
Bioinformatics

Follow-up & Support

GENOMICS CORELEUVEN

RNA-Seq at Genomics Core

A typical RNA-Seq analysis at the Genomics Core delivers:



KU LEUVEN GENOMICS CORE UZ LEUVEN

Herestraat 49 PO box 602,
3000 Leuven
tel: +32 16 33 08 21
mail: genomicscore@kuleuven.be

Project identification

Project Type RNA seq
Number Of Samples 16
Number Of Conditions 2
Condition 1 MMP-9_KO_water
Condition 2 MMP-9-KO_DSS

Used Read Information

Sample Code	Condition	Used Read Count	Size Factor
GC124962	MMP-9_KO_water	4962826	0.481151327775072
GC124963	MMP-9_KO_water	7140682	1.27716186751061
GC124964	MMP-9_KO_water	6036545	1.0640993138704
GC124965	MMP-9_KO_water	4594610	0.931302710277977
GC124970	MMP-9_KO_DSS	6985397	0.951266717763552
GC124971	MMP-9_KO_DSS	5971470	0.931648601332944
GC124972	MMP-9_KO_DSS	7308085	1.12041456893931
GC124973	MMP-9_KO_DSS	6299628	1.11078799446147
GC124978	MMP-9_KO_water	5922351	1.20096919253668
GC124979	MMP-9_KO_water	3319976	0.543371024301831
GC124980	MMP-9_KO_water	5768708	1.21066952669347
GC124981	MMP-9_KO_water	6813443	1.45836741122796
GC124986	MMP-9_KO_DSS	5734337	0.813923637998365
GC124987	MMP-9_KO_DSS	10632059	1.63676485961185
GC124988	MMP-9_KO_DSS	7968021	1.25323666394006
GC124989	MMP-9_KO_DSS	6823721	1.28413830085937

Sample Relations

Data quality assesment and quality control are essential steps of any data analysis. Here we define the term quality a fitness for purpose. Our purpose is the detection of differentially expressed genes, and we are looking in particular fo samples whose experimental treatment suffers from an anormality that renders the data points obtained from thes particular samples detrimental to our purpose.

Variance stabilized data is used to create sample-to-sample distances. With these distances sample clustering becomes possible. The clustering should reflect the experimental design correctly, samples are more similar when they have th same treatment. The heatmap with tree should show this effect (Figure 1). The Principal Component Analysis plo (PCA-plot) (Figure 2) is a 2 dimentional version of these analysis. Expected is that samples with a same treatmen cluster together. Outliers and possible bias are easy to detect.

Top 30 highly expressed genes

By taking a look at the top 30 highly expressed genes, a first impression of the data can be made. The heatmap belo shows this expression data (Figure 3). The data is normalised by using the variance stabilisation transformatio Sample should cluster together according the experimental factor. However if this is not the case, this doesn't impl that there is no difference. This plot shows only the highest expressed genes, not the differently expressed genes.

20 report files: differentially expressed genes, reads quality report, counting report,…

3GB of raw and analyzed data: Fastq, BAM, counts, normalized counts, gene expression files, heatmaps…

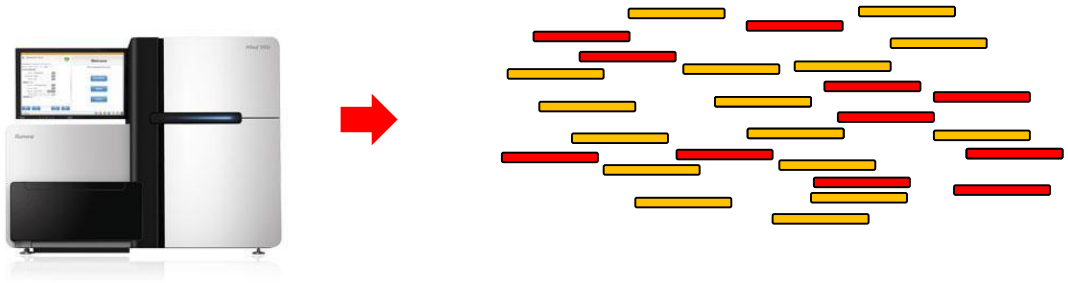
GENOMICS CORELEUVEN

RNA-Seq Data Interpretation & Analysis

GENOMICS CORELEUVEN

Reads and Differential Expression

- Sequencer produces millions of reads and a qualification for each base call



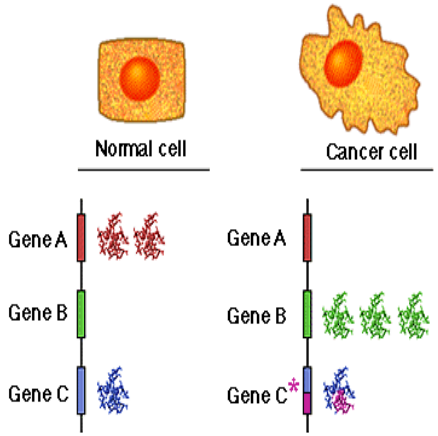
- What reads have to do with differential expression of genes?
  - Observations: read counts
  - Goal: discover changes in abundance between conditions

Read count is linearly related to the abundance of the target transcript

Estimating Differential Expression of Genes

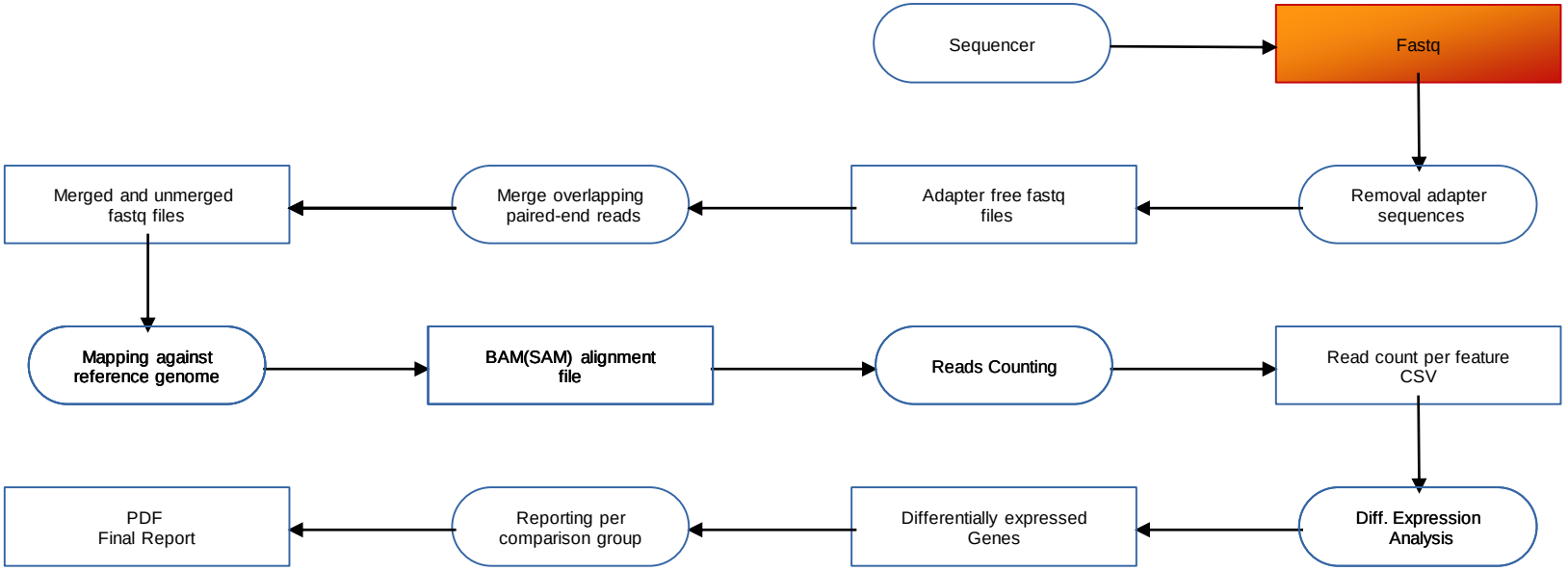
- Count the number of reads that map into annotated genes
  - Map: Align reads w.r.t. a reference genome (transcriptome)
  - Count: number of aligned reads per feature (genes).

