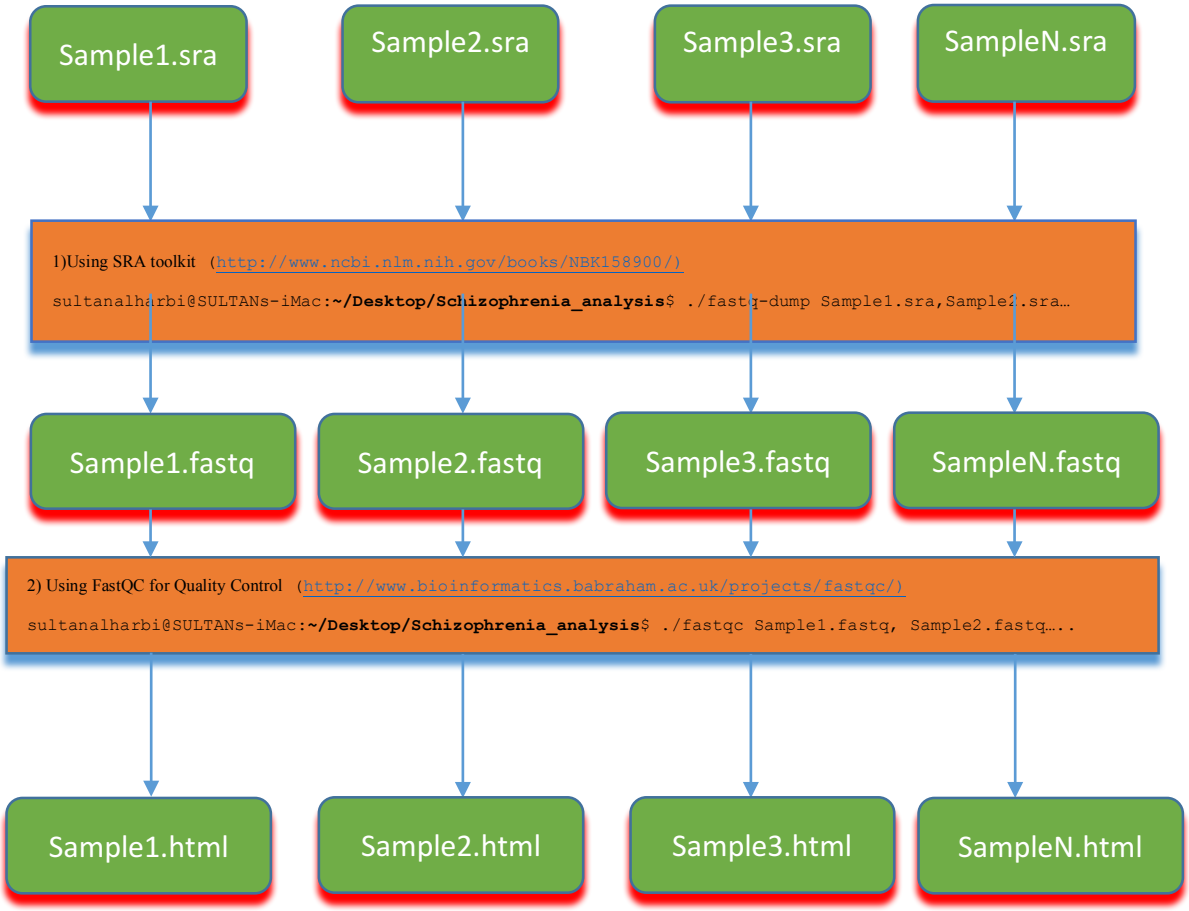
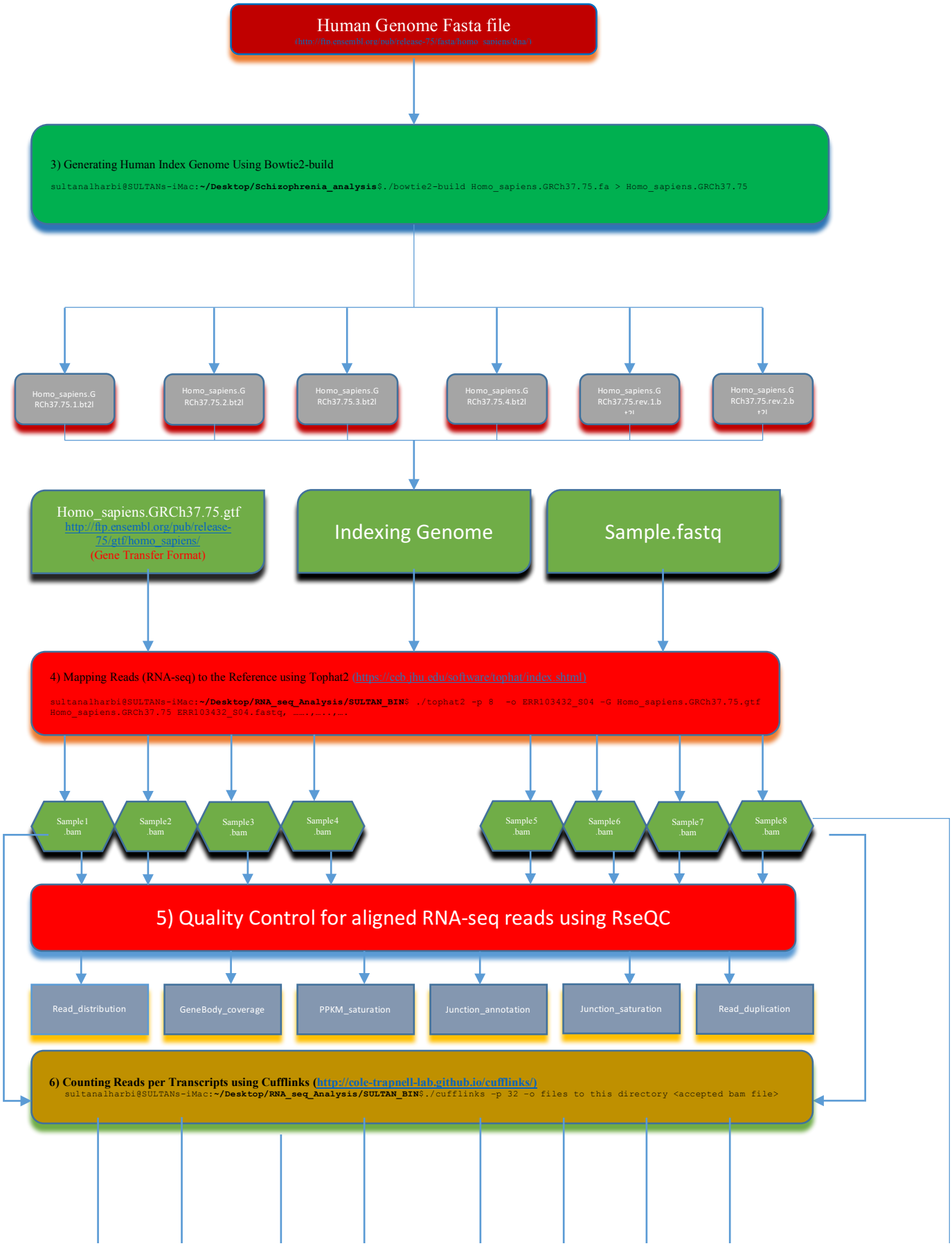
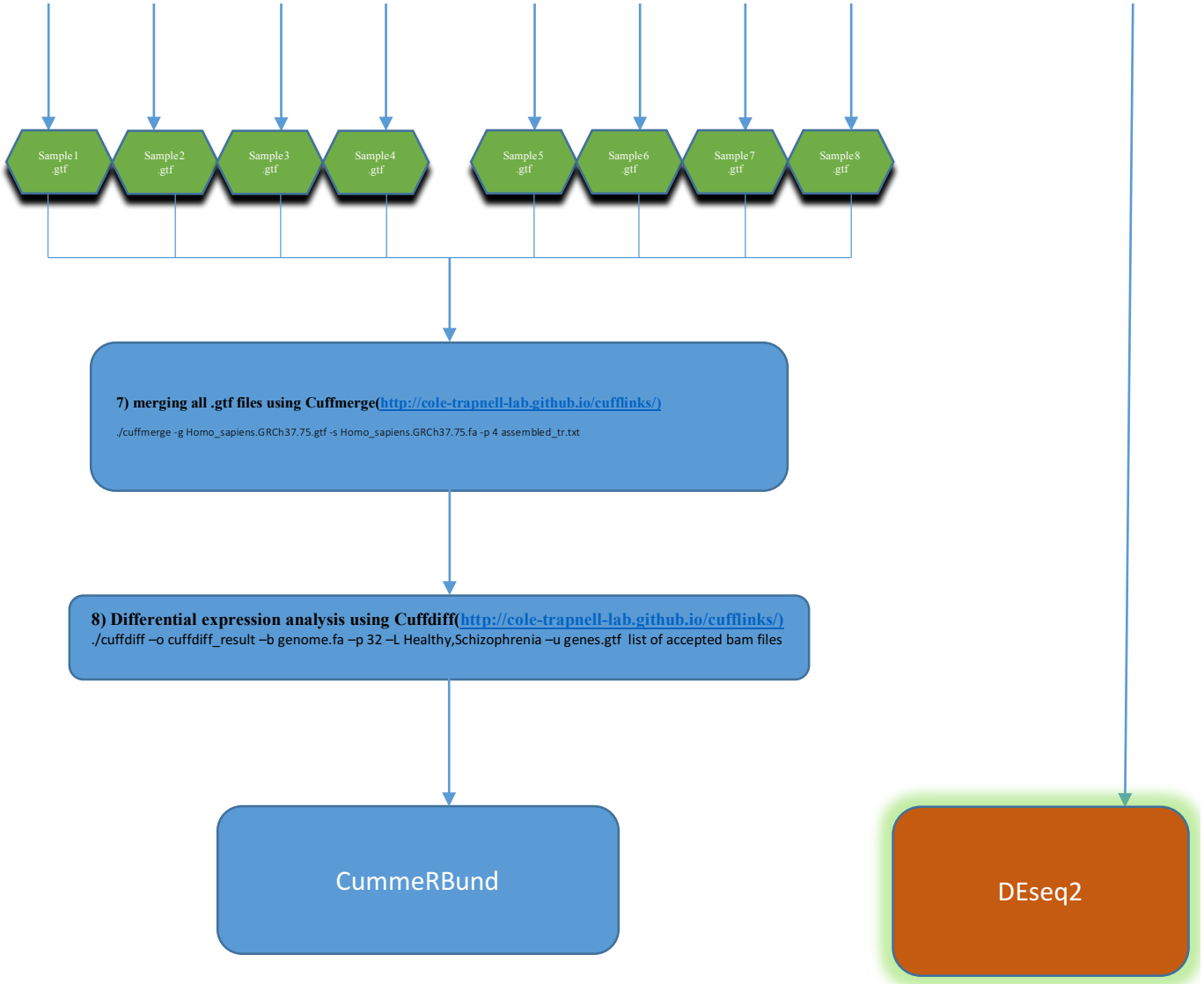

