eurofins Genomics

GATC Biotech AG, Jakob-Stadler-Platz 7, 78467 Konstanz

Data Analysis Report: Expression Analysis v3.3

Project / Study: GATC-Demo-Human

Date: February 27, 2018

Table of Contents

1 Samples **1**

2 Reference **1**

3 Analysis Summary **2**

3.1 Workflow . . . 2

3.2 Expression Analysis . . . 3

3.3 Variant Analysis . . . 3

4 Results **4**

4.1 Read Statistics . . . 4

4.2 Genome / transcriptome alignments . . . 6

4.3 Differential gene expression . . . 6

4.3.1 Sample wise expression (FPKM) tables . . . 6

4.3.2 Pair-wise (control vs. case) differential expression (fold change) tables . . . 6

4.3.3 Quality Metrics . . . 7

4.3.4 Scatter plot(s) . . . 9

4.3.5 Volcano plot(s) . . . 11

4.4 Single Nucleotide Variants and InDel Report . . . 14

5 Deliverables **16**

6 Formats **16**

7 Software Tools **17**

8 Tables **18**

9 FAQ **25**

Bibliography **27**

1 Samples

Table 1: Analysed samples (SE = single end, PE = paired end).

Sample	Read Type	File Name
sample_1	PE	GATC-Demo-Human_sample_1_lib00001_1.fastq GATC-Demo-Human_sample_1_lib00001_2.fastq
sample_2	PE	GATC-Demo-Human_sample_2_lib00001_1.fastq GATC-Demo-Human_sample_2_lib00001_2.fastq
sample_3	PE	GATC-Demo-Human_sample_3_lib00001_1.fastq GATC-Demo-Human_sample_3_lib00001_2.fastq
sample_4	PE	GATC-Demo-Human_sample_4_lib00001_1.fastq GATC-Demo-Human_sample_4_lib00001_2.fastq
sample_5	PE	GATC-Demo-Human_sample_5_lib00001_1.fastq GATC-Demo-Human_sample_5_lib00001_2.fastq
sample_6	PE	GATC-Demo-Human_sample_6_lib00001_1.fastq GATC-Demo-Human_sample_6_lib00001_2.fastq

2 Reference

ORGANISM: Human
GENOME: hg19 / GRC37, UCSC
ANNOTATIONS: Gencode v19, Ensembl 75

3 Analysis Summary

3.1 Workflow

Schematic diagram of the data analysis.

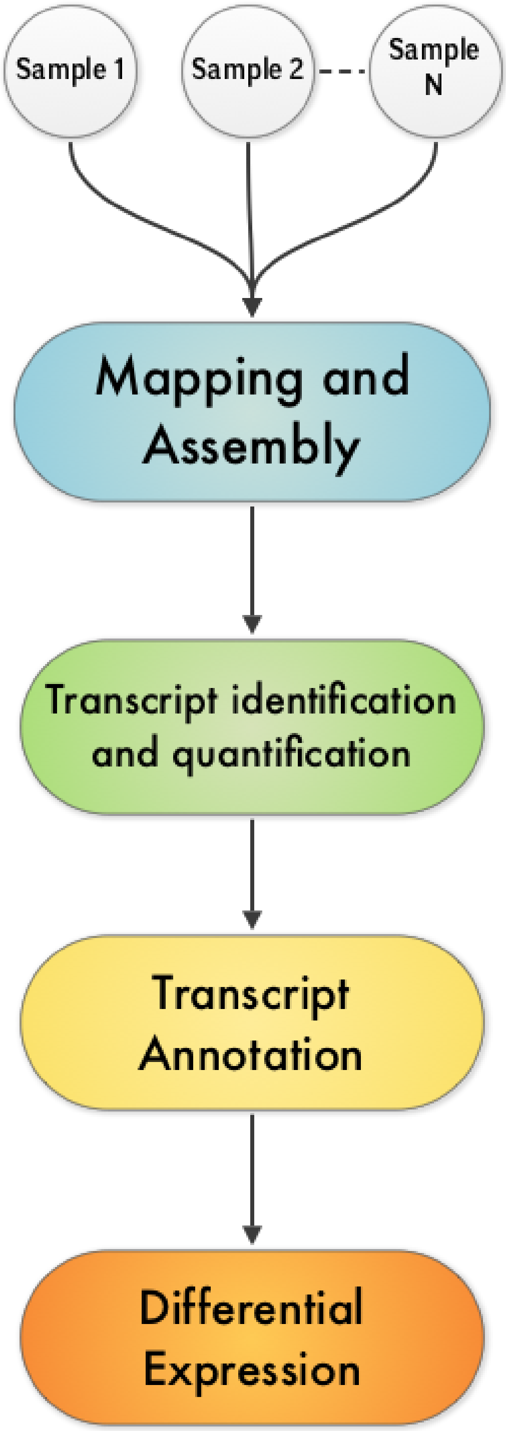

Figure 1: RNA-Seq Workflow

3.2 Expression Analysis

The RNA-Seq reads are aligned to the reference genome or reference transcriptome using Bowtie generating genome / transcriptome alignments. TopHat identifies the potential exon-exon splice junctions of the initial alignment. Then Cufflinks identifies and quantifies the transcripts from the preprocessed RNA-Seq alignment-assembly. After this, Cuffmerge merges the identified transcript pieces to full length transcripts and annotates the transcripts based on the given annotations. Finally, merged transcripts from two or more samples / conditions are compared using Cuffdiff to determine the differential expression levels at transcript and gene level including a measure of significance between samples / conditions.

More information about the tools can be found here [1].

3.3 Variant Analysis

The SNP and InDel calling is done using GATK's Haplotype Caller [2, 3].
Variants detected are annotated based on their gene context using snpEff. The available annotations and their description is described in the tables 17 and 18. Several metrics, that are used to evalutate the quality of a variant, are annotated using GATK's VariantAnnotator module.
Customised flters are applied to the variants to flter false positive variants using GATK's VariantFilteration module. Filters used are described in tables 20 and 21.

Please note the variants reported are NOT VALIDATED and provided as it is reported from the programs mentioned above. Therefore it is highly recommended to inspect the variants thoroughly and validate using alternative methods.

4 Results

4.1 Read Statistics

The total amount of raw sequence data and the results of the quality filtering is collected and reported in the following table. Single Reads are reads without mates (discarded poor quality mate reads). They are not included in further analysis.

The following table contains the number of reads mapped to the reference genome / transcriptome for each of the samples in the experiment. The accuracy of the reference (genome / transcriptome) and better quality of mapped reads lead to a higher percentage of reads mapped to the reference.