- Perform statistical analysis on the counts to discover quantitative changes in expression levels between experimental groups
  - Normalization of counts
  - Probabilistic modeling of read counts
  - Estimate differential expression



GENOMICS CORELEUVEN

RNA-Seq Workflow at Genomics Core



GENOMICS CORELEUVEN

RNA-Seq Data Interpretation & Analysis

- What is a "read"?
  - A raw sequence (ordered collection) of nucleotides names A,C,G,T, or N.
  - Typical differential Expression Analysis: 50 characters long (single end).

- Fastq format?
  - Plain-text file, where each read and complementary information occupies 4 consecutive lines
  - Typical size RNA-Seq 50bp, 15M reads: 500M compressed, 2000M unzipped

```
@HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
GGGGGATTCTCACTAGGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1560:2177 1:N:0:ATCACG
ATGGTCCAGCAAGGGGTATGCTGAGAAGGGGAGCAGTTCAGAACCCATCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1583:2223 1:N:0:ATCACG
CTACCTTCACTATCAACATAGCAAACACACCTTTAGCTCCAGCTATTAACA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1609:2245 1:N:0:ATCACG
AGCTTAAGAGGCACTAGACACAGCCAGCTTCTTCAGGTGATCCATGAACAC
```

  - Counting reads in fastq files: zcat my.fastq.gz | echo $((`wc -l`/4)) : 12.748.143

Reads Fastq files

Reads are the raw diamonds of any NGS bioinformatics analysis.

Reads are collected in fastq files

One or two files per sample:

for RNA-Seq, DE, typically one: Single-End sequencing.

How many reads need to be produced?

Experiment and Library preparation kit dependent:

Differential Expression, typically

Illumina: 10-20M reads

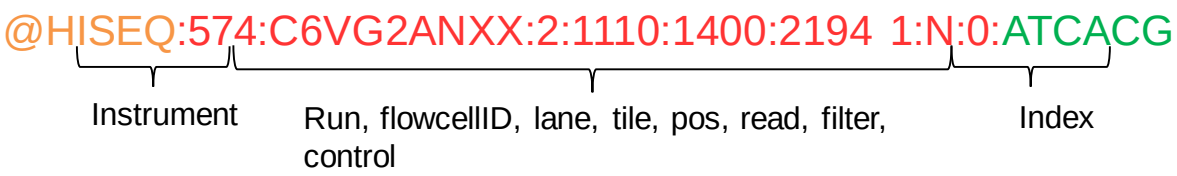
QuantSeq: 1-2M reads

SNP Calling Illumina: >100M reads.

Reads and Fastq Files

```
1 @HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
2 GGGGGATTCTCACTAGGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
3 +
4 CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

- Line 1: Read identifier and is followed by a sequence that
  - Unique, platform dependent
  - Begins with a '@' character



- Line 2: Raw sequence of nucleotides: read
- Line 3: begins with a '+' character and is optionally followed by the same sequence identifier.
- Line 4: Quality values for the sequence in Line 2

Read and Fastq: Base Calling Quality Control

```
@HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
GGGGGATTCTCACTAGGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

- A quality score (Q-score) is a prediction of the probability of an error in base calling.
- It serves as a compact way to communicate very small error probabilities
  - P = 10^(–Q/10)
  - Q = –10 log10(P)