Sample1.sra
Sample2.sra
Sample3.sra
SampleN.sra
1)Using SRA toolkit (http://www.ncbi.nlm.nih.gov/books/NBK158900/)
sultanalharbi@SULTANs-iMac:~/Desktop/Schizophrenia_analysis$ ./fastq-dump Sample1.sra,Sample2.sra…
Sample1.fastq
Sample2.fastq
Sample3.fastq
SampleN.fastq
2) Using FastQC for Quality Control (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
sultanalharbi@SULTANs-iMac:~/Desktop/Schizophrenia_analysis$ ./fastqc Sample1.fastq, Sample2.fastq…..
Sample1.html
Sample2.html
Sample3.html
SampleN.html


Human Genome Fasta file
(http://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/)
3) Generating Human Index Genome Using Bowtie2-build
sultanalharbi@SULTANs-iMac:~/Desktop/Schizophrenia_analysis$./bowtie2-build Homo_sapiens.GRCh37.75.fa > Homo_sapiens.GRCh37.75
Homo_sapiens.GRCh37.75.1.bt2l
Homo_sapiens.GRCh37.75.2.bt2l
Homo_sapiens.GRCh37.75.3.bt2l
Homo_sapiens.GRCh37.75.4.bt2l
Homo_sapiens.GRCh37.75.rev.1.bt2l
Homo_sapiens.GRCh37.75.rev.2.bt2l
Homo_sapiens.GRCh37.75.gtf
http://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/
(Gene Transfer Format)
Indexing Genome
Sample.fastq
4) Mapping Reads (RNA-seq) to the Reference using Tophat2 (https://ccb.jhu.edu/software/tophat/index.shtml)
sultanalharbi@SULTANs-iMac:~/Desktop/RNA_seq_Analysis/SULTAN_BIN$ ./tophat2 -p 8 -o ERR103432_S04 -G Homo_sapiens.GRCh37.75.gtf Homo_sapiens.GRCh37.75 ERR103432_S04.fastq, ……,……,…..
Sample1 .bam
Sample2 .bam
Sample3 .bam
Sample4 .bam
Sample5 .bam
Sample6 .bam
Sample7 .bam
Sample8 .bam
5) Quality Control for aligned RNA-seq reads using RseQC
Read_distribution
GeneBody_coverage
PPKM_saturation
Junction_annotation
Junction_saturation
Read_duplication
6) Counting Reads per Transcripts using Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/)
sultanalharbi@SULTANs-iMac:~/Desktop/RNA_seq_Analysis/SULTAN_BIN$./cufflinks -p 32 -o files to this directory <accepted bam file>


Sample1 .gtf
Sample2 .gtf
Sample3 .gtf
Sample4 .gtf
Sample5 .gtf
Sample6 .gtf
Sample7 .gtf
Sample8 .gtf
7) merging all .gtf files using Cuffmerge(http://cole-trapnell-lab.github.io/cufflinks/)
./cuffmerge -g Homo_sapiens.GRCh37.75.gtf -s Homo_sapiens.GRCh37.75.fa -p 4 assembled_tr.txt
8) Differential expression analysis using Cuffdiff(http://cole-trapnell-lab.github.io/cufflinks/)
./cuffdiff –o cuffdiff_result –b genome.fa –p 32 –L Healthy,Schizophrenia –u genes.gtf list of accepted bam files
CummeRBund
DEseq2

Why do we do RNA-seq analysis?

