Introduction to Next Generation Sequencing

RNA-Seq: Differential Expression of Genes

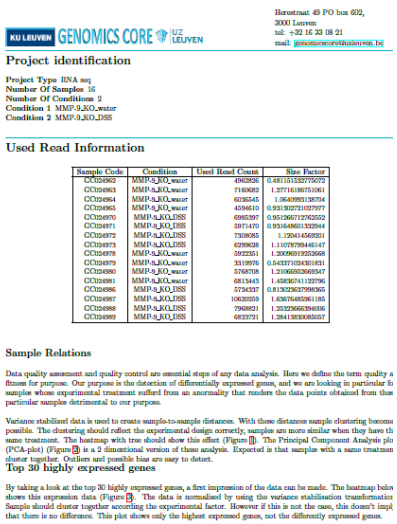
Bioinformatics Workshop
31 March 2017

GENOMICS CORELEUVEN

Overview

- Differential Expression Principles
- Splice-aware Alignment and Counting
- Differential Expression Table

A Typical RNA-Seq Analysis at GC



KU LEUVEN GENOMICS CORE UZ LEUVEN

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

Project identification

Project Type RNA seq
Number Of Samples 16
Number Of Conditions 2
Condition 1 MMP-9_KO_ester
Condition 2 MMP-9_KO_DSS

Used Read Information

Sample Code	Condition	Used Read Count	Size Factor
CO24962	MMP-9_KO_ester	4963826	0.481151552775072
CO24963	MMP-9_KO_ester	7160682	1.27716186751061
CO24964	MMP-9_KO_ester	6036545	1.06409931138704
CO24965	MMP-9_KO_ester	4184610	0.931302731027977
CO24970	MMP-9_KO_DSS	6983397	0.951266712762552
CO24971	MMP-9_KO_DSS	5971470	0.931648601332944
CO24972	MMP-9_KO_DSS	7308365	1.12941456829351
CO24973	MMP-9_KO_DSS	6299628	1.11078799446147
CO24978	MMP-9_KO_ester	5922351	1.20096919253668
CO24979	MMP-9_KO_ester	3319976	0.543371093631831
CO24980	MMP-9_KO_ester	5768708	1.21066952669347
CO24981	MMP-9_KO_ester	6813443	1.45836741122796
CO24986	MMP-9_KO_DSS	5734337	0.813029267998365
CO24987	MMP-9_KO_DSS	10632059	1.63676485961185
CO24988	MMP-9_KO_DSS	7968821	1.35323666394006
CO24989	MMP-9_KO_DSS	6823721	1.28413939083697

Sample Relations

Data quality assessment 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 suffurd 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 become
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.

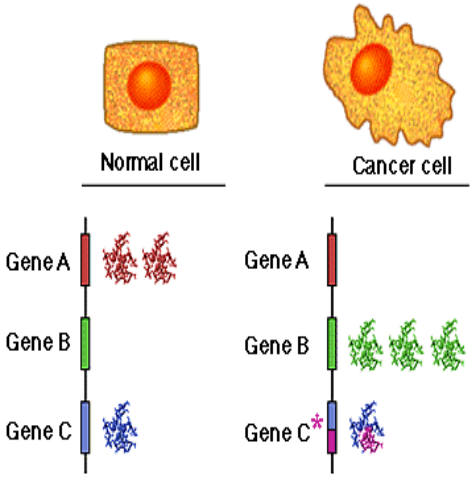
3GB of raw and analyzed data: Fastq, BAM, counts, normalized counts, gene expression files, heatmaps…

GENOMICS CORELEUVEN

RNA-Seq Differential Expression Analysis

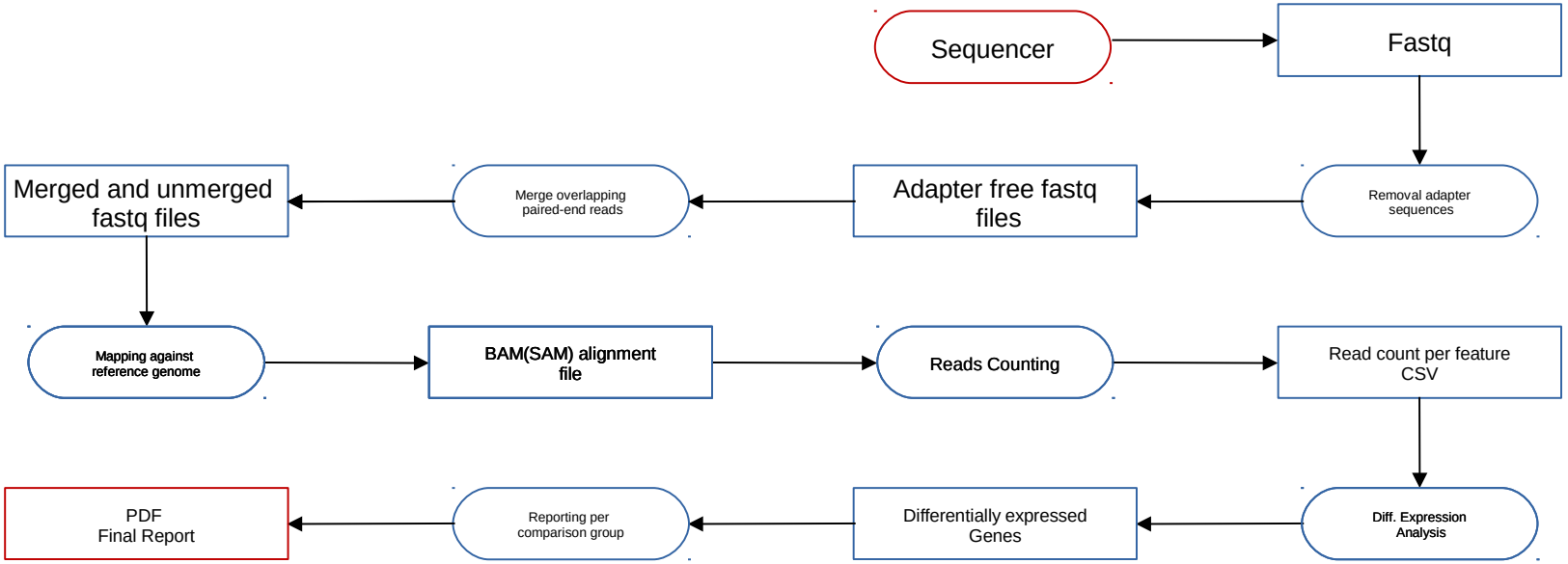
RNA-Seq is a technique that allows transcriptome studies based on next-generation sequencing technologies