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Reads: Base Calling Quality Control

```
@HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
GGGGGATTCTCACTAGGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

- Base: G
- Quality: C
- ASCII: 67
- Q: 34
- P: 0.00040

- Base: G
- Quality: G
- ASCII: 71
- Q: 38
- P: 0.00016

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

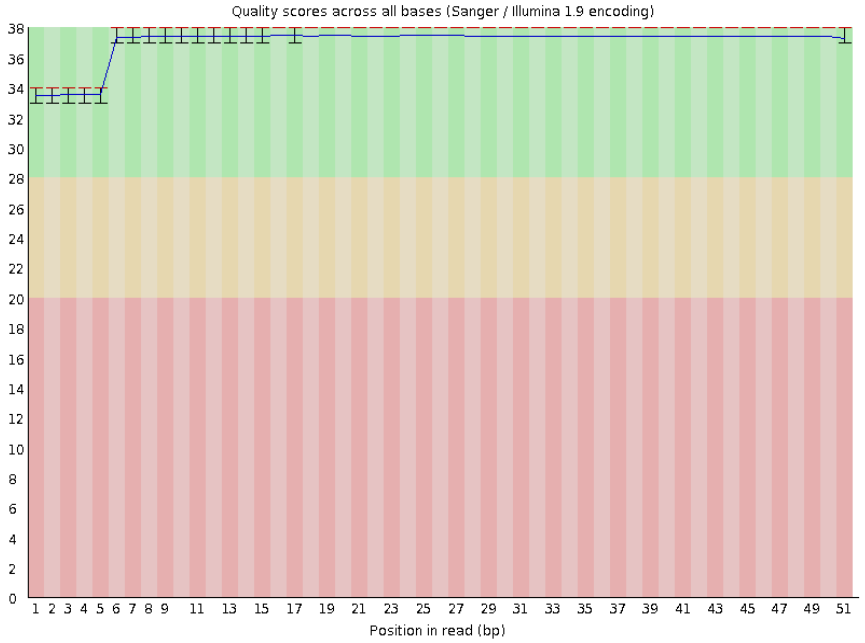
Reads: Base Calling Quality Control

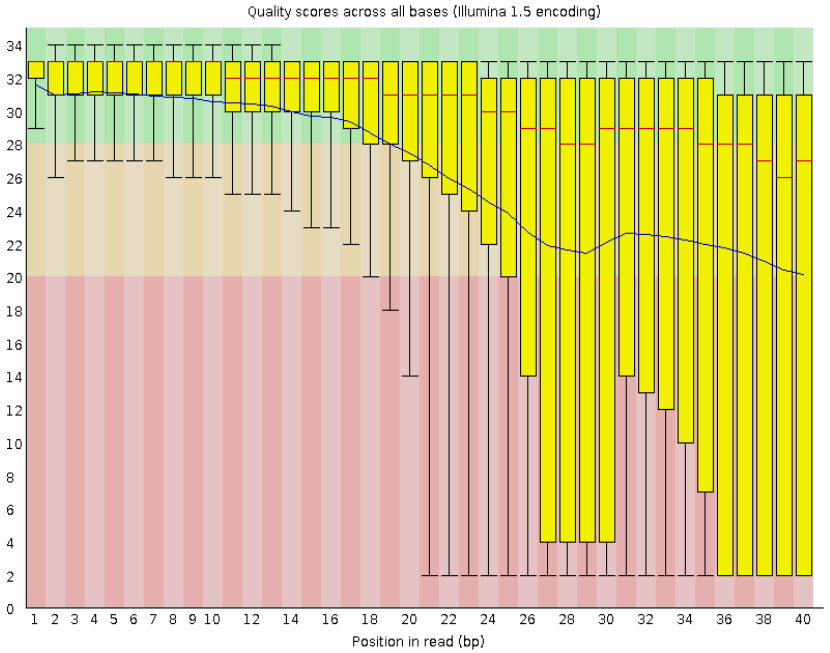
GENOMICS CORELEUVEN

Report: Aggregated Quality Control

Quick impression of whether reads as a whole present anomalies

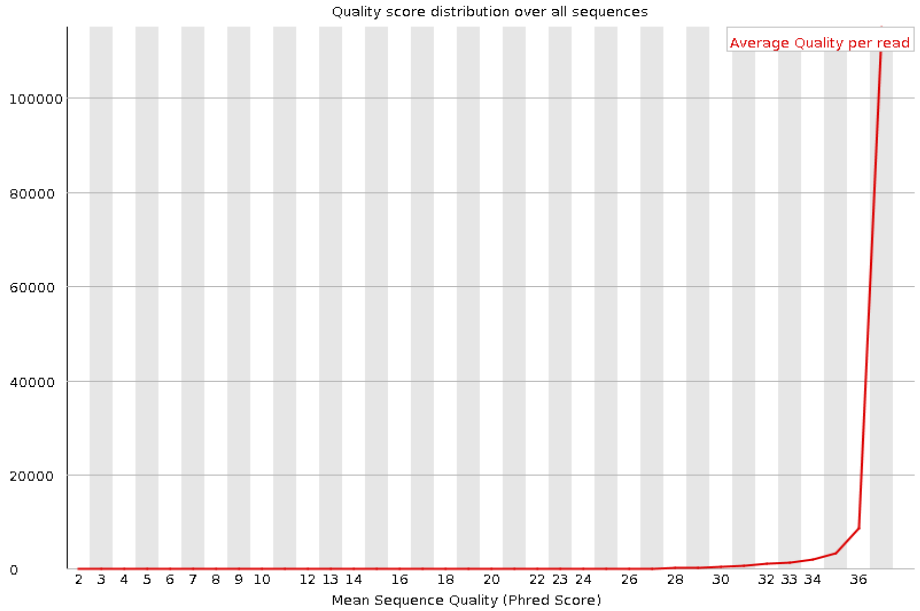
Per base sequence quality

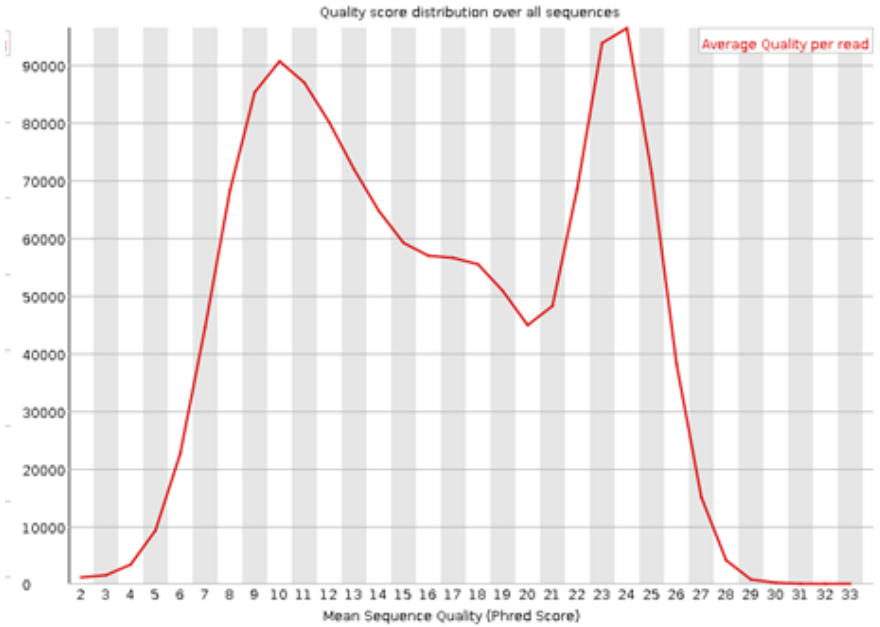




GENOMICS CORELEUVEN

Fastq Files Quality Control





GENOMICS CORELEUVEN

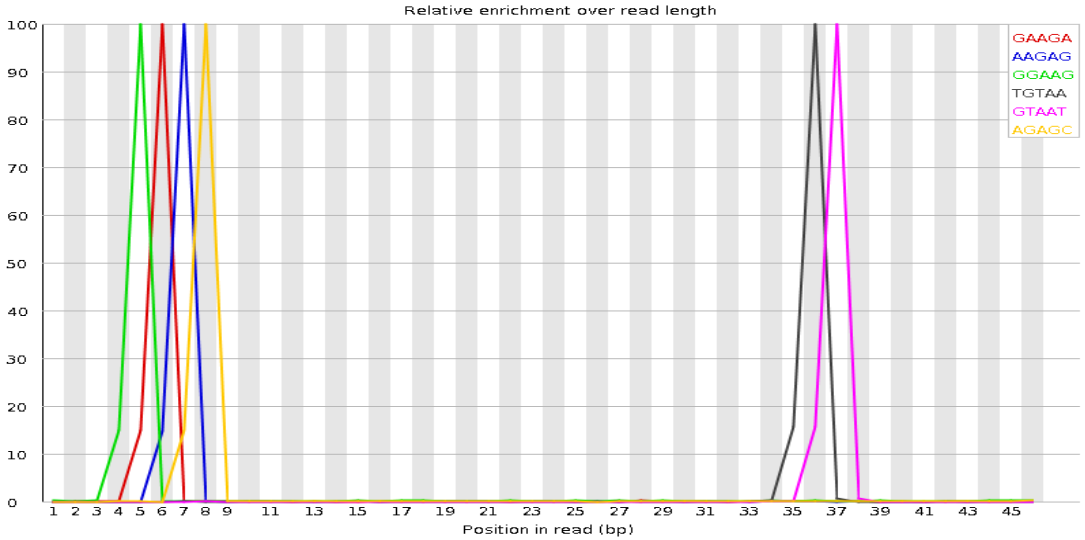
Fastq Files Quality Control

Overrepresented sequences:

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCC	40105	32.362577063361414	TruSeq Adapter, Index 12 (100% over 51bp)
ATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCG	5730	4.6238016849036505	TruSeq Adapter, Index 12 (100% over 51bp)

Kmer Content



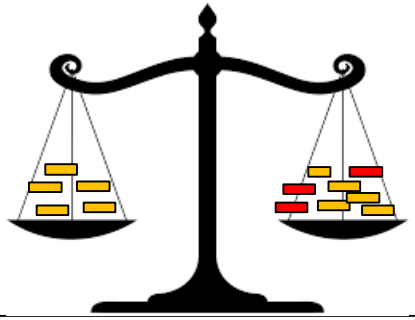
Number of reads: 14.103.785

Fastq Adapter Filter: 8.922602

GENOMICS CORELEUVEN

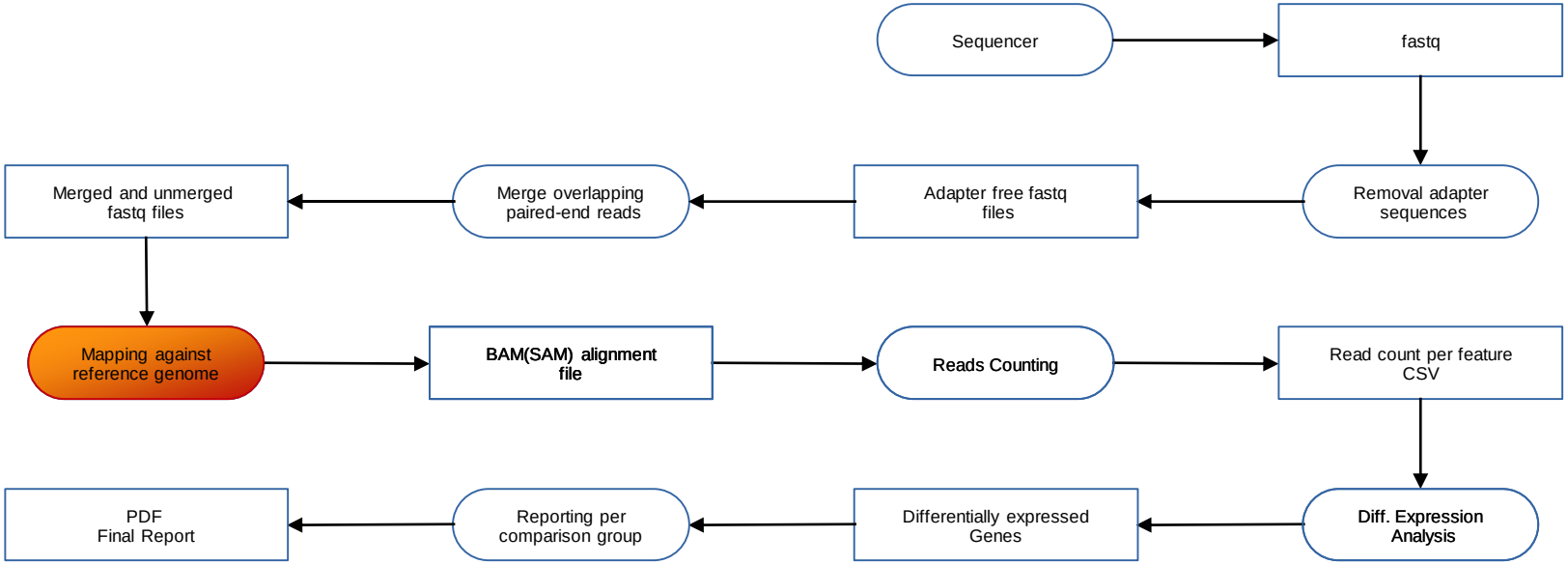
Preprocessing or not?

- Preprocessing remove/trimming reads with low quality data?
- Better quality input data with less noise…
  - Tradeoff: information is lost…can affect alignment



Quality scores are exploited by aligners

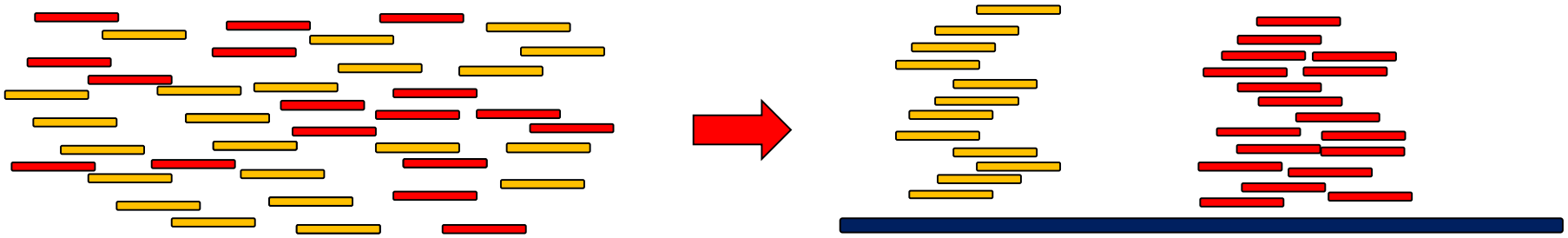
RNA-Seq Workflow Genomics Core



GENOMICS CORELEUVEN

From Raw Reads to Alignment

Where do reads come from?

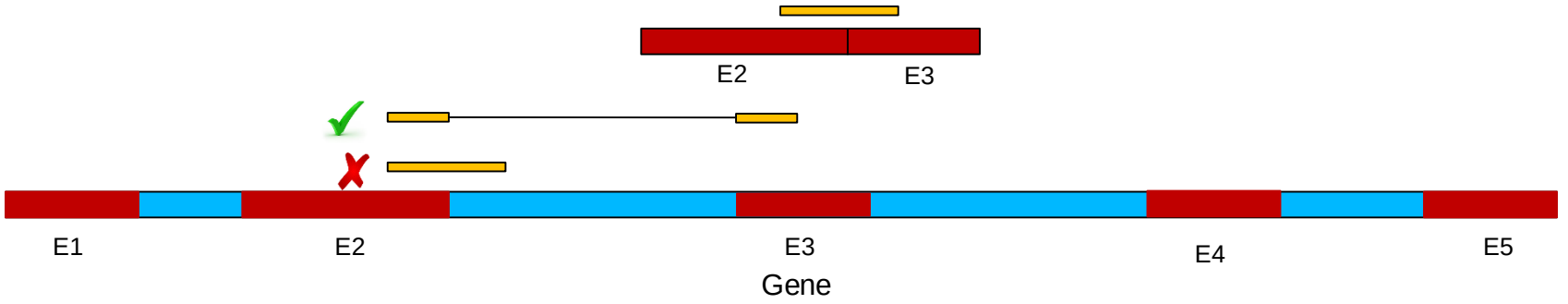


Alignment

GENOMICS CORELEUVEN