Table 2: Quality control statistics per sample

Sample	Total Reads	Discarded Reads	Clean Reads (single)	Clean Reads
sample_1	40,198,668	2,223,412 (5.5 %)	1,551,896 (3.9 %)	36,423,360 (90.6 %)
sample_2	41,459,392	2,301,151 (5.6 %)	1,838,601 (4.4 %)	37,319,640 (90.0 %)
sample_3	40,473,130	3,082,220 (7.6 %)	2,502,904 (6.2 %)	34,888,006 (86.2 %)
sample_4	40,024,310	2,549,136 (6.4 %)	1,934,196 (4.8 %)	35,540,978 (88.8 %)
sample_5	39,648,452	2,387,662 (6.0 %)	1,303,630 (3.3 %)	35,957,160 (90.7 %)
sample_6	36,756,522	8,377,829 (22.8 %)	7,154,137 (19.5 %)	21,224,556 (57.7 %)

Table 3: Mapped read statistics observed per sample

Sample	QC Passed Reads	Mapped Reads	% Mapped
sample_1	36,423,360	34,658,462	95.2
sample_2	37,319,640	36,400,504	97.5
sample_3	34,888,006	33,580,790	96.3
sample_4	35,540,978	34,798,857	97.9
sample_5	35,957,160	34,906,948	97.1
sample_6	21,224,556	20,541,009	96.8

4.2 Genome / transcriptome alignments

The alignments generated from mapping and assembling to the genome / transcriptome reference (see chapter 2) is provided as binary SAM (BAM) format. They can be easily visualized and the alignment can be inspected at gene level using the viewers mentioned in chapter 5.

4.3 Differential gene expression

Operating on the RNA-Seq alignments and Cufflinks processing, Cuffdiff tracks the mapped reads and determines the fragment per kilobase per million mapped reads (FPKM) for each transcript in all the samples. Primary transcripts and gene FPKMs are then computed by adding up the FPKMs of each primary transcript group or gene group. The results can be found in the files listed below.

4.3.1 Sample wise expression (FPKM) tables

For each sample, the genes are listed with the expression values (FPKM) in a tab separated text file. Additionally, combined expression (FPKM) tables are generated by merging all the samples into one table, which may be used for performing comparative analyses. The structure and description of the tables are listed in table 12
a. Sample.FPKM.expression_table.tsv
b. genes.FPKM.combined_expression_table.tsv

4.3.2 Pair-wise (control vs. case) differential expression (fold change) tables

For each pair of samples (control vs. case), the differential expression values such as fold change and p-value are computed at gene level and are listed in a tab separated text file. The genes which are identified as significant by the program are reported in a separate table. The structure and description of the tables are being detailed in table 11.
a. SampleA_SampleB.genes.FPKM.table.tsv
b. SampleA_SampleB.SIGNIFICANT.gene_expression_table.tsv

4.3.3 Quality Metrics

For inspecting the quality of RNA-Seq data, the 100 most abundant genes are taken from all the samples and heatmaps are generated to observe the relation between samples/conditions.
File: top_genes_expressed_fpkm_heatmap.png.

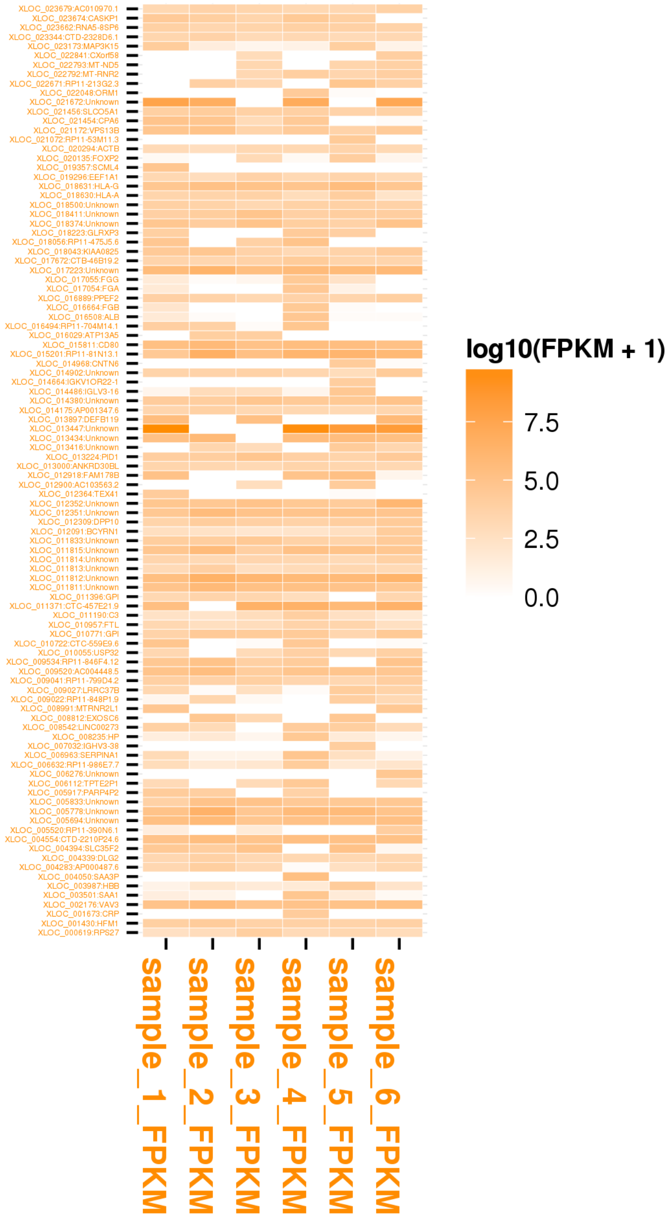

Figure 2: Heat map of top 100 gene(s)

Table 4: List of top 100 genes (*listing first 20 entries, file: top_genes_expressed_fpkm_table.tsv*)

gene_short_name	sample_1_FPKM	sample_2_FPKM	sample_3_FPKM	sample_4_FPKM
RPS27	343	771	10,371	1,149
HFM1	8,662	17,847	5,303	6,847
CRP	0	0	0	17,327
VAV3	102,584	379,952	70,531	106,333
SAA1	21	7	0	53,235
HBB	8	99	41	52
SAA3P	0	0	0	196,492
AP000487.6	2,065	10,662	1,620	1
DLG2	2,132	6,756	1,689	1,923
SLC35F2	31,816	12,808	88,647	0
CTD-2210P24.6	102,584	308,432	92,804	217,729
RP11-390N6.1	26	0	45	0
-	200,334	660,641	58,774	148,054
-	110,561	2,747,250	46,677	506,486
-	6,438	103,817	101,086	31,303
PARP4P2	18,548	10,081	0	5,739
TPTE2P1	1,112	0	963	28,632
-	0	0	0	0
RP11-986E7.7	694	48	92	16,116
SERPINA1	1,065	7	7	64,648