- SNP calling
- Fusion gene detection
- Differential expression (DE)



DE RNA-Seq: compare gene expression between conditions

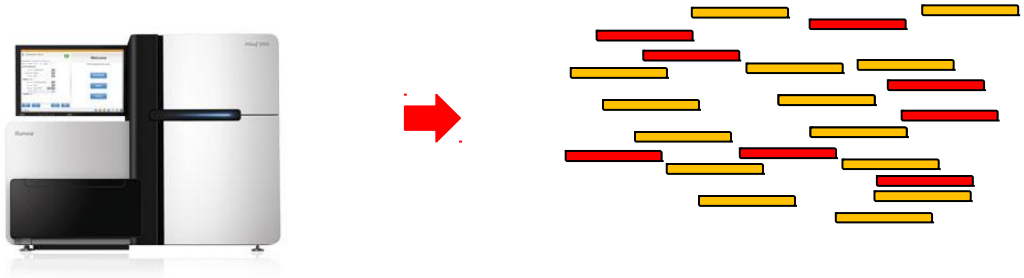
RNA-Seq Workflow (or pipeline) at GC



GENOMICS CORELEUVEN

RNA-Seq Workflow at GC

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



How reads relate differential expression of genes?

Goal: discover changes in abundance between conditions

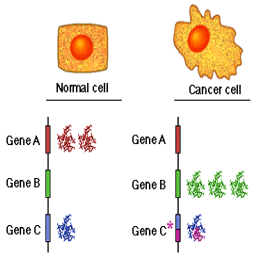
Read count is linearly related to the abundance of the target transcript

Differential Expression of Genes

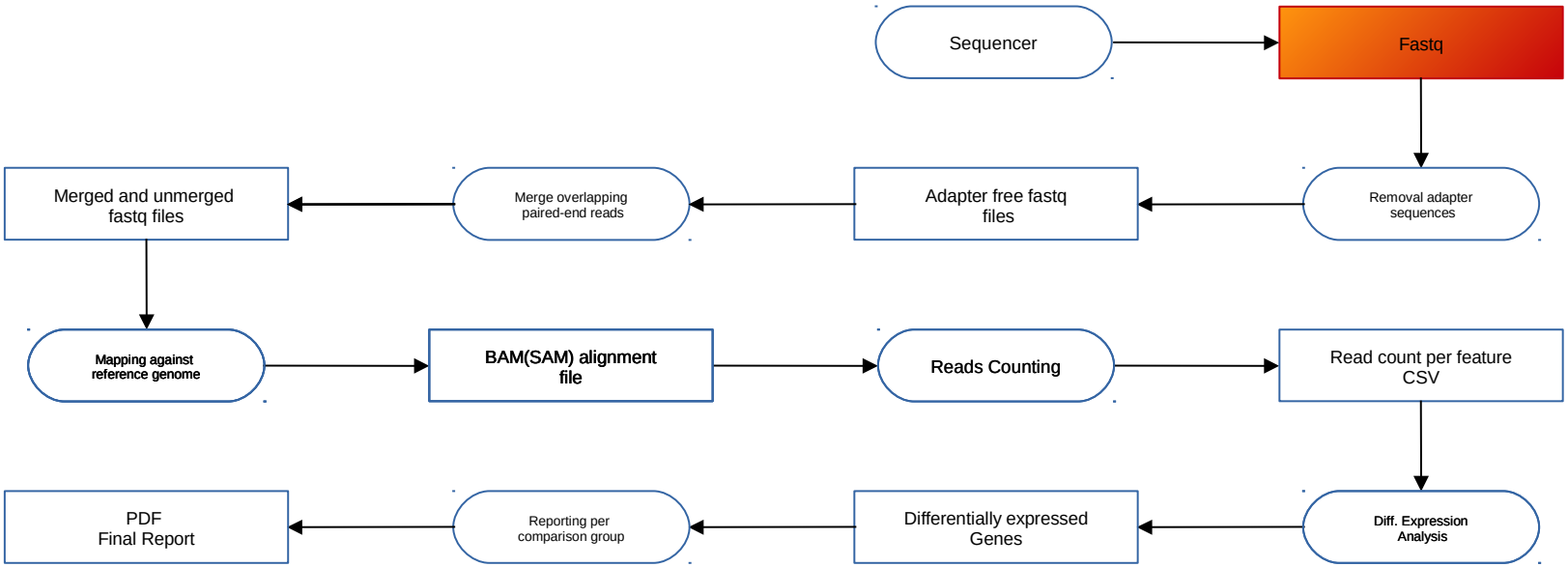
- Count the number of reads that fall into annotated genes
  - Fall: Align reads w.r.t. a reference genome (transcriptome)
  - Count: number of aligned reads per feature (genes).