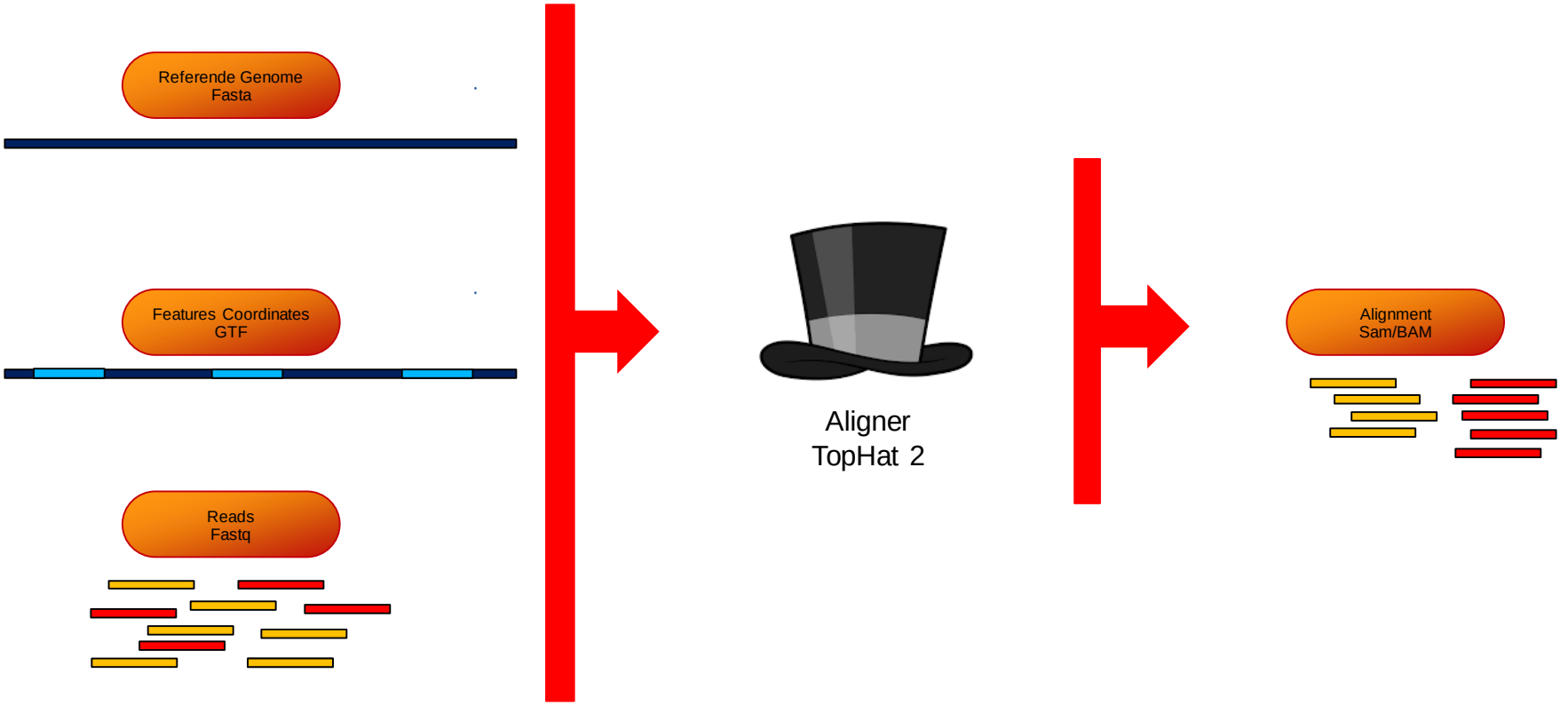
Splice-Aware Alignment

- Individual reads are aligned to a reference genome
- RNA-seq read alignment differs from standard aligment:
  - If an RNA-Seq read spans an exon boundary, part of the map will not map contiguosly to the reference
  - This causes the standard mapping procedure to fail



GENOMICS CORELEUVEN

Reference-Based Alignment with Tophat 2



GENOMICS CORELEUVEN

Reference Sequence in Fasta Format

- genome.fa human-readable nucleotide sequence
- Mouse genome: 2.6GB
- And it looks like this:

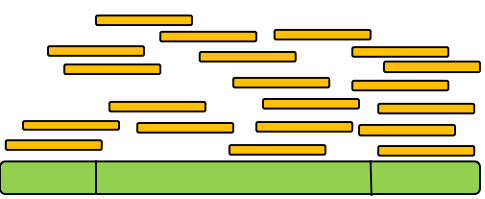
```
AATAAGTCAATGGCCTTTCTCTACACAAAGAATAAACAGGCTGAGAAAGAAATTAGGGAA
ACAACACCCTTCTCAATAGTCACAAATAATATAACATATCTCGGCGTGACTCTAACTAAG
GAAGTGAAAGATCTGTATGATAAAAACTTCAAGTCTCTGAAGAAAGAAATTAAAGAAGAT
CTCAGAAGATGGAAAGATCTCCCATGCTCATGGATTGGCAGGATCAATATTGTAAAAATG
GCTATCTTGCCAAAAGCAATCTACAGATTCAATGCAATCCCCATCAAAATTCCAACTCAA
TTCTTCAACGAATTAGAAGGAGCAATTTGCAAATTCATCTGTAATAACAAAAAACCTAGG
ATAGCAAAAAGTCTTCTCAAGGATAAAAGAACCTCTGGTGGAATCACCATGCCTGACCTA
AAGCTTTACTACAGAGCAATTGTGGTAAAAACTGCATGGTACTGGTATAGAGACAGACAA
GTAGACCAATGGAATAGAATTGAAGACCCAGAAATGAACCCACACACCTATGGTCACTTG
ATCTTCGACAAGGGAGCTAAAACCATCCAGTGGAAGAAAGACAGCATTTTCAACAAATGG
TGCTGGCACAACTGGTTGTTATCATGTAGAAGAATGCGAATCGATCCATACTTATCTCCT
TGTACTAAGGTCAAATCTAAATGGATCAAAGAACTTCACATAAAACCAGAGACACTGAAA
CTTATAGAGGAGAAAGTGGGGAAAAGCCTTGAAGATATGGGCACAGGGGAAAAATTCCTG
AACAGAACAGCAATGGCTTGTGCTGTAAGATTGAGAATTGACAAATGGGACCTAATGAAA
CTCCAAAGTTTCTGCAAGGCAAAAGACACCGTCAATAAGAGAAAGAGACCACCAACAGAT
TGGGAAAGGATCTTTACCTATCCTAAATCAGATAGGGGACTAATATCCAACATATATAAA
GAACTCAAGAAGGTGGACTTCAGAAAATCAAACAACCCCATTAAAAAATGGGGCTCAGAA
CTGAACAAAGAATTCTCACCTGAGTTATACCGAATGGCAGAGAAGCACCTGAAAAAATGC
TCAACATCCTTAATCATCAGGGAAATGCAAATCAAAACAACCCTGAGATTCCACCTCACA
CCAGTCAGAATGTCTAAGATCAAAAATTCAGGTGACAGCAGATGCTGGCGAGGATGTGGA