1- Quantify expression Genome-wide in a single assay.
2- Finding novel genes and splice variants (isoforms of known genes).
3- Compare genes and transcripts expression under two or more conditions.

A) Materials and Software:

- Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)
- Samtools (http://samtools.sourceforge.net)
- Tophat2 (http://ccb.jhu.edu/software/tophat/index.shtml)
- STAR (https://github.com/alexdobin/STAR)
- Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/)
- Rstudio (https://www.rstudio.com)
- CummeRbund package
- DESeq2 package

B) Datasets and Genome reference

- Reference Genome: Homo_Sapiens release (http://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/)

- SRA: ERP001304 (RNA-seq datasets) (http://www.ncbi.nlm.nih.gov/sra/?term=ERP001304)

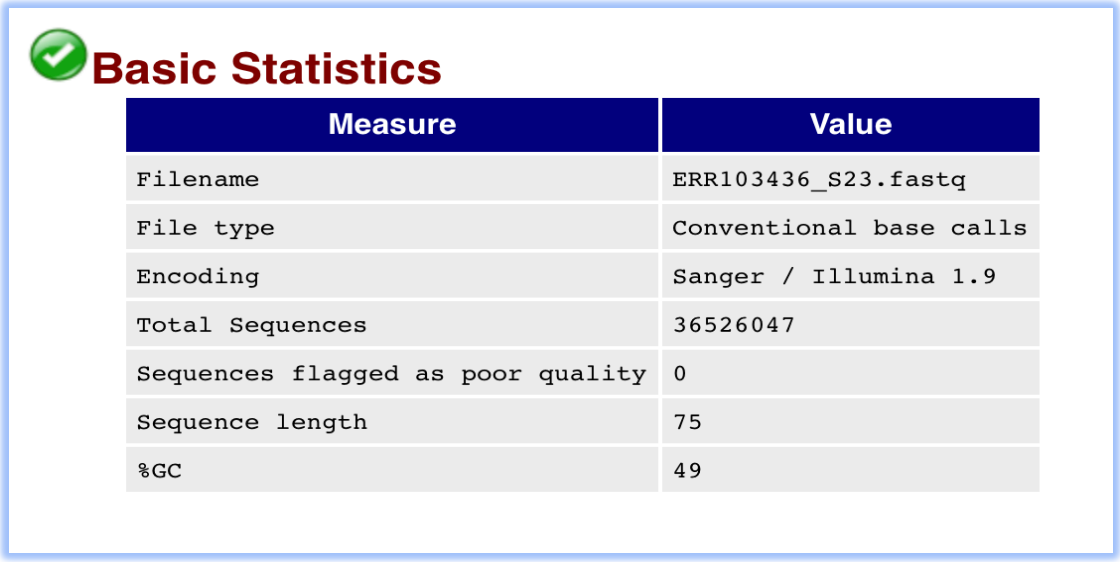
Results

A) Quality Control Results for Health samples

1) file:///Users/sultanalharbi/Desktop/Schizophrenia_analysis/ERR103422_C02_fastqc.html
2) file:///Users/sultanalharbi/Desktop/Schizophrenia_analysis/ERR103423_C04_fastqc.html
3) file:///Users/sultanalharbi/Desktop/Schizophrenia_analysis/ERR103424_C05_fastqc.html
4) file:///Users/sultanalharbi/Desktop/Schizophrenia_analysis/ERR103428_C25_fastqc.html

B) Quality Control Results for Schizophrenia Disease samples

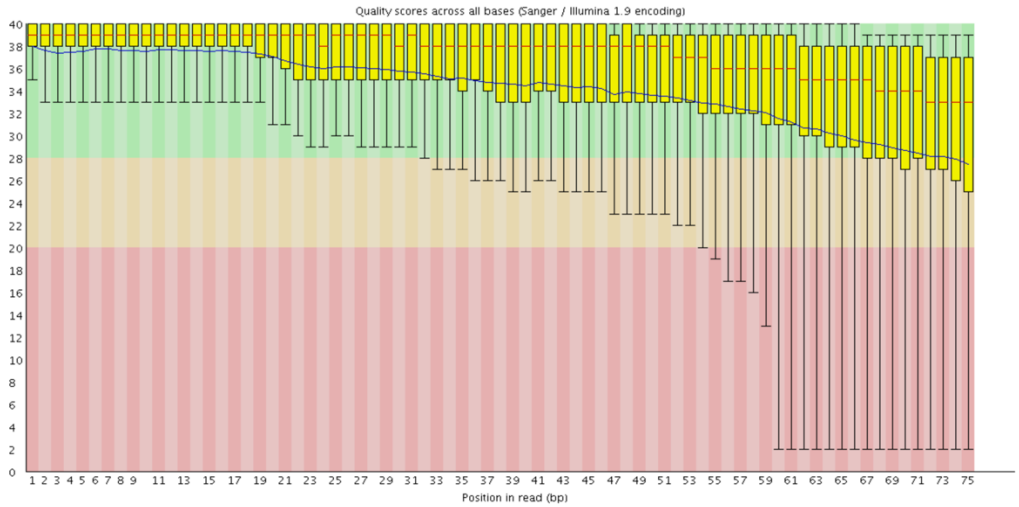
1) file:///Users/sultanalharbi/Desktop/Schizophrenia_analysis/ERR103431_S02_fastqc.html
2) file:///Users/sultanalharbi/Desktop/Schizophrenia_analysis/ERR103432_S04_fastqc.html
3) file:///Users/sultanalharbi/Desktop/Schizophrenia_analysis/ERR103433_S05_fastqc.html
4) file:///Users/sultanalharbi/Desktop/Schizophrenia_analysis/ERR103436_S23_fastqc.html



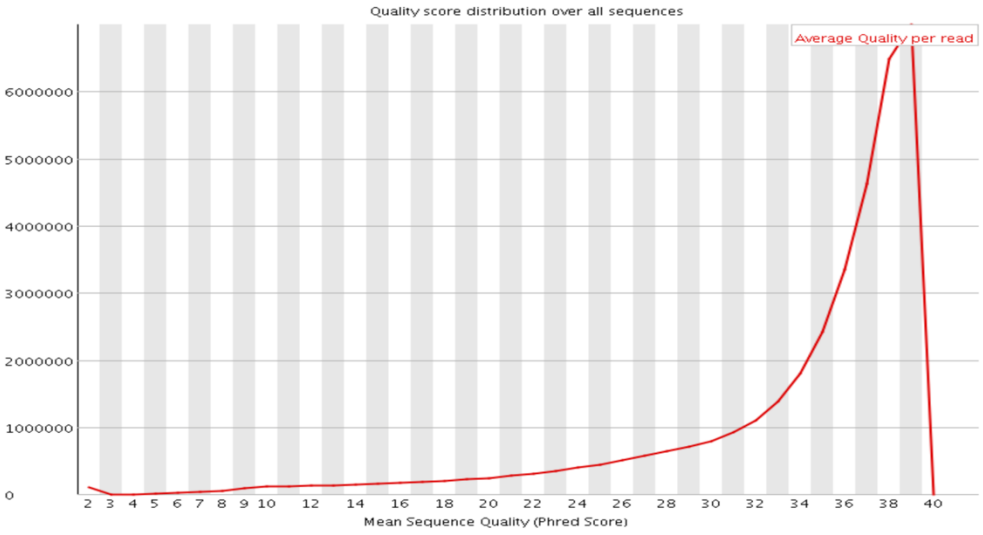
Basic Statistics

Measure	Value
Filename	ERR103436_S23.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	36526047
Sequences flagged as poor quality	0
Sequence length	75
%GC	49