Reference Genome

- 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



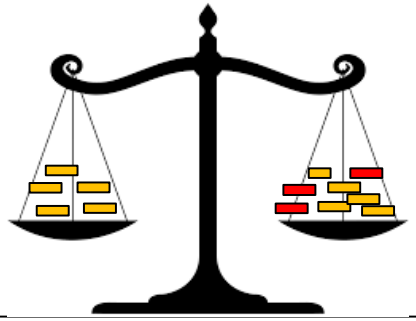
RNA-Seq Workflow at Genomics Core



GENOMICS CORELEUVEN

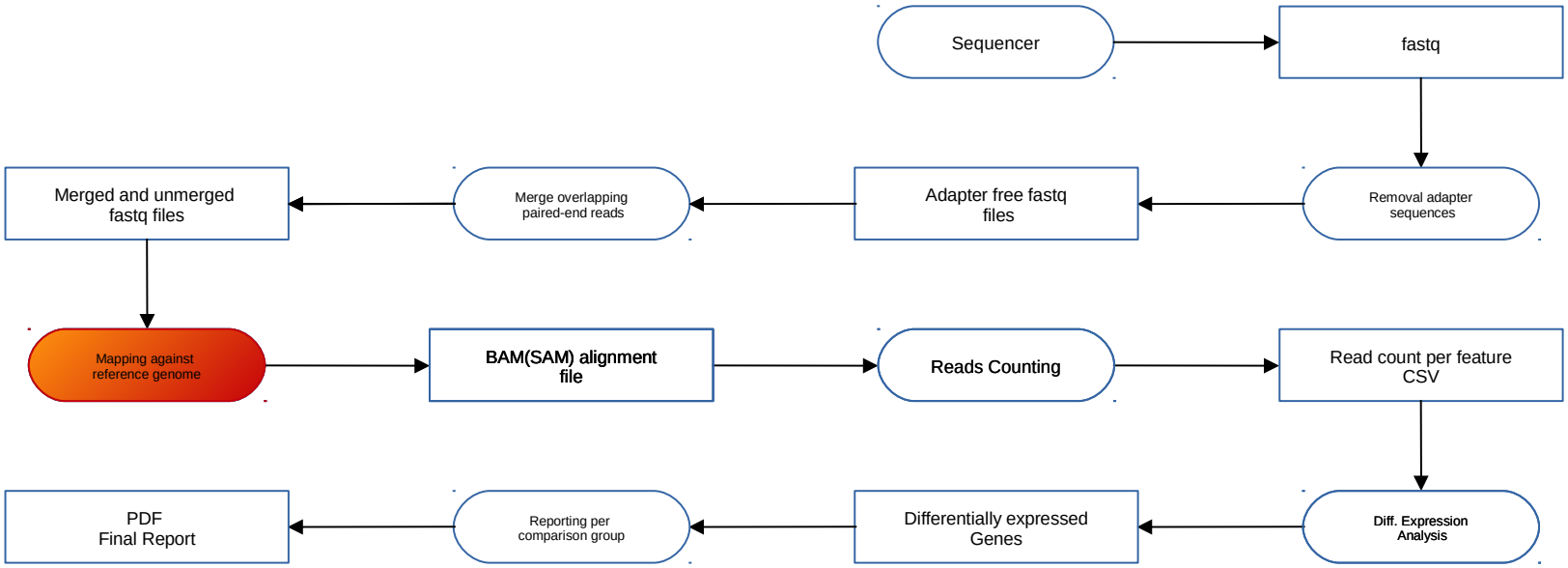
Preprocessing or not?

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



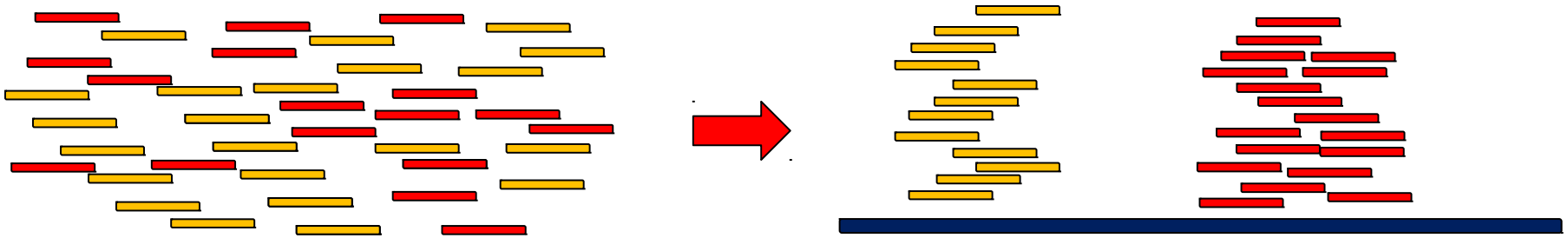
Quality scores are exploited by aligners

RNA-Seq Workflow at Genomics Core



GENOMICS CORELEUVEN

From Raw Reads to Alignment



GENOMICS CORELEUVEN

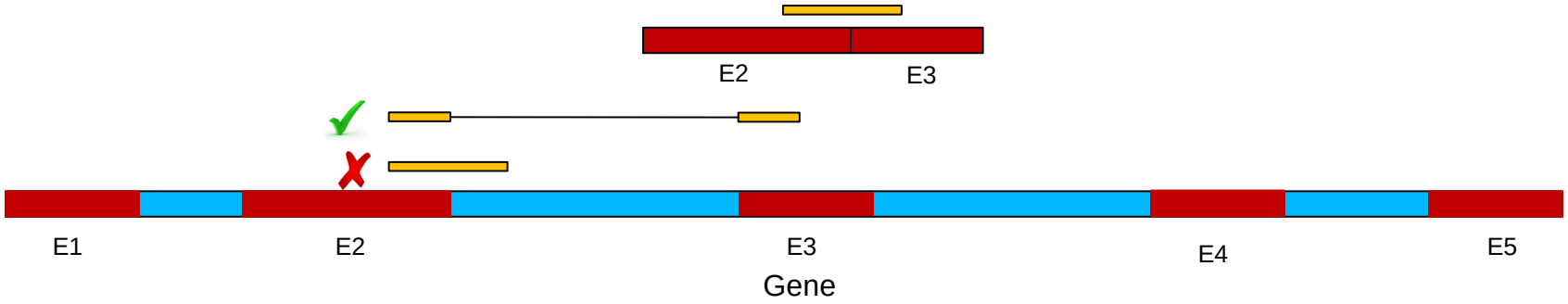
Tophat: Splice-Aware Alignment

Individual reads are aligned to a reference genome

RNA-seq read alignment differs from standard aligment:

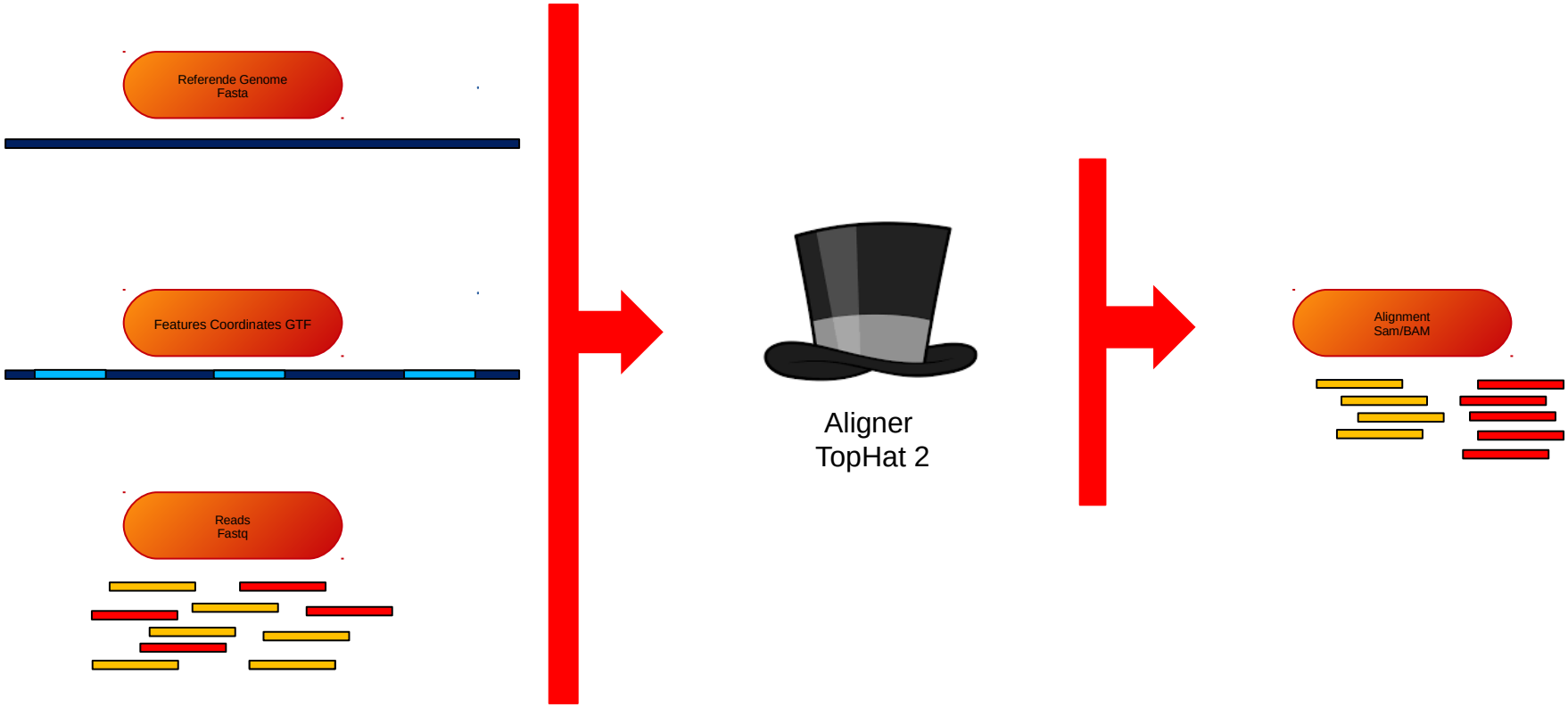
If an RNA-Seq read spans an exon boundary, part of the map will not map contiguosly to the reference