GAAAGAAGAACACTCCTCCATTGTTGGTGGGATTGCAGGCTTGTACAACCACTCTGGAAA
TCCGTCTGGCGGTTCCTCAGAAAATTGGACATAGTACTACCGGAGGATCCAGCAATACCT
CTCCTGGGCATATATCCAGAAGATGCCCCAACTGGTAAGAAGGACACATGCTCCACTATG
TTCATAGCAGCCTTATTTATAATAGCCAGAAGCTGGAAAGAACCCAGATGCCCCTCAACA
GAGGAATGGATACAGAAAATGTGGTACATCTACACAATGGAGTACTACTCAGCTATTAAA
AAGAATGAATTTATGAAATTCCTAGCCAAATGGATGGACCTGGAGGGCATCATCCTGAGT
```

GTF Files: Gene Transfer Format

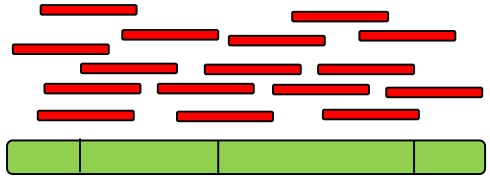
```
9   havana  gene   44107301  44108960  .  .  .  gene_id "ENSMUSG00000092258"; gene_version "1"; gene_name "Gm20444"; gene_source "havana"; gene_biotype "processed_transcript";
9   ensembl gene   44112923  44121947  .  +  .  gene_id "ENSMUSG00000097467"; gene_version "1"; gene_name "Gm26737"; gene_source "ensembl"; gene_biotype "lincRNA";
9   ensembl gene   44123768  44134485  .  -  .  gene_id "ENSMUSG00000097617"; gene_version "1"; gene_name "Gm10687"; gene_source "ensembl"; gene_biotype "lincRNA";
```

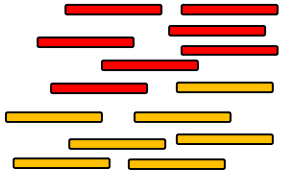
column-number	content	values/format
1	chromosome name	chr{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X,Y,M}
2	annotation source	{ENSEMBL,HAVANA}
3	feature-type	{gene,transcript,exon,CDS,UTR,start_codon,stop_codon,Selenocysteine}
4	genomic start location	integer-value (1-based)
5	genomic end location	integer-value
6	score (not used)	.
7	genomic strand	{+,-}
8	genomic phase (for CDS features)	{0,1,2,.}

Tophat Alignment Strategy

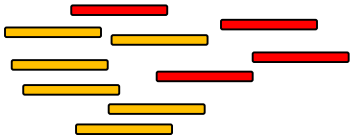


Transcriptome





Unmapped reads



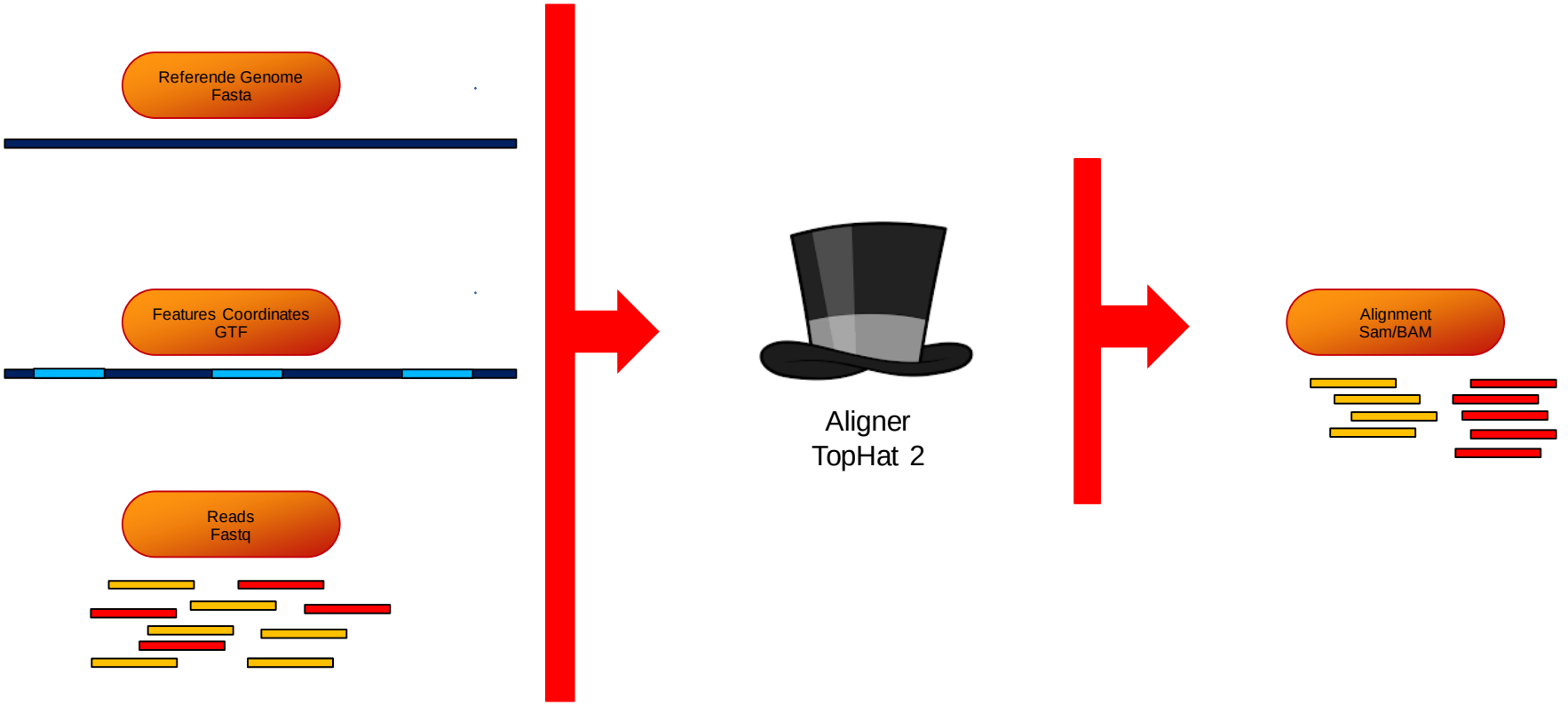
Unmapped reads

Reference Genome

Read are split-aligned

GENOMICS CORELEUVEN

Reference-Based Alignment with Tophat 2

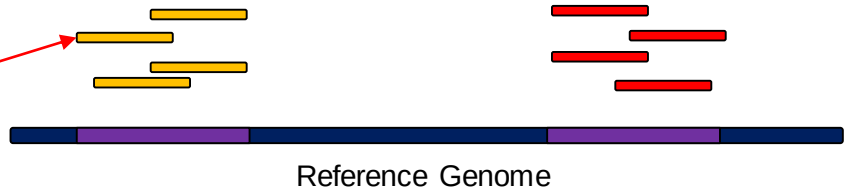


GENOMICS CORELEUVEN

SAM/BAM Files

Aligners produces SAM (Sequence Alignment Map) file