Table 5: List of top 100 genes (*listing first 20 entries, file: top_genes_expressed_fpkm_table.tsv*)

gene_short_name	sample_5_FPKM	sample_6_FPKM
RPS27	2,170	806
HFM1	4,222	6,929
CRP	0	0
VAV3	111,256	161,722
SAA1	42	2
HBB	18,550	151
SAA3P	0	0
AP000487.6	380	2,174
DLG2	1,860	2,377
SLC35F2	119,961	3
CTD-2210P24.6	147,415	138,055
RP11-390N6.1	0	11,456
-	115,249	176,401
-	510,082	95,986
-	31,915	43,115
PARP4P2	0	0
TPTE2P1	0	1,501
-	0	38,237
RP11-986E7.7	62	133
SERPINA1	385	8

4.3.4 Scatter plot(s)

Scatter plots highlight the general similarities and specific outliers between the conditions in the RNA-Seq experiment. They are generated from the expression data for genes using the cummeRbund package. Scatter plots can be used for inspecting overall quality of RNA-Seq data.
File(s): SampleA_SampleB_genes_scatterplot.png.

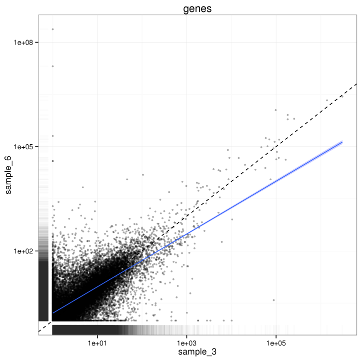

Figure 3: Scatter plot for sample_3 versus sample_6

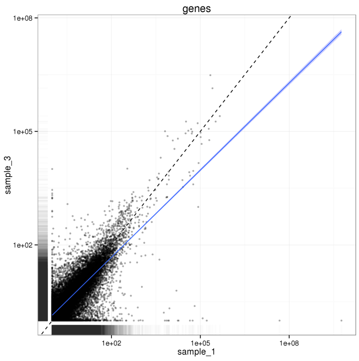

Figure 4: Scatter plot for sample_1 versus sample_3

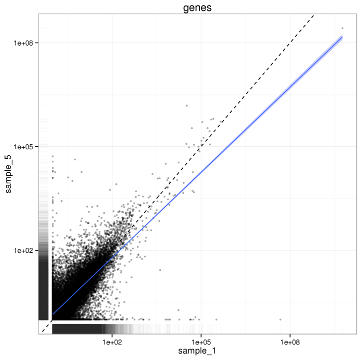

Figure 5: Scatter plot for sample_1 versus sample_5

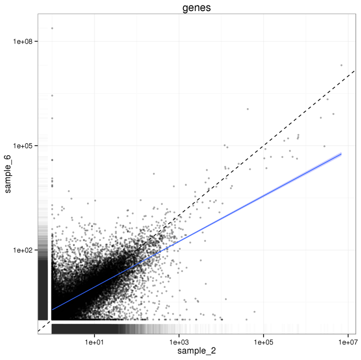

Figure 6: Scatter plot for sample_2 versus sample_6

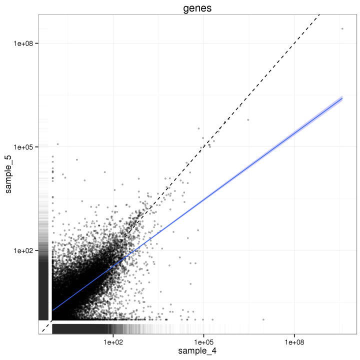

Figure 7: Scatter plot for sample_4 versus sample_5

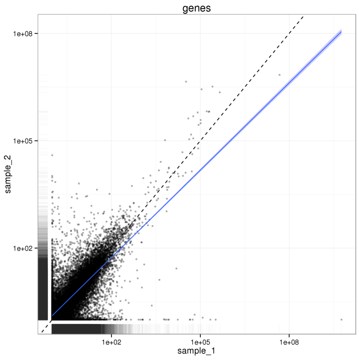

Figure 8: Scatter plot for sample_1 versus sample_2

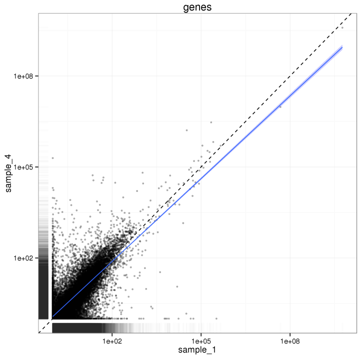

Figure 9: Scatter plot for sample_1 versus sample_4

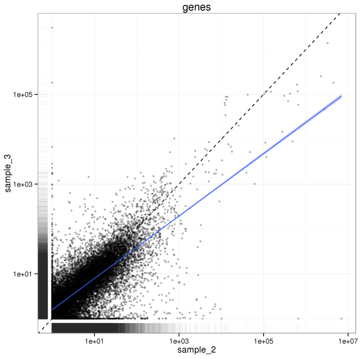

Figure 10: Scatter plot for sample_2 versus sample_3

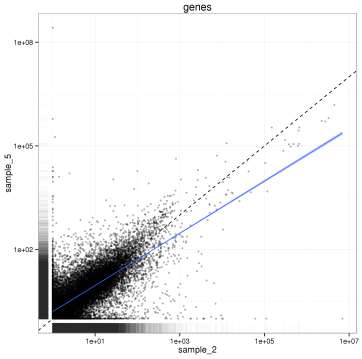

Figure 11: Scatter plot for sample_2 versus sample_5

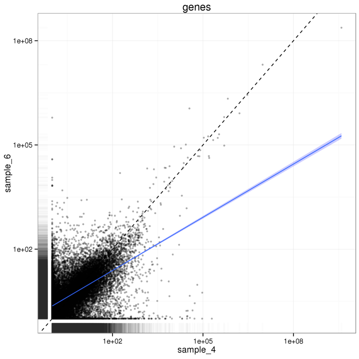

Figure 12: Scatter plot for sample_4 versus sample_6

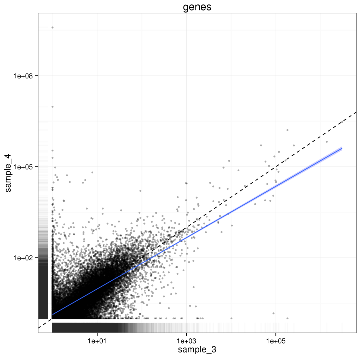

Figure 13: Scatter plot for sample_3 versus sample_4

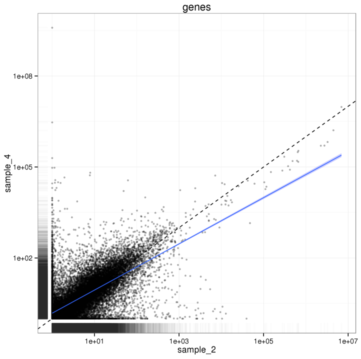

Figure 14: Scatter plot for sample_2 versus sample_4

4.3.5 Volcano plot(s)

Volcano plots highlight the genes that significantly differ between the conditions tested based on the fold change and test statistics performed on the RNA-Seq data between conditions. They are generated based on expression data of genes using the cummeRbund package. Volcano plots can be used for displaying the relationship between conditions at gene expression level.
File(s): SampleA_SampleB_genes_foldchange.png.