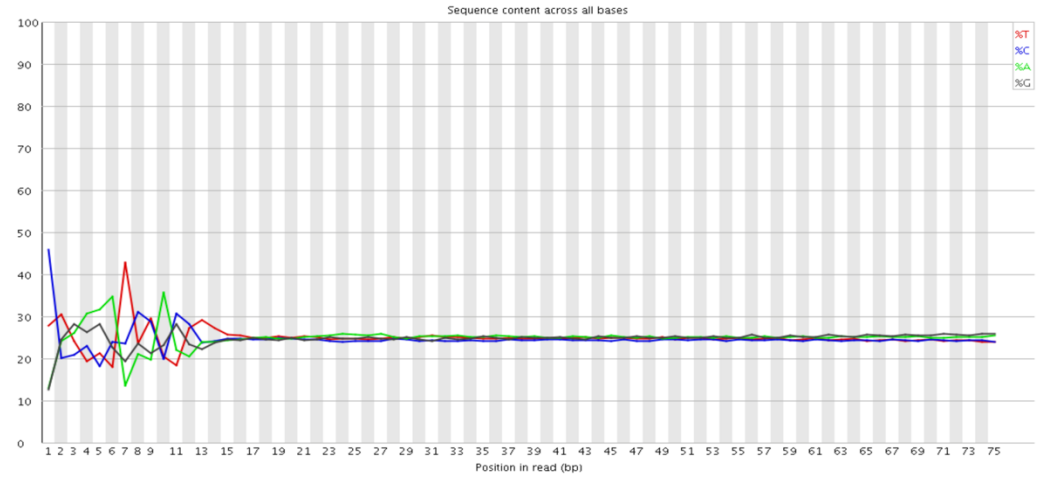
Per base sequence quality



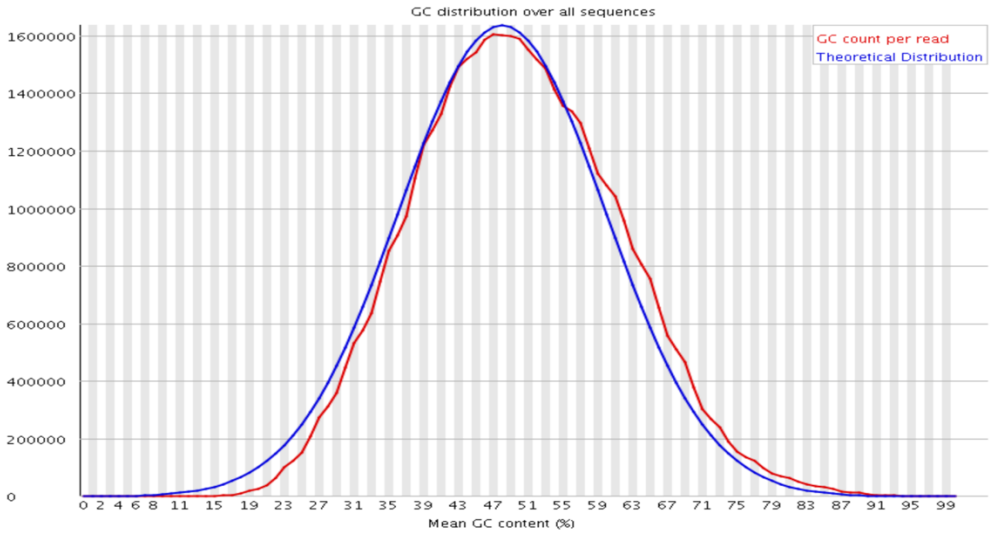
Per sequence quality scores



Per base sequence content



Per sequence GC content



C) Tophat2 Results (aligning results)

```
sultanalharbi — -bash — 119×24
[sultanalharbi@SULTANs-iMac:~$ head /Users/sultanalharbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/ERR103422_C02/align_summary
.txt
Reads:
          Input     :  14880770
           Mapped   :  10505850 (70.6% of input)
            of these:    985733 ( 9.4%) have multiple alignments (578 have >20)
70.6% overall read mapping rate.
```

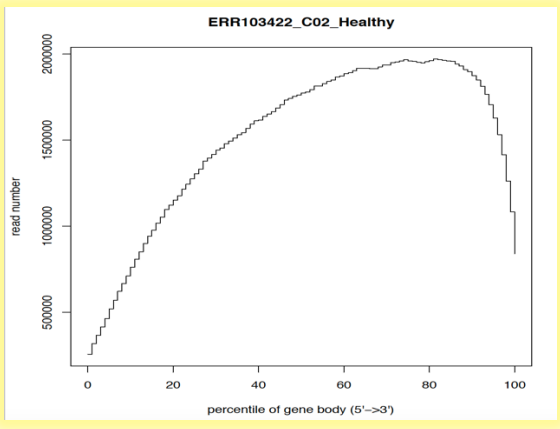
```
sultanalharbi — -bash — 119×24
sultanalharbi@SULTANs-iMac:~$ head /Users/sultanalharbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/ERR103423_C04/align_summary
.txt
Reads:
          Input     :  31819015
           Mapped   :  28779389 (90.4% of input)
            of these:   2594075 ( 9.0%) have multiple alignments (1640 have >20)
90.4% overall read mapping rate.
```