```
HISEQ:574:C6VG2ANXX:3:1307:21149:38188  0       1       3216913 50      51M     *       00
CTGGTAGGAGGCTAGGGCCCAAGCCAAGGACACAAGGGAGGCTGCTGCTGT     BBBCBGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGFGGGGGGGGGGG        AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:-  NH:i:1  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:2208:7076:98530   16      1       3532557 3       51M     *       00
TGTTCGGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTAT     GGBGGGGGGGGGGGGGGBGGGGGGGGG
GFCGGGGGGGGGGGGGGGGGBCBBC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10 CP:i:117736175  HI:i:0  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:2110:2743:48952   272     1       3532562 3       51M     *       00
GGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTATTTAAG     GGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGCCCCC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10 CP:i:117736170  HI:i:0  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:1115:10182:55408  16      1       3532562 3       51M     *       00
GGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTATTTAAG     GGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGCCCCC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10 CP:i:117736170  HI:i:0  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:1107:5697:49696   16      1       3532562 3       51M     *       00
GGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTATTTAAG     GGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGCCCCC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10 CP:i:117736170  HI:i:0  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:1315:2671:58947   256     1       3592903 1       51M     *       00
TTAAGACTGAATTCTGACATAGCTAAAGCCTTCCGCCAGTGTTCCAACAGT     BBCCCGGGGGGGGGGGGGGGGGGFGG
GGGGGGGGGGGGGGGGGGGGGGGGG       AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD
:Z:6T1A42       YT:Z:UU XS:A:-  NH:i:3  CC:Z:15 CP:i:10497071   HI:i:0  RG:Z:GC024982
```



BAM files are machine-readable versions of SAM files

GENOMICS CORELEUVEN

SAM Format Explained