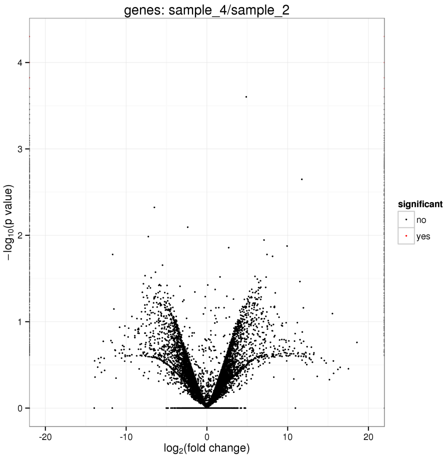

Figure 15: Volcano plot for sample_2 versus sample_4

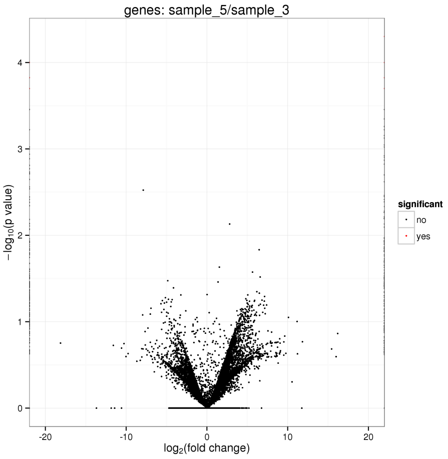

Figure 16: Volcano plot for sample_3 versus sample_5

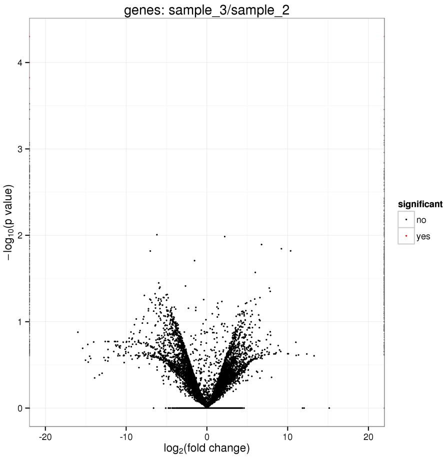

Figure 17: Volcano plot for sample_2 versus sample_3

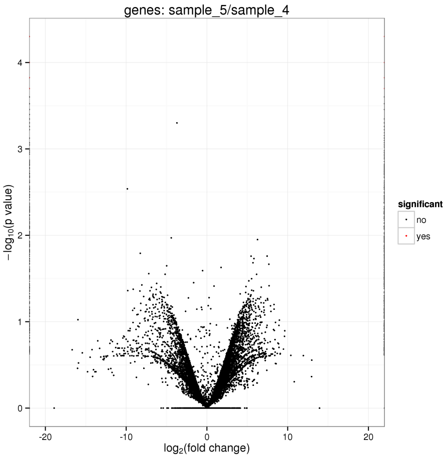

Figure 18: Volcano plot for sample_4 versus sample_5

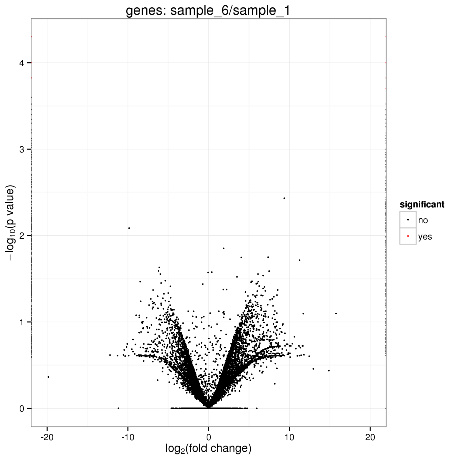

Figure 19: Volcano plot for sample_1 versus sample_6

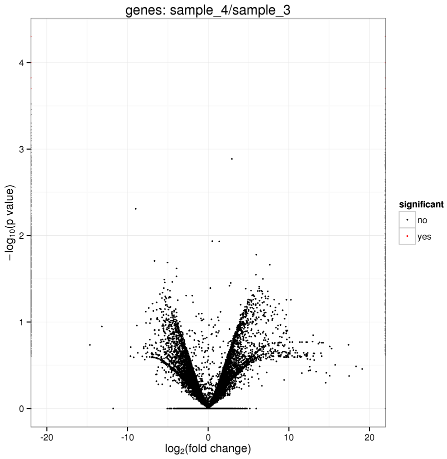

Figure 20: Volcano plot for sample_3 versus sample_4

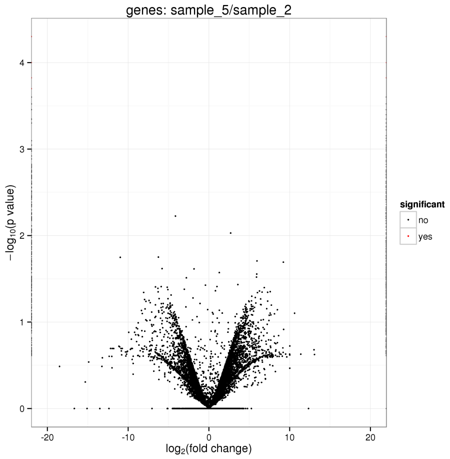

Figure 21: Volcano plot for sample_2 versus sample_5

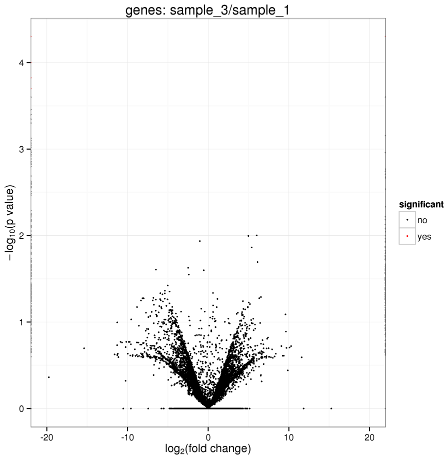

Figure 22: Volcano plot for sample_1 versus sample_3

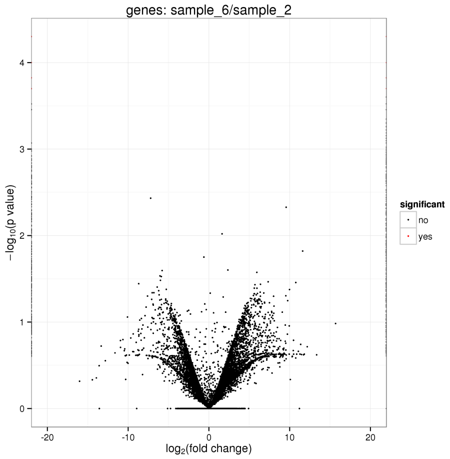

Figure 23: Volcano plot for sample_2 versus sample_6

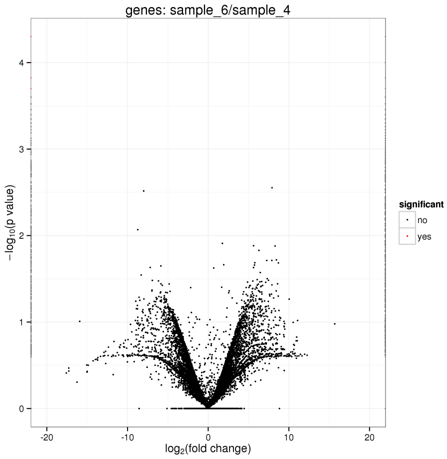

Figure 24: Volcano plot for sample_4 versus sample_6

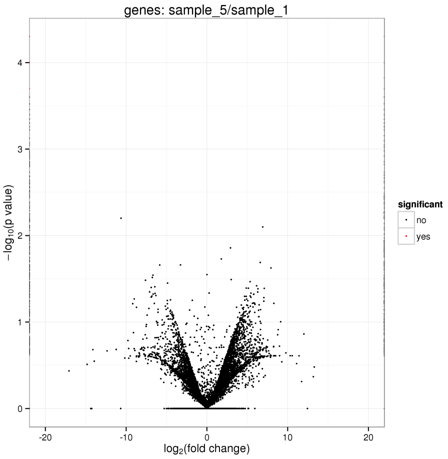

Figure 25: Volcano plot for sample_1 versus sample_5

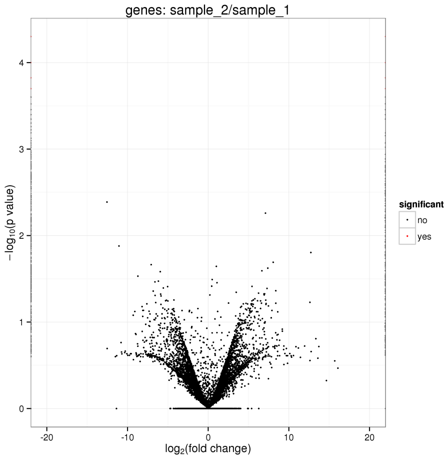

Figure 26: Volcano plot for sample_1 versus sample_2

4.4 Single Nucleotide Variants and InDel Report

The complete list of variants, stratified in single or few nucleotide(s) variants and insertion/deletion (InDel) variants is contained in the delivery package (see chapter 5) in the corresponding VCF and TSV files. The TSV format is described in tables 13 and 14 (fixed fields) and tables 15 and 16 (samplewise fields). The variants (SNV and InDels) detected are summarized in the following table(s).

Table 6: Variant Statistics for sample_1,sample_2,sample_3,sample_4,sample_5