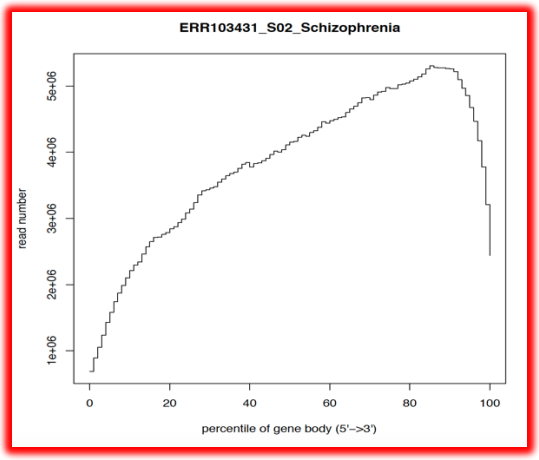
D) Calculating the distribution of reads to different genomic feature types

```
sultanalharbi@SULTANs-iMac:~$ cd /Users/sultanalharbi/Desktop/RSeQC-2.3.6/scripts
sultanalharbi@SULTANs-iMac:~/Desktop/RSeQC-2.3.6/scripts$ python read_distribution.py -r
/Users/sultanalharbi/Desktop/Genomes_Indexes/Homo_sapiens_GRCh37.bed -i /Users/sultanalha
rbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/ERR103422_C02/accepted_hits.bam
processing /Users/sultanalharbi/Desktop/Genomes_Indexes/Homo_sapiens_GRCh37.bed ... Done
processing /Users/sultanalharbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/ERR103422_C02/accepte
d_hits.bam ... Finished

Total Reads                   10505850
Total Tags                    12752907
Total Assigned Tags           9964920
=====================================================================
Group                 Total_bases         Tag_count           Tags/Kb
CDS_Exons             36860499            6140726             166.59
5'UTR_Exons           35730397            459558              12.86
3'UTR_Exons           57806939            2718461             47.03
Introns               1453670930          542622              0.37
TSS_up_1kb            31788498            7804                0.25
TSS_up_5kb            141452483           18410               0.13
TSS_up_10kb           252939815           25671               0.10
TES_down_1kb          33557126            20260               0.60
TES_down_5kb          145246303           61836               0.43
TES_down_10kb         255858394           77882               0.30
=====================================================================
```

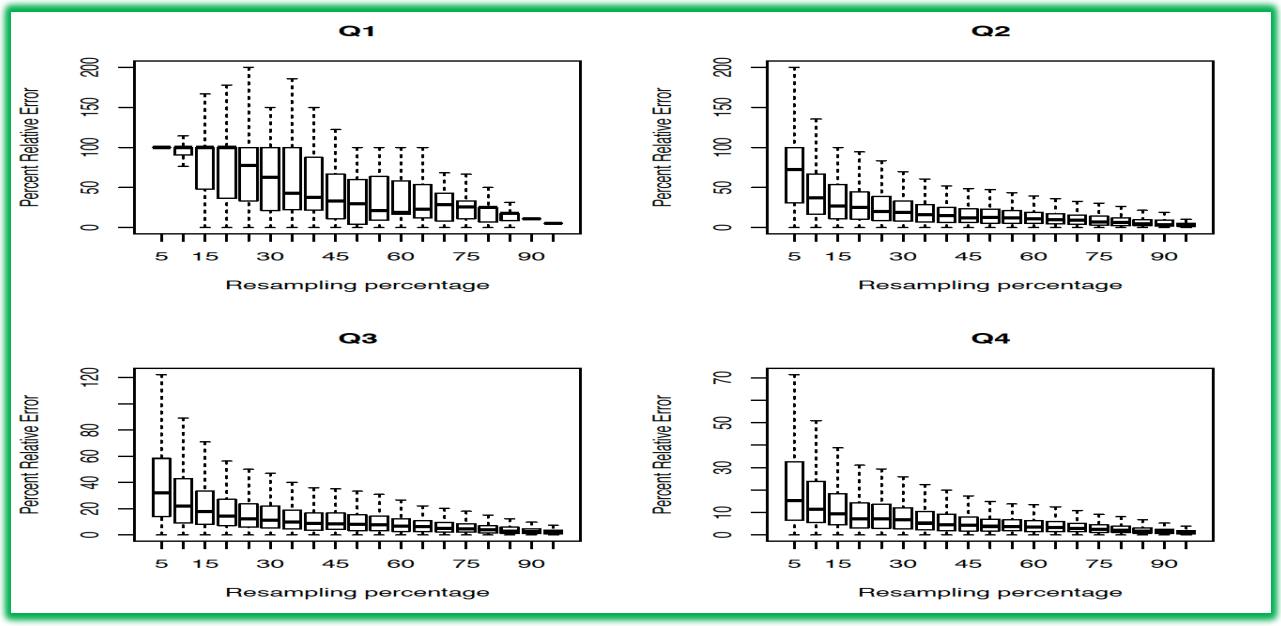
```
sultanalharbi@SULTANs-iMac:~/Desktop/RSeQC-2.3.6/scripts$ python read_distribution.py -r
/Users/sultanalharbi/Desktop/Genomes_Indexes/Homo_sapiens_GRCh37.bed -i /Users/sultanalha
rbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/ERR103431_S02/accepted_hits.bam
processing /Users/sultanalharbi/Desktop/Genomes_Indexes/Homo_sapiens_GRCh37.bed ... Done
processing /Users/sultanalharbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/ERR103431_S02/accepte
d_hits.bam ... Finished

Total Reads                   28469282
Total Tags                    33097762
Total Assigned Tags           24670495
=====================================================================
Group                 Total_bases         Tag_count           Tags/Kb
CDS_Exons             36860499            12374885            335.72
5'UTR_Exons           35730397            1456686             40.77
3'UTR_Exons           57806939            9146534             158.23
Introns               1453670930          1381162             0.95
TSS_up_1kb            31788498            24125               0.76
TSS_up_5kb            141452483           57201               0.40
TSS_up_10kb           252939815           78667               0.31
TES_down_1kb          33557126            66040               1.97
TES_down_5kb          145246303           186425              1.28
TES_down_10kb         255858394           232561              0.91
=====================================================================
```