This causes the standard mapping procedure to fail



GENOMICS CORELEUVEN

Reference-Based Alignment with Tophat 2



GENOMICS CORELEUVEN

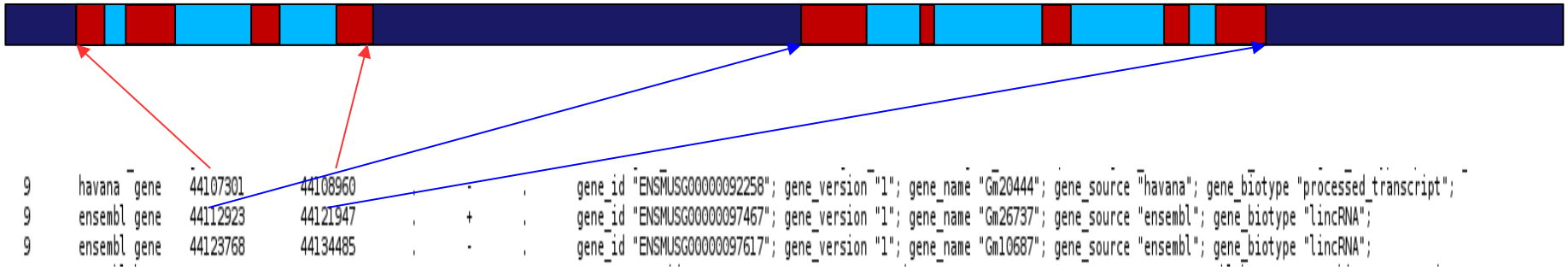
Reference Sequence in Fasta Format

- 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
```


GENOMICS CORELEUVEN

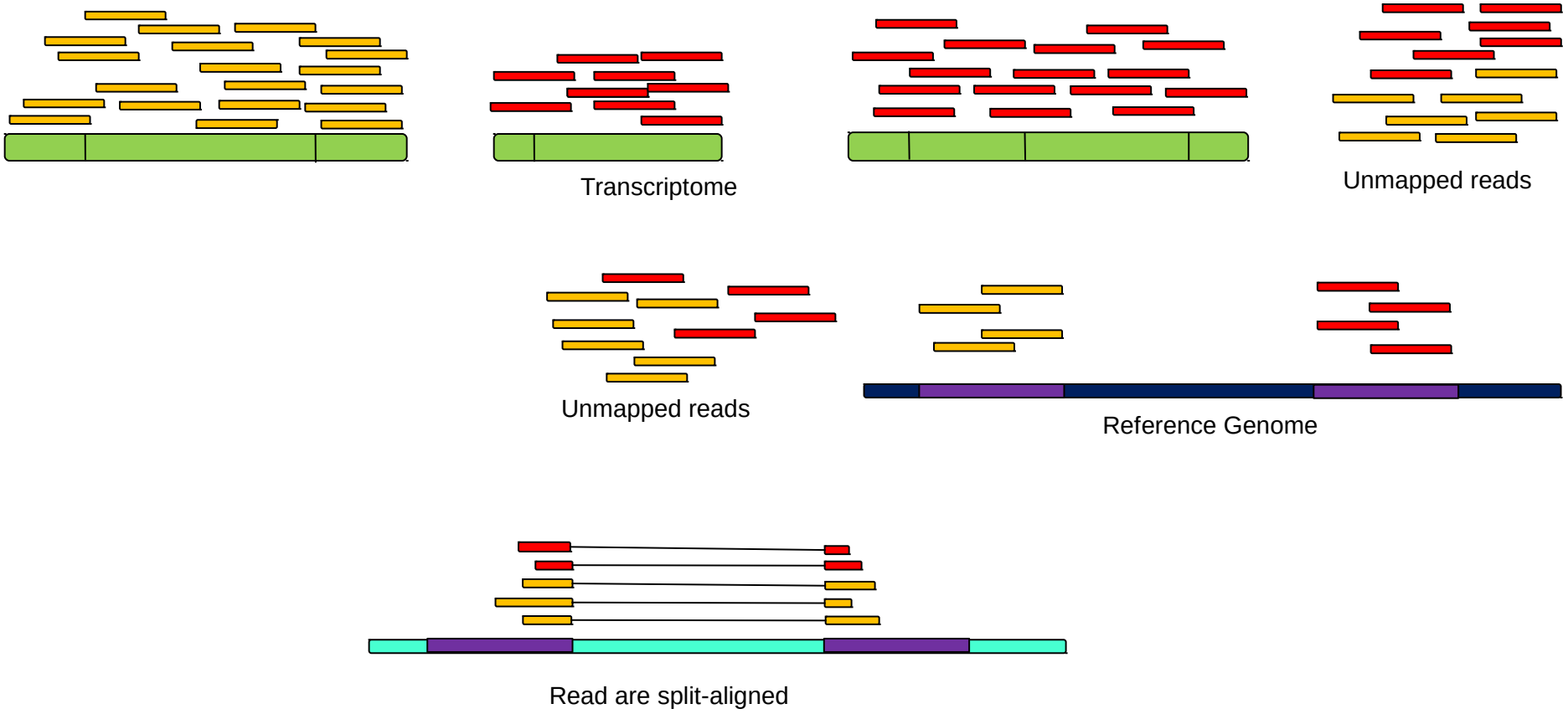
GTF Files: Gene Transfer Format



```
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";
```

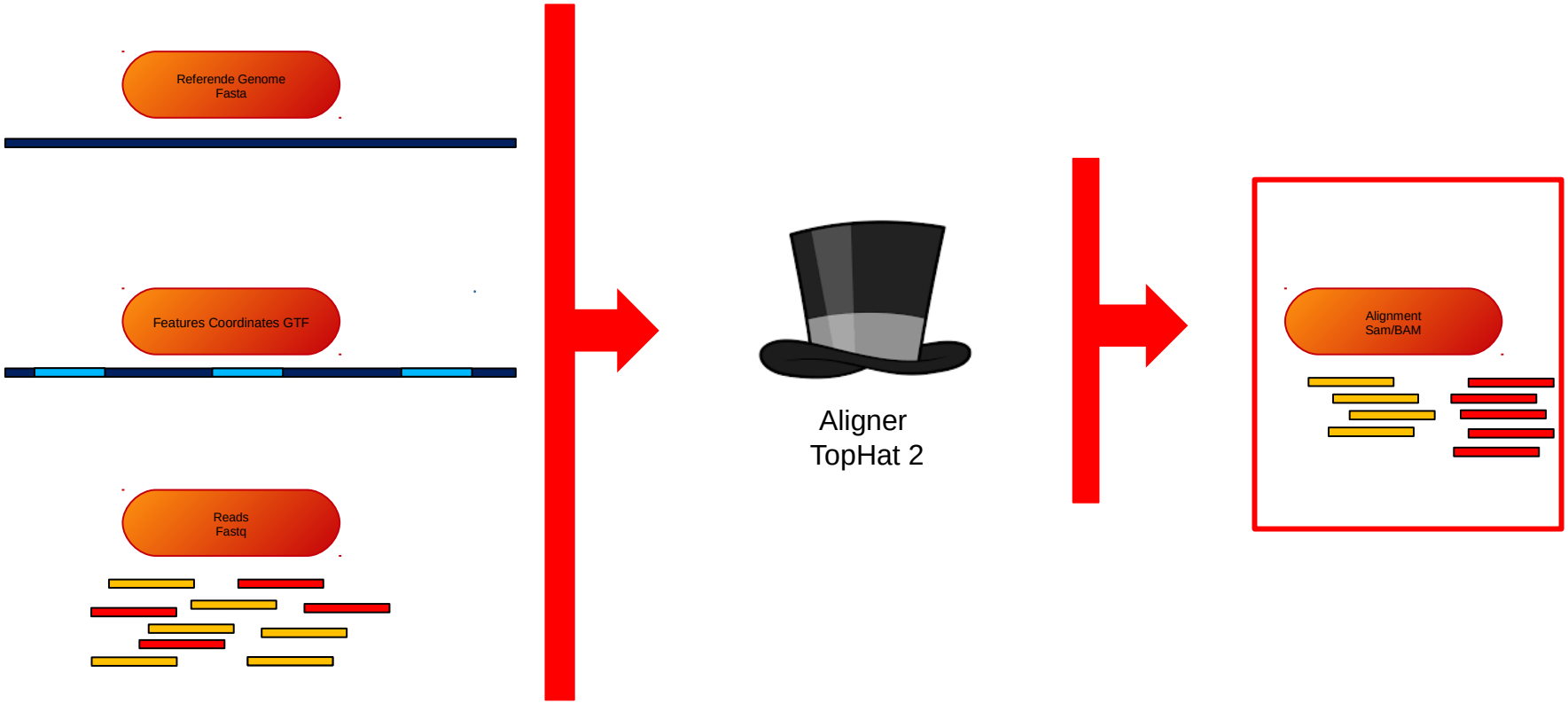
column-number	content	values/format
1	chromosome name	chr{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X,Y,M}
2	annotation source	{ENSEMBL,HAVANA}
3	feature-type	{gene,transcript,exon,CDS,UTR,start_codon,stop_codon,Selenocysteine}
4	genomic start location	integer-value (1-based)
5	genomic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

TopHat Alignment Strategy – GTF



GENOMICS CORELEUVEN

Reference-Based Alignment with Tophat 2



GENOMICS CORELEUVEN

RNA-Seq BAM/SAM

Extract SAM (Sequence Alignment Map) file