```
HISEQ:574:C6VG2ANXX:3:2208:7076:98530   16      1       3532557 3       51M     *       00
TGTTCGGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTAT     GGBGGGGGGGGGGGGGBGGGGGGGGG
GFCGGGGGGGGGGGGGGGGGBCBBC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10 CP:i:117736175  HI:i:0  RG:Z:GC024982
```

HISEQ:574:C6VG2ANXX:3:2208:7076:98530: Query Name

16 00000010000 Decimal bitwise: Mapped? Strand?

1 Chromosome

3532557 Position

3 Mapping Quality

51M CIGAR: match? Skipped bases on reference?

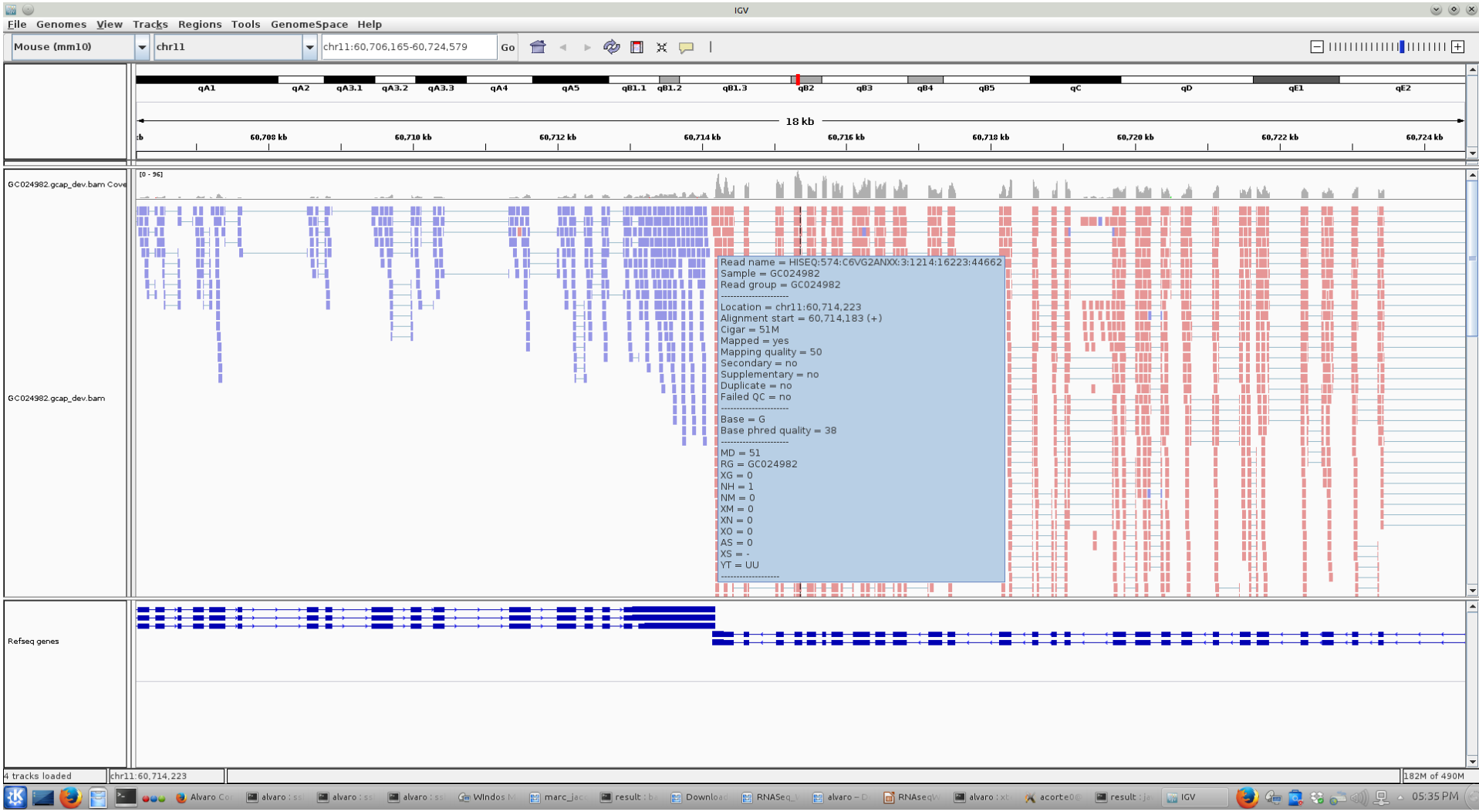
- Unpaired or paired (=)?

TGTTCGGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTAT
Reference

GGBGGGGGGGGGGGGGGBGGGGGGGGGGFCGGGGGGGGGGGGGGGGGGGGGGG Base Qualities

GENOMICS CORELEUVEN

Visualizing Alignment on IGV



GENOMICS CORELEUVEN

Counting Reads on HTSeq

Method

- Count each read only once
- Multi-mapping reads and reads overlapping multiple features discarded

	union	intersection _strict	intersection _nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

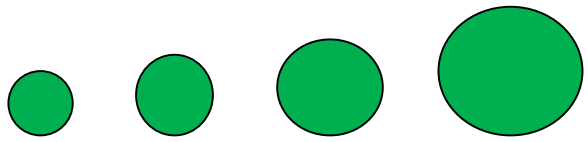
Simon Anders

GENOMICS CORELEUVEN

Counting Reads on HTSeq

Resulting table of absolute counts for each sample and feature:

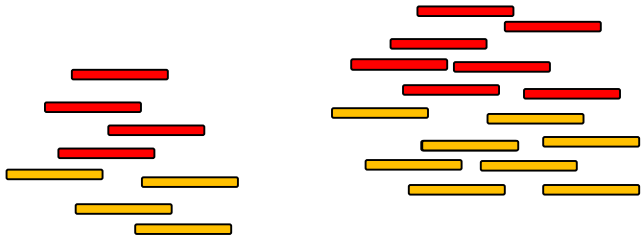
	Sample 1	Sample 2	Sample 3	Sample4
FBgn0000003	0	0	0	1
FBgn0000008	76	70	88	70
FBgn0000014	0	0	0	0
FBgn0000015	1	2	0	0
FBgn0000017	3564	3150	3072	3334
FBgn0000018	245	310	299	308



Library sizes

Counting Reads on HTSeq

Normalization of Counts: Impact Library Size



A bigger library produces more reads...But in slide 5:

“Read Counts is linearly related to the abundance of the target transcript”

GENOMICS CORELEUVEN

Counting Reads

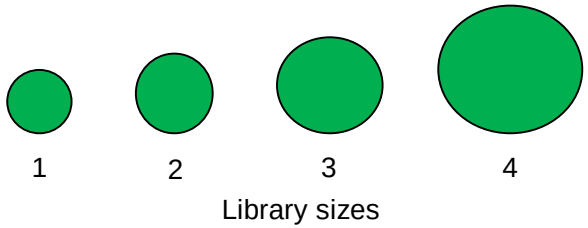
Normalization of Counts: Size Factors

Adjust for library sizes to produce count values in a common scale:

Sample 1	Sample 2	Sample 3	Sample4
0.873	1.011	1.022	1.115

→ Size factors

	Sample 1	Sample 2	Sample 3	Sample4
FBgn0000003	0	0	0	1
FBgn0000008	76	70	88	70
FBgn0000014	0	0	0	0
FBgn0000015	1	2	0	0
FBgn0000017	3564	3150	3072	3334
FBgn0000018	245	310	299	308



Library sizes

Normalized counts

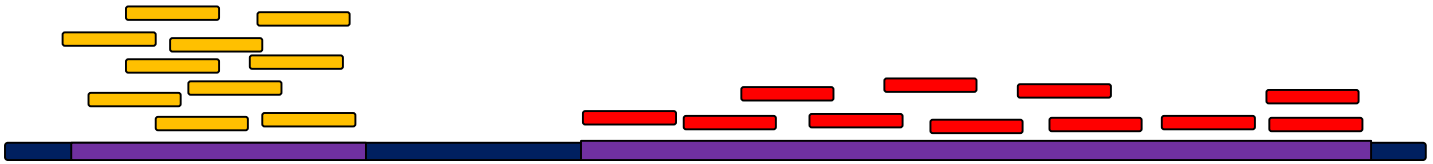
Feature	Sample 1	Sample 2	Sample 3	Sample4
FBgn0000003	0.00	0.00	0.0	0.897
FBgn0000008	87.05	69.27	86.1	62.803
FBgn0000014	0.00	0.00	0.0	0.000
FBgn0000015	1.15	1.98	0.0	0.000
FBgn0000017	4082.02	3116.93	3004.5	2991.238
FBgn0000018	280.61	306.75	292.4	276.335

76/0.873 = 87.05

GENOMICS CORELEUVEN

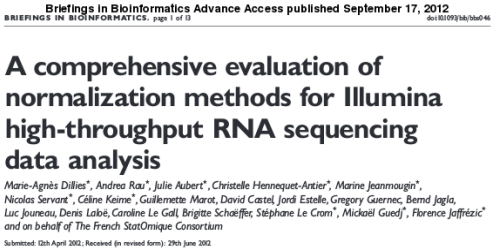
Counting Reads

More on Normalization



- Possible extra normalization:
  - Longer transcripts are more likely to have sequences mapped to their genes
    - Higher counts, biasing comparisons between transcripts of different lengths.
  - *RPKM - Reads per kilo base per million mapped reads*

"based on three real mRNA and one miRNA-seq datasets,we confirm previous observations that RPKM and TC, both of which are still widely in use [40,41], are ineffective and should be definitively abandoned in the context of differential analysis"



Briefings in Bioinformatics Advance Access published September 17, 2012

BRIEFINGS IN BIOINFORMATICS. page 1 of 13

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

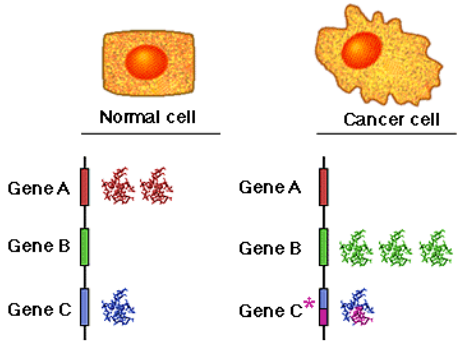
Marie-Agnès Dillies*, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