E) Checking the uniformity of transcripts along 3' and 5' ends

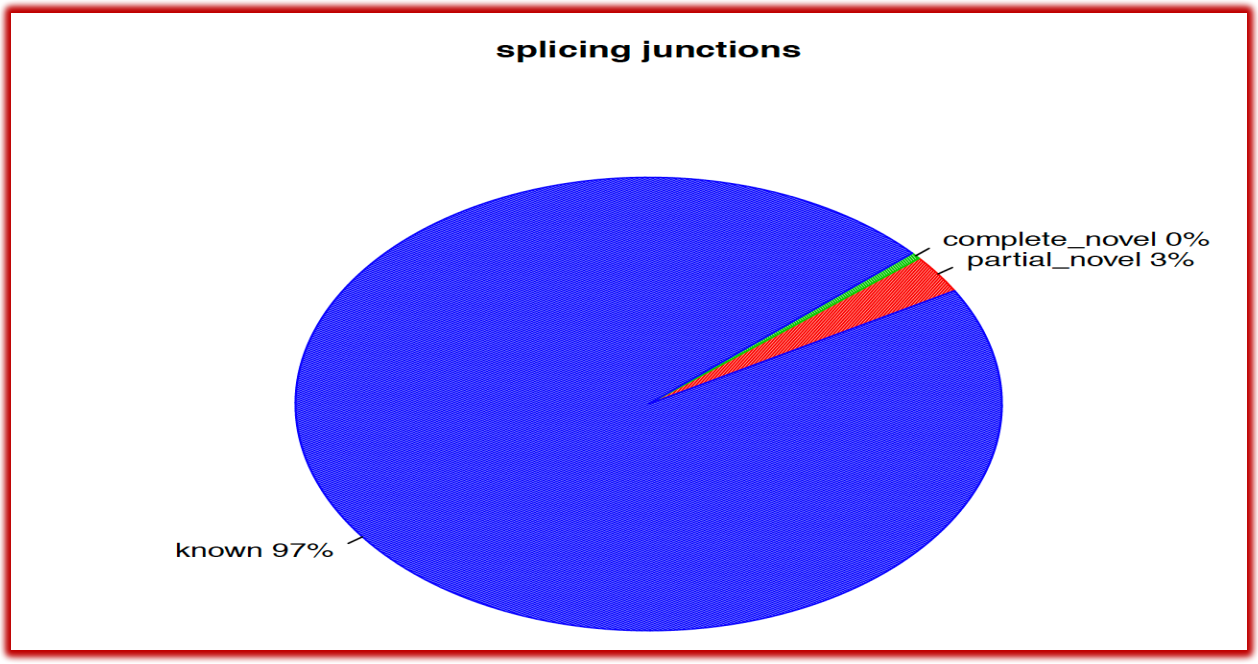




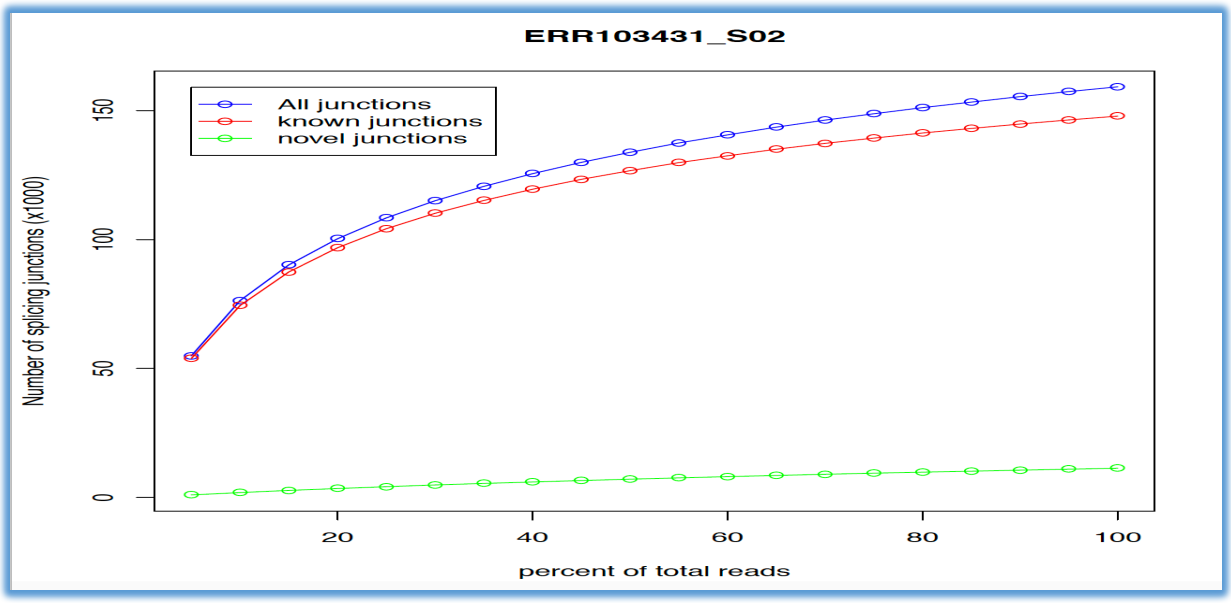
F) Calculating abundance in RPKM units



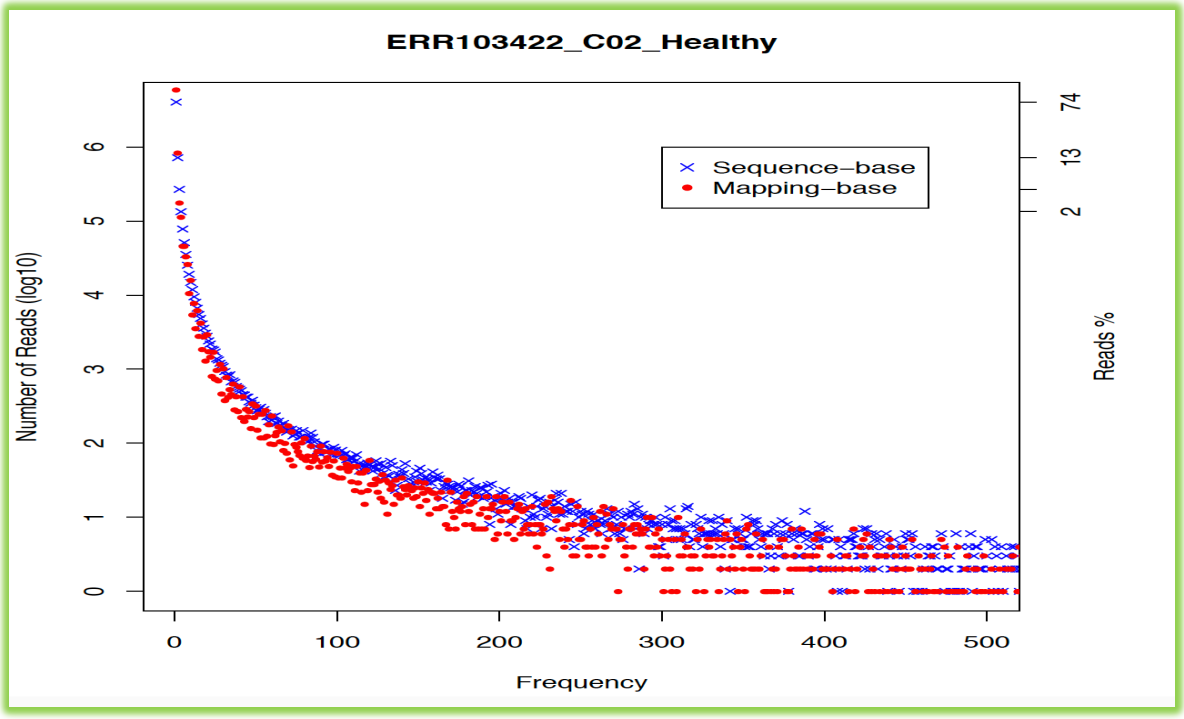
G) Detecting splicing junctions as novel, partially and known