```
HISEQ:574:C6VG2ANXX:3:1307:21149:38188  0       1       3216913 50      51M     *       00
CTGGTAGGAGGCTAGGGCCCAAGCCAAGGACACAAGGGAGGCTGCTGCTGT     BBBCBGGGGGGGGGGGGGGGGGGGG
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     GGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGCCCCC        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     GGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGCCCCC        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     GGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGCCCCC        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
GGGGGGGGGGGGGGGGGGGGGGGG        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

RNA-Seq BAM/SAM

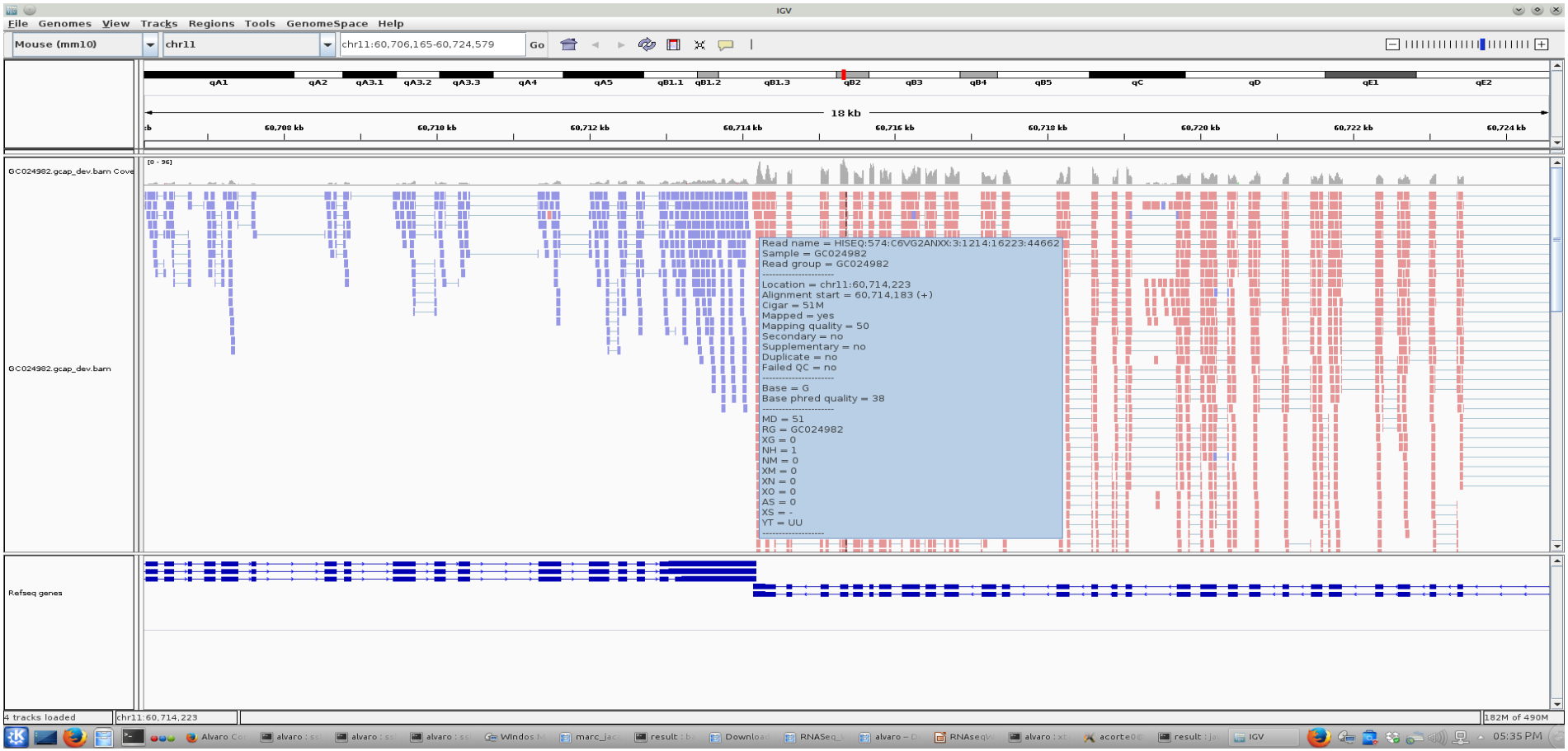
```
HISEQ:574:C6VG2ANXX:3:2208:7076:98530   16      1       3532557 3       51M     *       00
TGTTCGGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTAT     GGBGGGGGGGGGGGGGGBGGGGGGGGG
GFCGGGGGGGGGGGGGGGGGBCBBC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10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

HISEQ:574:C6VG2ANXX:3:2208:7076:98530: 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