Differential Expression Analysis

•Statistical test: Decide whether for a given gene, an observed difference in reads counts is significant

•Comparison between different biological conditions



- Read counts per gene are modeled by a probability distribution

Deseq or EdgeR?

Box 2 | Differences between DESeq and edgeR

The two packages described in this protocol, DESeq and edgeR, have similar strategies to perform differential analysis for count data. However, they differ in a few important areas. First, their look and feel differs. For users of the widely used limma package[60] (for analysis of microarray data), the data structures and steps in edgeR follow analogously. The packages differ in their default normalization: edgeR uses the trimmed mean of M values[56], whereas DESeq uses a relative log expression approach by creating a virtual library that every sample is compared against; in practice, the normalization factors are often similar. Perhaps most crucially, the tools differ in the choices made to estimate the dispersion. edgeR moderates feature-level dispersion estimates toward a trended mean according to the dispersion-mean relationship. In contrast, DESeq takes the maximum of the individual dispersion estimates and the dispersion-mean trend. In practice, this means DESeq is less powerful, whereas edgeR is more sensitive to outliers. Recent comparison studies have highlighted that no single method dominates another across all settings[27,61,62].

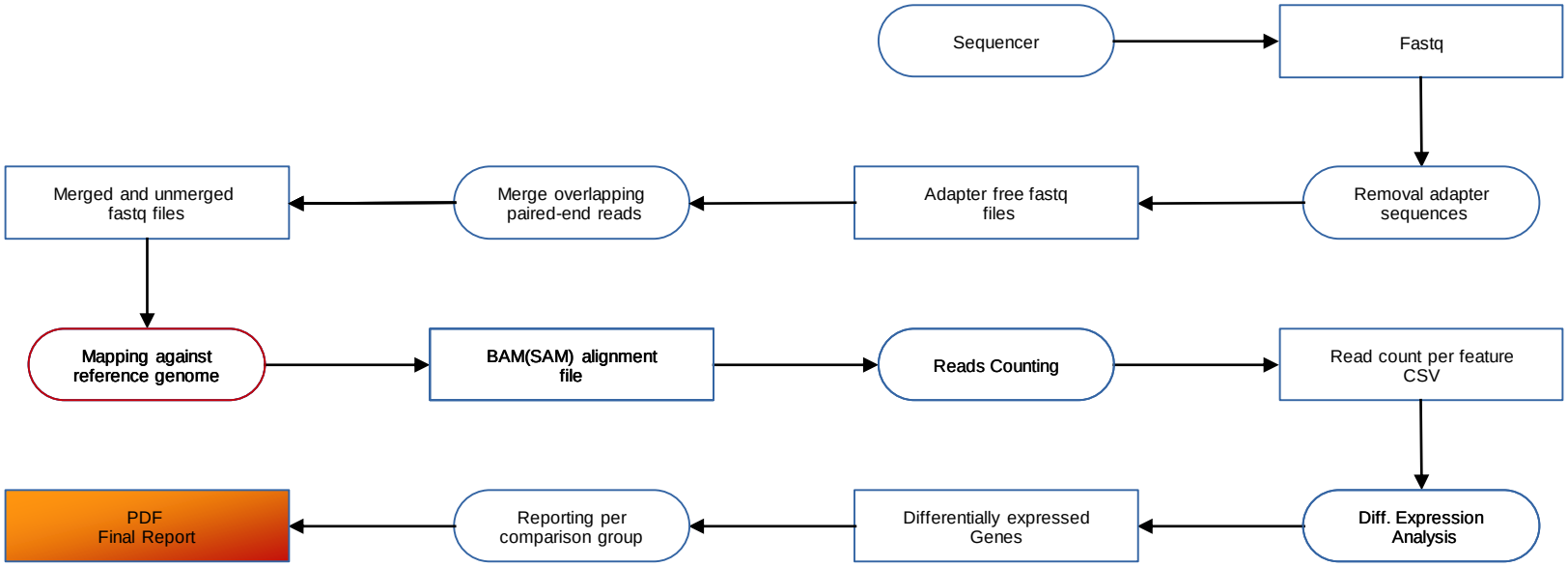
PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]

GENOMICS CORELEUVEN

RNA-Seq Workflow at Genomics Core



GENOMICS CORELEUVEN

Understanding DeSeq Results

id	baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj
TTp9	61.2142079613	114.5039786047	7.9244373179	0.0692066548	-3.8529454185	3.18278695454205E-038	7.8910836963961E-034
Roums4	111.7253852962	3.3863424303	220.064428162	64.9858756734	6.0220542852	3.14004006066747E-016	3.89255066120643E-012
Serinc3	5049.0292624521	2783.4853979737	7314.5731269304	2.6278467752	1.3938811573	8.88017567331908E-011	7.33887318228667E-007
Apoba	687.9674505131	478.6234549012	897.3114461251	1.8747753311	0.9067177166	3.02197214770456E-005	0.1833853013
Psiga	318.5920273479	219.6643392529	417.519715443	1.9007168704	0.926543645	3.69832818348108E-005	0.1833853013
Igdskv12-98	3.4254621585	0.697231647	6.1536926701	8.8258940866	3.1417424346	7.3351076233012E-005	0.2589065108
Ssdlc12a5	17.8775194124	28.617735224	7.1373036008	0.249401413	-2.0034584562	8.04863286246797E-005	0.2589065108
Treb3l3	31.0676067843	13.9815013501	48.1537122185	3.444101675	1.7841277338	8.35418096347497E-005	0.2589065108
Tm16316	1.205992866	2.4119857321	0	0	-Inf	0.0001326348	0.3653795179
Ces1d	553.0344051691	424.629437072	681.4393732662	1.604785994	0.6823809198	0.0002487961	0.6168402322
Vsig1	134.705417607	168.1043611548	101.3064740591	0.6026403679	-0.7306307796	0.0005224394	1

RNA-Seq at Genomics Core

Experimental Design & Project Management

Céline Helsmoortel
Celine.helsmoortel@uzleuven.be

Gregory Maes
Gregory.maes@kuleuven.be

Alvaro Cortes C.
alvaro.cortes@uzleuven.be

Laboratory Support

Vanessa Brys
vanessa.brys@uzleuven.be

Bioinformatics Analysing & Reporting

Álvaro Cortés C.
alvaro.cortes@uzleuven.be

GENOMICS CORELEUVEN