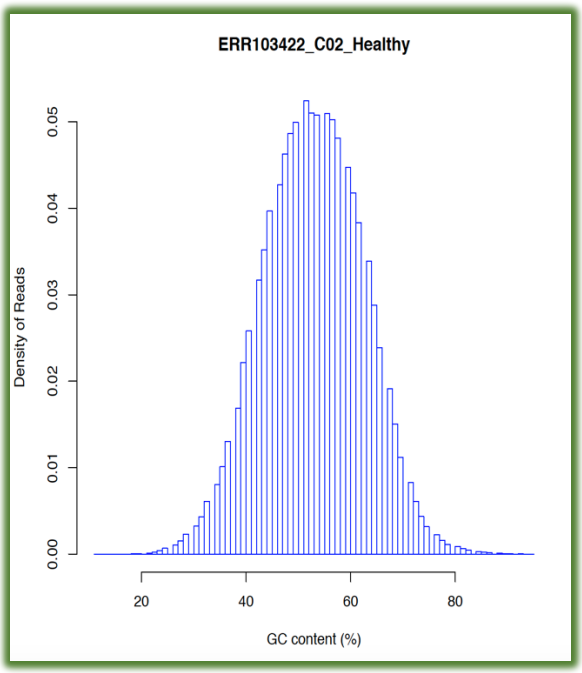
K) Detecting junctions in each subset

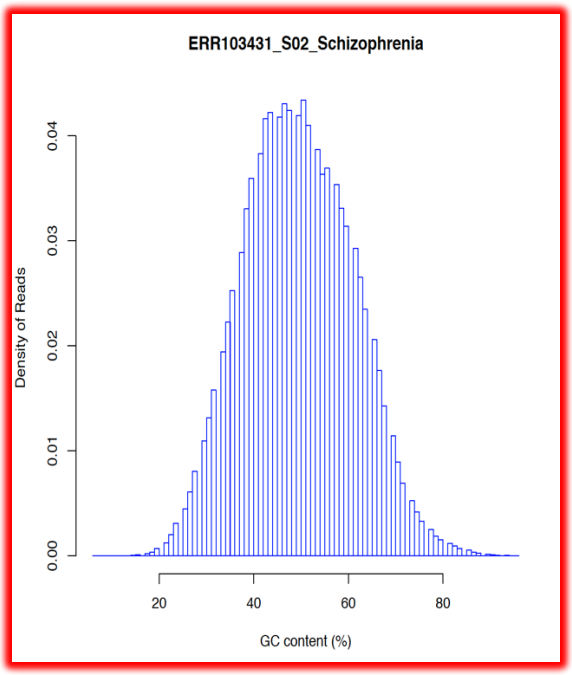


L) Checking Read Duplication

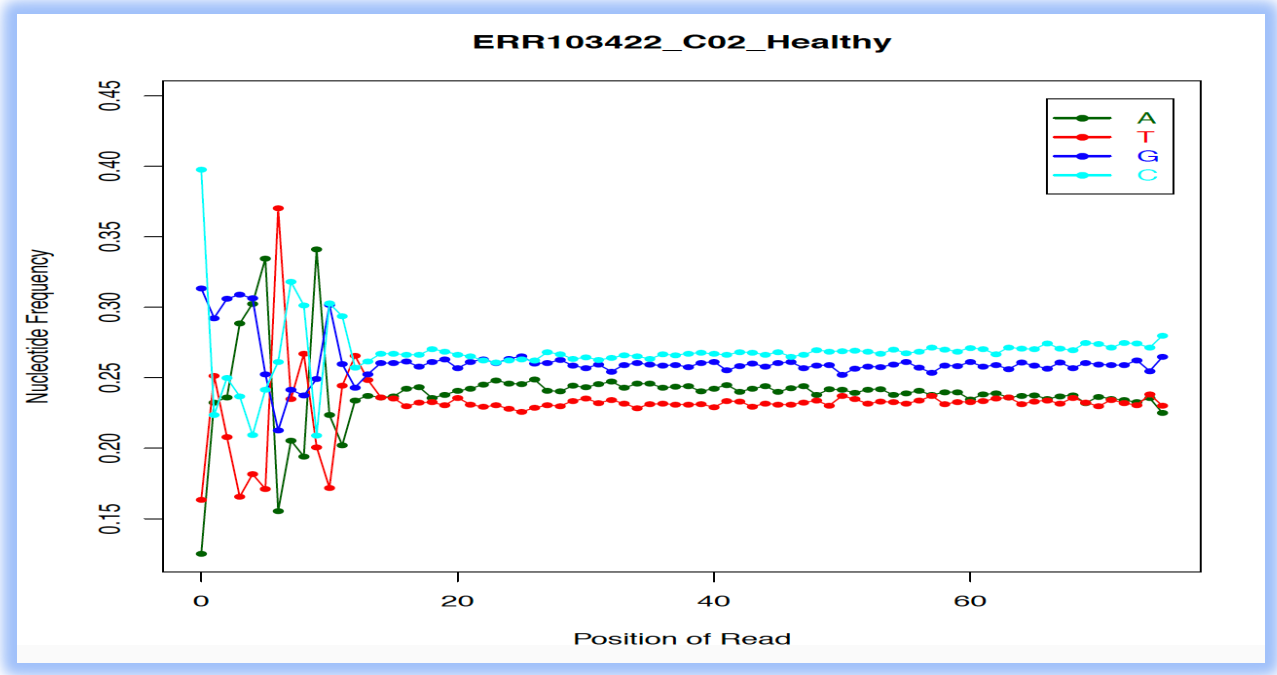


M) Distribution of read GC content (%):