GGBGGGGGGGGGGGGGGBGGGGGGGGGGFCGGGGGGGGGGGGGGGGGGGGGG Base Qualities
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:51 YT Base Qualities

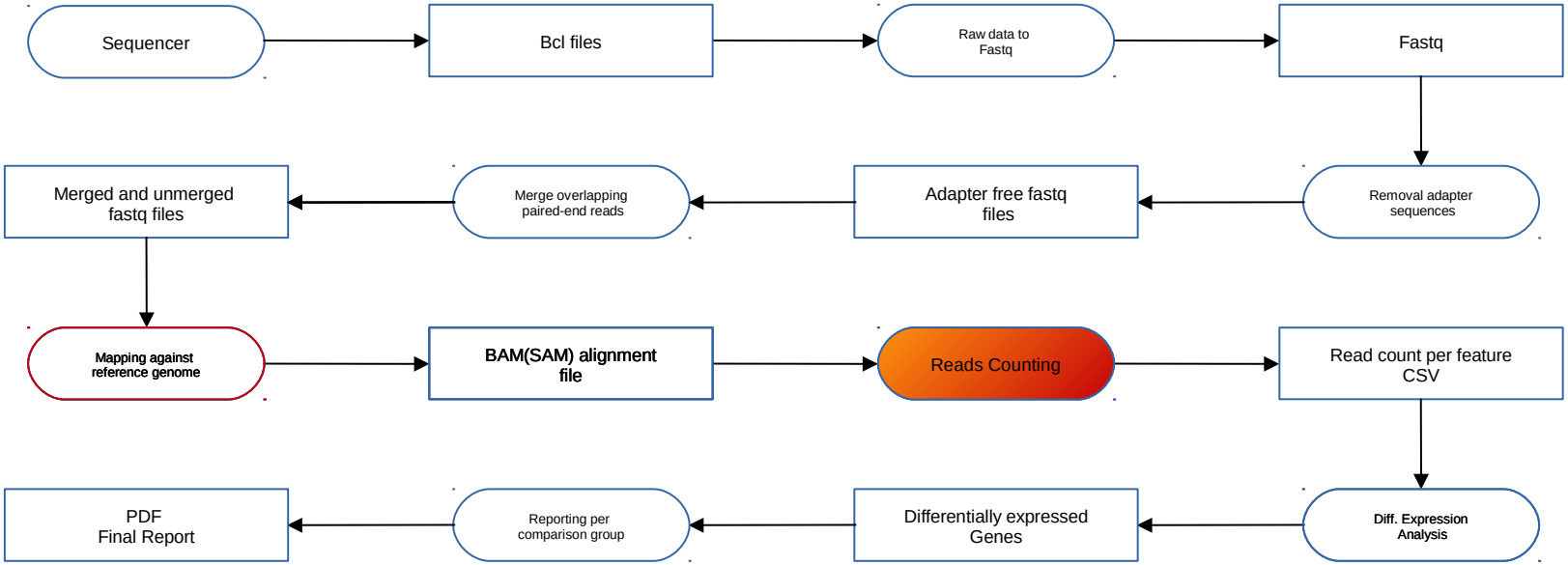
GENOMICS CORELEUVEN

RNA-Seq BAM/SAM



GENOMICS CORELEUVEN

RNA-Seq Workflow



GENOMICS CORELEUVEN

Counting Reads in 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

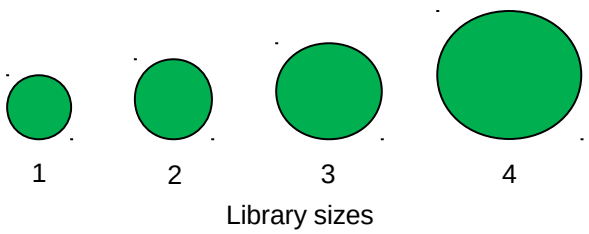
GENOMICS CORELEUVEN

Counting Reads in HTSeq(*)

(*) Simon Anders, Wolfgang Huber (EMBL)

Resulting table of absolute counts for each sample and feature:

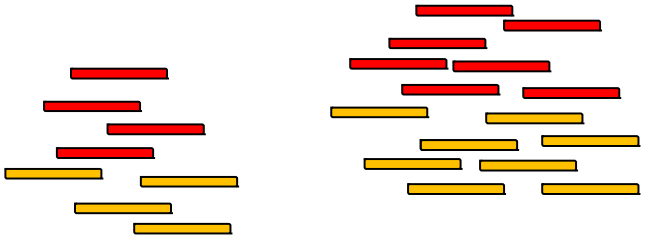
	Sample 1		Sample 2	Sample 3	Sample4
FBgn0000003	0	0	0	1	
FBgn0000008	76	70	88	70	
FBgn0000014	0	0	0	0	
FBgn0000015	1	2	0	0	
FBgn0000017	3564	3150	3072	3334	
FBgn0000018	245	310	299	308	



Library sizes

GENOMICS CORELEUVEN

Normalization of Counts: Impact Library Size



Library 1 Library 2

A bigger library produces more reads...But in slide 3:

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

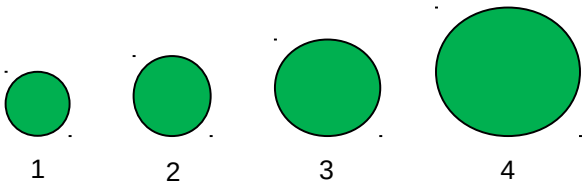
Normalization of Counts: Size Factors

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