Variant Type	Feature	sample_1	sample_2	sample_3	sample_4	sample_5
ALL [1]	TOTAL	33104	30418	53068	22031	31064
SNV	TOTAL	30308	28026	48849	20325	28647
	KNOWN	27584	26177	44063	18535	25416
	UNKNOWN	2724	1849	4786	1790	3231
	MISSENSE	2852	2853	4064	2232	2870
	NONSENSE	17	13	15	7	16
	SILENT	3990	3979	5650	2903	4016
	NONE	21685	19451	35804	13929	20041
	PASSED	5065	4974	8736	3980	5089
	FAILED	25243	23052	40113	16345	23558
	PASSED KNOWN	4764	4743	8224	3688	4601
	PASSED UNKNOWN	301	231	512	292	488
	PASSED MISSENSE	583	701	967	522	620
	PASSED NONSENSE	4	3	3	1	5
	PASSED SILENT	918	1077	1744	723	951
	PASSED NONE	3398	3017	5756	2529	3119
INDEL	TOTAL	2796	2392	4219	1706	2417
	INS TOTAL	1658	1296	2456	900	1276
	DEL TOTAL	1138	1096	1763	806	1141
	KNOWN	1917	1766	2827	1290	1833
	UNKNOWN	879	626	1392	416	584
	INS MAX SIZE	17	17	16	11	11
	DEL MAX SIZE	107	126	169	107	130
	PASSED	2527	2212	3755	1572	2235
	FAILED	269	180	464	134	182
	PASSED KNOWN	1884	1734	2760	1261	1801
	PASSED UNKNOWN	643	478	995	311	434

[1] Excluding complex sites (i.e. multiallelic calls).

Table 7: Variant Statistics for sample_6

Variant Type	Feature	sample_6
ALL [2]	TOTAL	31601
SNV	TOTAL	29258
	KNOWN	26538
	UNKNOWN	2720
	MISSENSE	2261
	NONSENSE	11
	SILENT	3455
	NONE	21269
	PASSED	3940
	FAILED	25318
	PASSED KNOWN	3722
	PASSED UNKNOWN	218
	PASSED MISSENSE	360
	PASSED NONSENSE	2
	PASSED SILENT	667
	PASSED NONE	2775
INDEL	TOTAL	2343
	INS TOTAL	1163
	DEL TOTAL	1180
	KNOWN	1819
	UNKNOWN	524
	INS MAX SIZE	11
	DEL MAX SIZE	107
	PASSED	2214
	FAILED	129
	PASSED KNOWN	1793
	PASSED UNKNOWN	421

[2] Excluding complex sites (i.e. multiallelic calls).

5 Deliverables

Table 8: List of deliverable files, format and recommended programs to access.

File	Format	Program To Open File
Sample.alignment.bam	BAM	IGV, Tablet
Sample.alignment.bam.bai	BAI	None
Sample.unmapped_1.fastq	FASTQ	Text editor
Sample.unmapped_2.fastq	FASTQ	Text editor
Sample.snp.bed	BED	USCS Genome Browser
Sample.indel.bed	BED	USCS Genome Browser
Sample.snp.tsv	TSV	Spreadsheet editor
Sample.indel.tsv	TSV	Spreadsheet editor
Sample.snp.vcf	VCF	Text Editor
Sample.indel.vcf	VCF	Text Editor
genes.FPKM.combined_expression_table.tsv	TSV	Spreadsheet editor
top_genes_expressed_fpkm_table.tsv	TSV	Spreadsheet editor
top_genes_expressed_fpkm_heatmap.png	PNG	Image viewer
SampleA_SampleB.gene_expression_table.tsv	TSV	Spreadsheet editor
SampleA_SampleB.SIGNIFICANT.gene_expression_table.tsv	TSV	Spreadsheet editor
SampleA_SampleB_genes_scatterplot.png	PNG	Image viewer
SampleA_SampleB_genes_foldchange.png	PNG	Image viewer
Expression_Analysis_Report.pdf	PDF	PDF reader

6 Formats

Table 9: References and descriptions of file formats

Format	Description
FASTQ[4]	Text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are encoded with a single ASCII character for brevity.
BAM[5]	Compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments.
TSV	Tab separated table style text file. Can be imported into spreadsheet processing software like MS OFFICE Excel.
PNG	Visual representation in Portable Network Graphics format.
BED	Browser Extensible Data (BED) is a text file compatible with genome browsers.
VCF[6]	Variant Call Format (VCF) is a format to describe and report the variants.

7 Software Tools

Table 10: Name, Version, Reference and Description of relevant programs

Program	Version	Description
Bowtie[7]	2.2.9	Bowtie is a ultrafast, memory-efficient short read aligner. It is based on Burrows-Wheeler transform algorithm.
CummeRbund[8]	2.0.0	CummeRbund is an R package used for post processing Cufflinks-Cuffdiff results to generate various plots.
GATK[2, 3]	3.7	GATK is a java-based command-line toolkit that process SAM / BAM / VCF files.
Picard[9]	1.131	Picard is a java-based command-line utilities for processing SAM / BAM files.
R[10]	2.15.3	R is a programming language and environment for statistical computing.
SAMTools[11]	0.1.18	SAMtools provide various utilities for manipulating alignments in the SAM format.
TopHat[12]	2.0.14	TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to the reference genome / transcriptome using the ultra-high-throughput short read aligner Bowtie, and analyses the mapping results to identify splice junctions between exons.
Trimmomatic[13]	0.33	Trimmomatic performs a variety of useful trimming tasks for Illumina paired-end and single-end data.
sambamba[14]	0.6.6	Sambamba is a high performance modern robust and fast tool (and library), for working with SAM and BAM files.
snpEff[15]	4.3	snpEff is a variant annotation and effect prediction tool.

8 Tables

Table 11: Structure and description of differential expression (fold change) table. Columns 3 and 4 may not be present if the analysis was done with a custom reference.

No.	Name	Example	Description
1	test_id	XLOC_000001	A unique identifier describing the transcript, gene, primary transcript, or CDS being tested.
2	gene	Lypla1	The gene_name(s) or gene_id(s) being tested.
3	refseq_id	NM_008866	Nearest RefSeq ID of the identified transcript based on the location on genome and the corresponding annotation features.
4	alternative_refseq_ids	-	List of alternative RefSeq IDs sharing the same location and features.
5	locus	chr1:4797771-4835363	Genomic coordinates for easy browsing to the genes or transcripts being tested.
6	sample_1	Liver	Label (or number if no labels provided) of the first sample being tested.
7	sample_2	Brain	Label (or number if no labels provided) of the second sample being tested.
8	status	NOTEST	Can be one of OK (test successful), NOTEST (not enough alignments for testing), LOWDATA (too complex or shallowly sequenced), HIDATA (too many fragments in locus), or FAIL, when an ill-conditioned covariance matrix or other numerical exception prevents testing.
9	value_1	8.01089	FPKM of the gene in sample 1.
10	value_2	8.551545	FPKM of the gene in sample 2.
11	log2(fold_change)	0.06531	The (base 2) log of the fold change y/x.
12	test_stat	0.860902	The value of the test statistics used to compute significance of the observed change in FPKM.
13	p_value	0.389292	The uncorrected p-value of the test statistic.
14	q_value	0.985216	The False Discovery Rate (FDR) adjusted p-value of the test statistic.
15	significant	no	Can be either yes or no, depending on whether p is greater than the FDR after Benjamini-Hochberg correction for multiple testing.

Table 12: Structure and description of expression (FPKM) table. Columns 2 and 3 may not be present if the analysis was done with a custom reference.

No.	Name	Example	Description
1	gene_short_name	Lypla1	The gene_short_name(s) associated with the object.
2	refseq_id	NM_008866	Nearest RefSeq ID of the identified transcript based on the location on genome and the corresponding annotation features.
3	altern_refseq_ids	-	List of alternative RefSeq IDs sharing the same location and features.
4	locus	chr1:4797771-4835363	Genomic coordinates for easy browsing to the object.
5	Sample1_FPKM	8.01089	FPKM of the object in sample 1.
6	Sample1_status	OK	Quantification status for the transcript in sample 1. Can be one of OK (deconvolution successful), LOWDATA (too complex or shallowly sequenced), HIDATA (too many fragments in locus), or FAIL, when an ill-conditioned covariance matrix or other numerical exception prevents deconvolution.

Table 13: Examples of fixed fields of the tab delimited variant report table.

CHROMOSOME	POSITION	DBSNP ID	REFERENCE BASE	OBSERVED BASE	QUALITY SCORE	FILTER	COVERAGE
chr3	46399798	rs1799865	T	C	9551.17	PASS	254
chr3	164777677	rs9290259	T	G	9919.08	PASS	306
chr11	55873024	rs2449148	A	G	9104.32	PASS	240
chr12	55945119	rs7313899	A	G	9616.99	PASS	281
chr12	10570965	rs2682495	C	G	9476.45	PASS	278
chr17	66039350	rs4638	A	G	9077.84	PASS	253
chr19	53911973	rs10425136	A	G	9853.53	PASS	252
chr19	55378008	rs3745902	C	T	9066.27	PASS	297

Table 14: Defintion of fixed fields of the tab delimited variant report table

Name	Meaning
CHROMOSOME	Name of reference contig or chromosome where the variant occurs
POSITION	Position of reference contig or chromosome where the variant occurs
DBSNP ID	The dbSNP rs identifier of the SNP based on the contig or chromosome position of the call. If there is an entry in the dbSNP then the respective rs id will be displayed. Dot ('.') indicates no entry in the dbSNP.
REFERENCE BASE	The reference base at the variant site
OBSERVED BASE	Alternative (observed) base in the samples in general [VARIANT]
QUALITY SCORE	The Phred scaled probability of OBSERVED BASE is correct at this site given sequencing data. The value is computed based on error models designed by Broad Institute. Since the Phred scale is -10 * log(1-p), a value of 10 indicates a 1 in 10 chance of error, while a 100 indicates a 1 in 10^10 chance. The higher the value the more accurate is the variant call.
FILTER	In addition to quality score, several filters can be defined to filter the SNPs by considering factors other than quality score alone. For e.g., SNP with low quality score threshold of < 30 could be tagged as LowQual SNPs and the ones which pass this filter will be tagged as PASS. More than one filter can be defined and applied to the variant calls. Default filters are SnpCluster (more than 2 SNPs found in cluster of size=10), LowQual (SNP with quality score < 30), LowCov (SNP with coverage < 20), Mask (SNP is at least 10 base near to indel location) and HardToValidate (Not enough evidence to validate). Variant passing the default filters will be tagged "PASS"
COVERAGE	Sequencing depth or coverage at the variant position. More accurate is to see the SAMPLE:COVERAGE

Table 15: Example of sample wise fields.