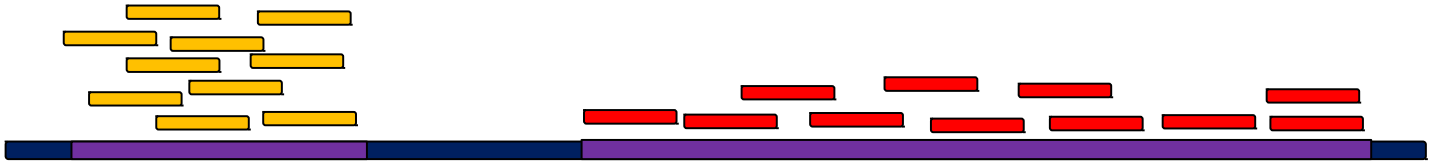
Normalized counts

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

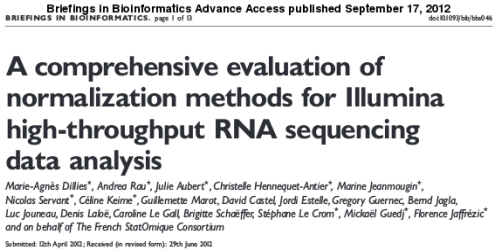
GENOMICS CORELEUVEN

More on Counting and 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

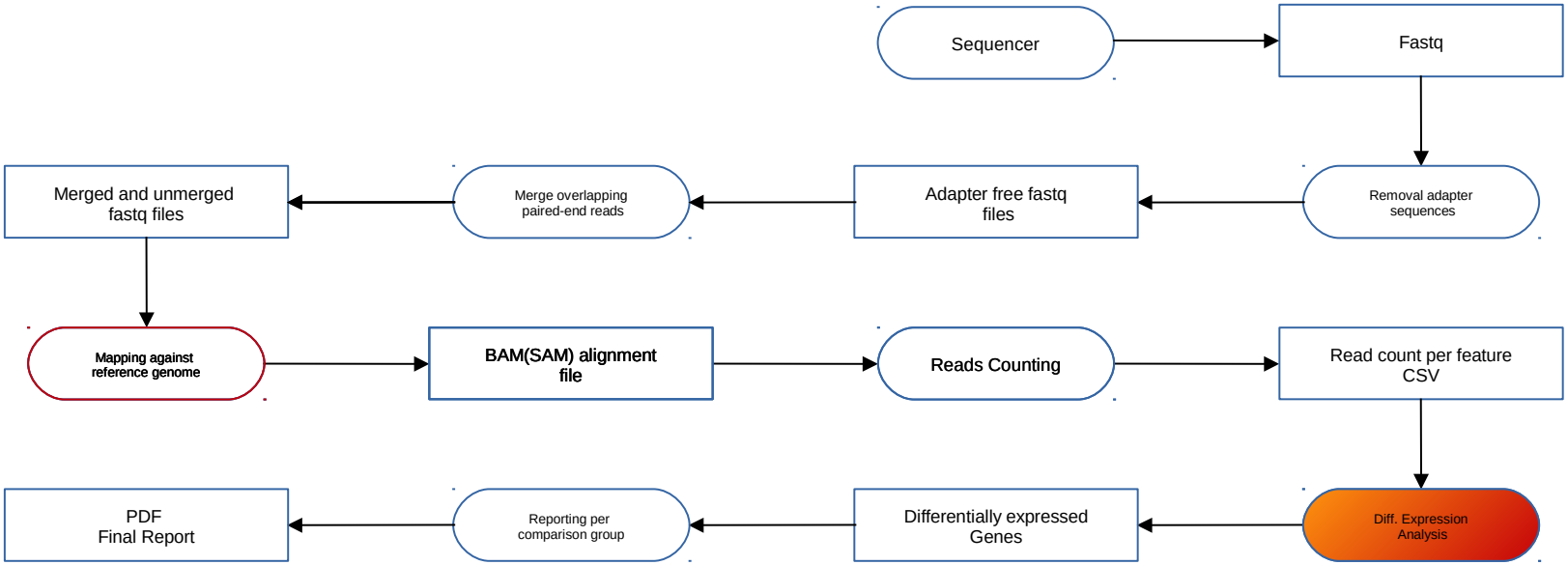
BRIEFINGS IN BIOINFORMATICS. page 1 of 13 doi:10.1093/bib/bbs046

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

RNA-Seq Workflow

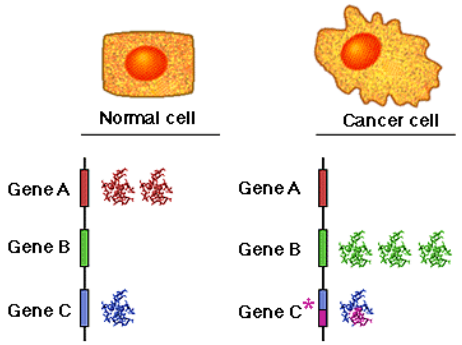


GENOMICS CORELEUVEN

Differential Expression

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

GENOMICS CORELEUVEN

Read Count Distribution Assumption

Method		Read Count Distribution Assumption
EdgeR		Negative binomial distribution
DESeq		Negative binomial distribution
Cuffdiff2 (CuffLinks)		Beta negative binomial distribution

Package Selection

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

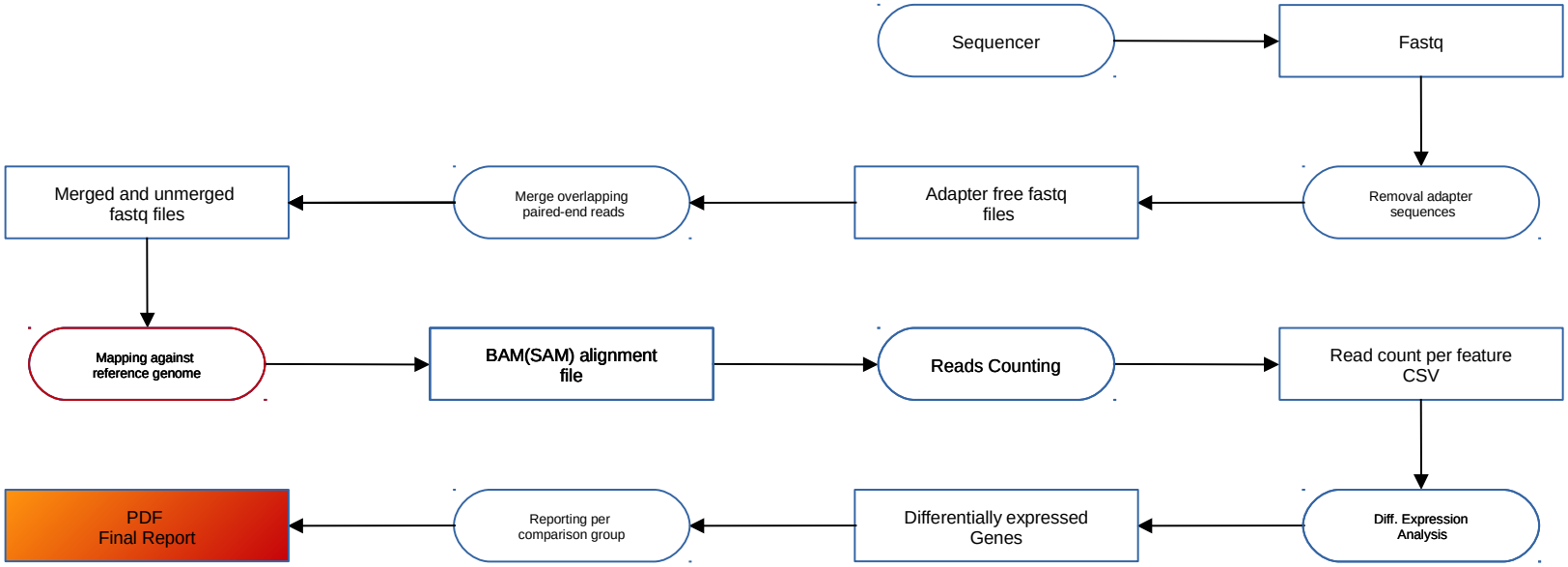
Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]

Differential Expression Table (DeSeq)

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

GENOMICS CORELEUVEN

RNA-Seq Workflow



GENOMICS CORELEUVEN

Thanks!