SAMPLE: GENOTYPE	SAMPLE:GQ	SAMPLE: ALLELE DEPTH	SAMPLE: ALELLE BALANCE	SAMPLE: COVERAGE
1/1	96.11	0.68		65
0/1	99	86.26	0.77	110
0/0	48.35	72.50	.	77

Table 16: Definition of sample wise fields.

Name	Meaning
SAMPLE:GENOTYPE	The genotype of the sample. For a diploid genome, the GENOTYPE indicates the two alleles carried by the sample, encoded by a 0 for the REFERENCE allele, 1 for the first ALTERNATIVE (OBSERVED) allele. Possible GENOTYPEs are 0/0 (the sample is homozygous to reference), 0/1 (the sample is heterozygous, carrying 1 copy of each of the REFERENCE and ALTERNATIVE alleles) and 1/1 (the sample is homozygous alternate i.e., completely opposite to the REFERENCE)
SAMPLE:GQ	The phred scaled genotype quality.
SAMPLE:ALLELE DEPTH	The allele depth, one for each REFERENCE and ALTERNATIVE (OBSERVED), is the count of all reads that carried with them the respective alleles. The read counts also include the poor mapping quality reads, unlike the COVERAGE counts.
SAMPLE:ALELLE BALANCE	Allele balance is a ratio of the REFERENCE bases to the total bases observed in the give position. This applies for only heterozygous calls and value ranges from > 0.0 to < 1.0
SAMPLE:COVERAGE	The total depth of the reads that passed the internal quality control metrics (for eg., mapping quality >17) from all reads present at this site.

Table 17: Examples of genomic annotations as produced by snpEff.

AMINO ACID CHANGE	CODON CHANGE	EFFECT	EXON ID	FUNCTIONAL CLASS	GENE NAME	IMPACT	TRANSCRIPT
R44S	agG/agT	NON SYNONYMOUS CODING	exon_1_935072_935552	MISSENSE	HES4	MODERATE	ENST00000428771
L615	Ctg/Ttg	SYNONYMOUS CODING	exon_1_881553_881666	SILENT	NOC2L	LOW	ENST00000327044
		FRAME_SHIFT	exon_1_877939_878438	NONE	SAMD11	HIGH	ENST00000342066
P605PG	cca/ccCGGa	CODON CHANGE PLUS CODON INSERTION	exon_1_35653574_35653691	NONE	SFPQ	MODERATE	ENST00000357214
-409G	-/GGG	CODON INSERTION	exon_1_1683910_1684499	NONE	NADK	MODERATE	ENST00000342348
Y205*	taT/taG	STOP GAINED	exon_1_25167264_25170815	NONSENSE	CLIC4	HIGH	ENST00000374379
154	tAa/tGa	SYNONYMOUS STOP	exon_4_41621205_41621953	SILENT	LIMCH1	LOW	ENST00000509638
		INTERGENIC	NONE			MODIFIER	
		UPSTREAM	NONE		AL669831.1	MODIFIER	ENST00000358533
		UTR_5_PRIME	exon_1_948803_948956	NONE	ISG15	MODIFIER	ENST00000379389
		SPLICE SITE ACCEPTOR		NONE	RP11-34P13.2	HIGH	ENST00000538476
		SPLICE SITE DONOR		NONE	SAMD11	HIGH	ENST00000342066

Table 18: Definition of genomic annotations as produced by snpEff.

Name	Meaning
AMINO ACID CHANGE	The exact position and the change of the amino acid.
CODON CHANGE	The change of the nucleotide within the context of the Codon.
EFFECT	The predicted effect the change implies.
EXON ID	The Exon Id the variant belongs to.
FUNCTIONAL CLASS	Functional class of the SNP - silent (synonoymous), missense (non-synonymous), nonsense (stop-gaining),readthrough (stop-loss) and NA (unclassified)
GENE NAME	The gene entry associated with the location of the variant call. If present, gene name will be displayed. ifnot, ”NA” will be displayed
IMPACT	Effect impact. Can be one of High, Moderate, Low, Modifier.
TRANSCRIPT ID	The transcript Id.

Table 19: Impact, Description and Examples of Effects as reported by snpEff.

Impact	Effects	Description	Examples
High	SPLICE_SITE_ACCEPTOR	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).	
	SPLICE_SITE_DONOR	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).	
	START_LOST	Variant causes start codon to be mutated into a non-start codon.	aTg/aGg, M/R
	EXON_DELETED	A deletion removes the whole exon.	
	FRAME_SHIFT	Insertion or deletion causes a frame shift	An indel size is not multple of 3
	STOP_GAINED	Variant causes a STOP codon	Cag/Tag, Q/*
	STOP_LOST	Variant causes stop codon to be mutated into a non-stop codon	Tga/Cga, */R
Moderate	NON_SYNONYMOUS _CODING	Variant causes a codon that produces a different amino acid	Tgg/Cgg, W/R
	CODON_CHANGE	One or many codons are changed	An MNP of size multiple of 3
	CODON_INSERTION	One or many codons are inserted	An insert multiple of three in a codon boundary
	CODON_CHANGE_PLUS _CODON_INSERTION	One codon is changed and one or many codons are inserted	An insert of size multiple of three, not at codon boundary
	CODON_DELETION	One or many codons are deleted	A deletion multiple of three at codon boundary
	CODON_CHANGE_PLUS _CODON_DELETION	One codon is changed and one or more codons are deleted	A deletion of size multiple of three, not at codon boundary
	UTR_5_DELETED	The variant deletes and exon which is in the 5'UTR of the transcript	
	UTR_3_DELETED	The variant deletes and exon which is in the 3'UTR of the transcript	
Low	SYNONYMOUS_START	Variant causes start codon to be mutated into another start codon.	Ttg/Ctg, L/L (TTG and CTG can be START codons)
	NON_SYNONYMOUS_START		
	START_GAINED	A variant in 5'UTR region produces a three base sequence that can be a START codon.	
	SYNONYMOUS_CODING	Variant causes a codon that produces the same amino acid	Ttg/Ctg, L/L
	SYNONYMOUS_STOP	Variant causes stop codon to be mutated into another stop codon.	taA/taG, */*
	NON_SYNONYMOUS_STOP		
Modifier	UTR_5_PRIME	Variant hits 5'UTR region	
	UTR_3_PRIME	Variant hits 3'UTR region	
	REGULATION		
	UPSTREAM	Upstream of a gene (default length: 5K bases)	
	DOWNSTREAM	Downstream of a gene (default length: 5K bases)	
	GENE	The variant hits a gene.	
	TRANSCRIPT	The variant hits a transcript.	
	EXON	The vairant hits an exon.	
	INTRON_CONSERVED	The variant is in a highly conserved intronic region	
	INTRON	Variant hist and intron. Technically, hits no exon in the transcript.	
	INTRAGENIC	The variant hits a gene, but no transcripts within the gene	
	INTERGENIC	The variant is in an intergenic region	
	INTERGENIC_CONSERVED	The variant is in a highly conserved intergenic region	
	NONE		
	CHROMOSOME		
	CUSTOM		
	CDS	The variant hits a CDS.	

Table 20: Filters applied for single nucleotide variant sites.

Name	Expression	Description
LowCovFilter	≤ 20	Depth of Coverage.
QDFilter	<2.0	Quality by read depth.
MQFilter	<-12.5	Root Mean Square of the Mapping quality of the reads across all samples.
FSFilter	>60.0	Phred-scaled p-value using Fisher's Exact Test to detect strand bias.
HaplotypeFilter	>13.0	Consistency of the site with two (and only two) segregating haplotypes.
MQFilter	<-12.5	The phred-scaled p-value (u-based z-approximation) from the Mann-Whitney Rank Sum Test for mapping qualities.
ReadPosFilter	<-8.0	The phred-scaled p-value (u-based z-approximation) from the Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with the alternate allele.

Table 21: Filter applied for small Insertion / Deletion variant sites.

Name	Expression	Description
QDFilter	<2.0	Quality by read depth.
ReadPosFilter	<-20.0	The phred-scaled p-value (u-based z-approximation) from the Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with the alternate allele.
FSFilter	>200.0	Phred-scaled p-value using Fisher's Exact Test to detect strand bias.

9 FAQ

Q: What is the difference between FPKM and RPKM?
A: RPKM stands for Reads Per Kilobase of transcript per Million mapped reads. FPKM stands for Fragments Per Kilobase of transcript per Million mapped reads. In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it.

Q: How are the top genes in heat map determined?
A: Top genes are selected based on the reported FPKM values. The 100 most abundant genes are selected from each group of samples and a heat map is drawn. Such heat maps are helpful to give a quick overview about the samples under investigation by highlighting any outliers in the experiments performed.

Q: Why do I find some gene entries in the comparative tables but not in the FPKM tables?
A: The applied statistical model to compute FPKM values takes into account and corrects the final FPKM values based on the distribution of transcripts in the sample. In other words, the FPKM values reported will be corrected for fragment size selection during the library preparation step. So, the shorter transcript fragments will get increased FPKM values because of the fact that the size selection during the library preparation avoids very short fragments being represented in the RNA-Seq data. This compensation was designed to improve accuracy for transcripts that are in the 500bp-1kb range. Until there is a better model for quantifying shorter transcripts, the transcripts which are shorter than 300bp are ignored and not reported in the comparative expression tables. This might cause the missing entries in the comparative table even though they are reported in the sample FPKM table.

Q: How does Cuffdiff 2 test for differentially expressed and regulated genes?
A: To identify a gene or transcript as differentially expressed, Cuffdiff 2 tests the observed log-fold-change in expression against the null hypothesis of no change (i.e. a true log-fold-change of zero). Because measurement error, technical variability, and cross-replicate biological variability might result in an observed log-fold-change that is not zero, Cuffdiff assesses significance using a model of variability in the log-fold-change under the null hypothesis. This model is described in detail in Trapnell and Hendrickson et al. Briefly, Cuffdiff 2 constructs for each condition a table that predicts how much variance there is in the number of reads originating from a gene or transcript. The table is keyed by the average reads across replicates, so to look up the variance for a transcript using the table, Cuffdiff estimates how many reads originated from that transcript, and then queries the table to retrieve the variance for that number of reads. Cuffdiff 2 then accounts for read mapping and assignment uncertainty by simulating probabilistic assignment of the reads mapping to a locus to the splice isoforms for that locus. At the end of the estimation procedure, Cuffdiff 2 obtains an estimate of the number of reads that originated from each gene and transcript, along with variances in those estimates. The read counts are reported along with FPKM values and their variances. Change in expression is reported as the log-fold-change in FPKM and the FPKM variances allow the program to estimate the variance in the log-fold-change itself. Naturally, a gene that has highly variable expression will have a highly variable log-fold-change between two conditions. *(From Cufflinks website)*

Q: How can I open a TSV file in Excel?
A: Start Excel and click File -> Open and select the TSV file you want to open. Next an assistant dialog should show up. Make sure that you select tab as separator. Set the format of all rows without numbers to text. The TSV files use the dot as decimal mark and comma as thousands separator. Make sure that you set both correctly.

Q: My gene names are screwed up after opening a file in Excel. What can I do?
A: This is a common problem. For further information read this publication: *Mistaken Identifiers: Gene name*

errors can be introduced inadvertently when using Excel in bioinformatics [16].

Bibliography

[1] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, May 2010.

[2] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[3] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43:491–498, 2011.

[4] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010.

[5] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[6] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[7] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25–10, March 2009.

[8] L Goff, C. Trapnell, and D. Kelley. http://www.bioconductor.org/packages/release/bioc/html/cummeRbund.html, 2012.

[9] Picard. http://picard.sourceforge.net.

[10] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[11] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[12] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–1111, May 2009.

[13] Marc Lohse, Anthony M. Bolger, Axel Nagel, Alisdair R. Fernie, John E. Lunn, Mark Stitt, and Björn Usadel. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40(W1):W622–W627, July 2012.

[14] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, February 2015.

[15] Pablo Cingolani. ”snpEff: Variant effect prediction”. http://snpeff.sourceforge.net, 2012.

[16] Barry Zeeberg, Joseph Riss, David Kane, Kimberly Bussey, Edward Uchio, W. Marston Linehan, J. Carl Barrett, and John Weinstein. Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*, 5(1):80+, June 2004.

Eurofins Genomics' products, services and applications reach the best quality and safety levels. They are carried out under strict QM and QA systems and comply with the following standards:

ISO 9001	Globally recognised as the standard quality management certification	GLP	The gold standard to conduct non-clinical safety studies
ISO 17025	Accredited analytical excellence	GCP	Pharmacogenomic services for clinical studies
ISO 13485	Oligonucleotides according to medical devices standard	cGMP	Products and testing according to pharma and biotech requirements

Eurofins Genomics • Anzinger Str. 7a • 85560 Ebersberg • Germany