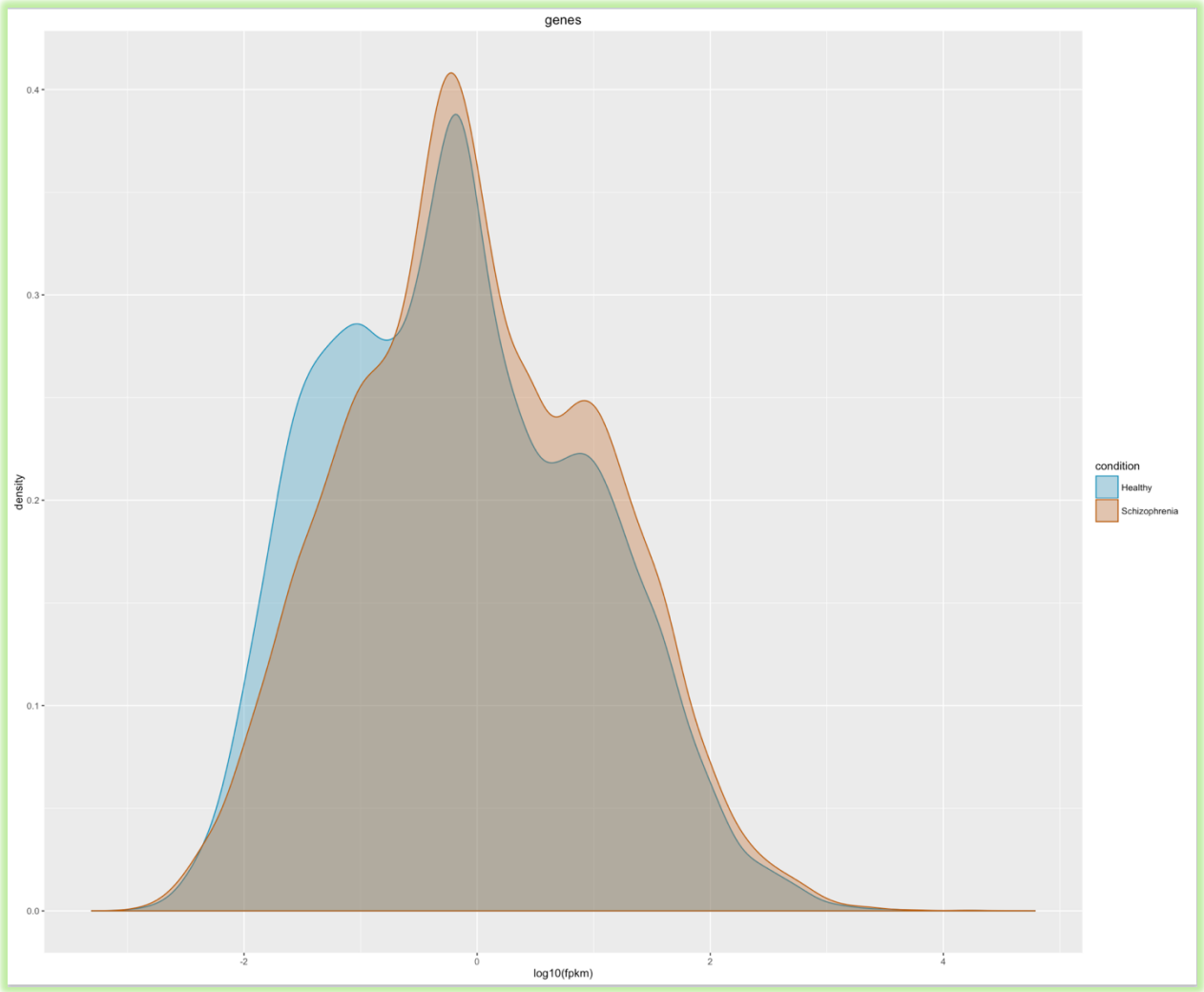




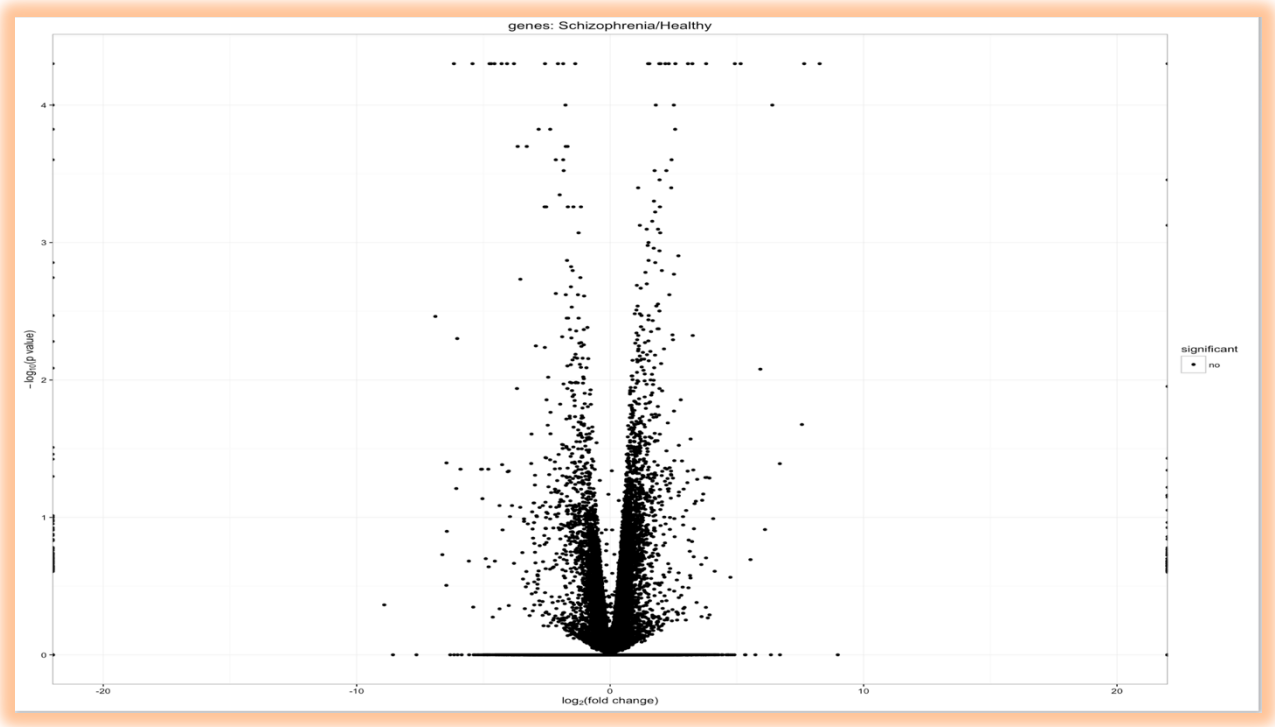
N) Read - Nucleotide vs Cycle (Phred base score vs. position in read):



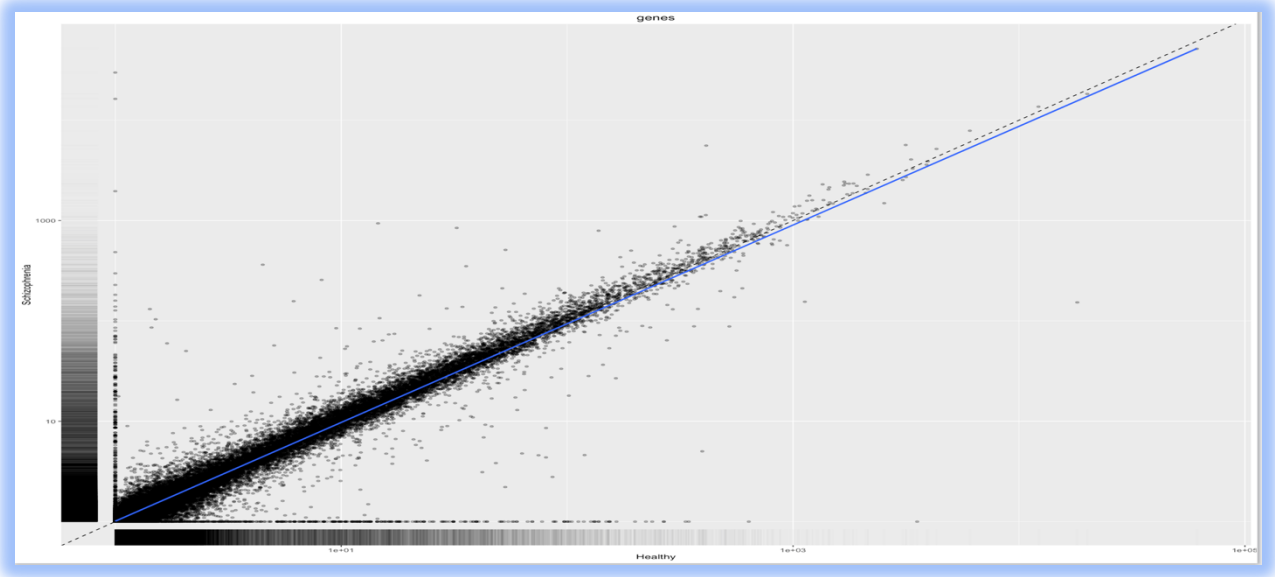
Results of Differential Expression analysis



1) A plot showing the expression levels for each sample, using the csDensity Function (CummeRbund package). Command line: `> csDensity(genes(cuffdata))`

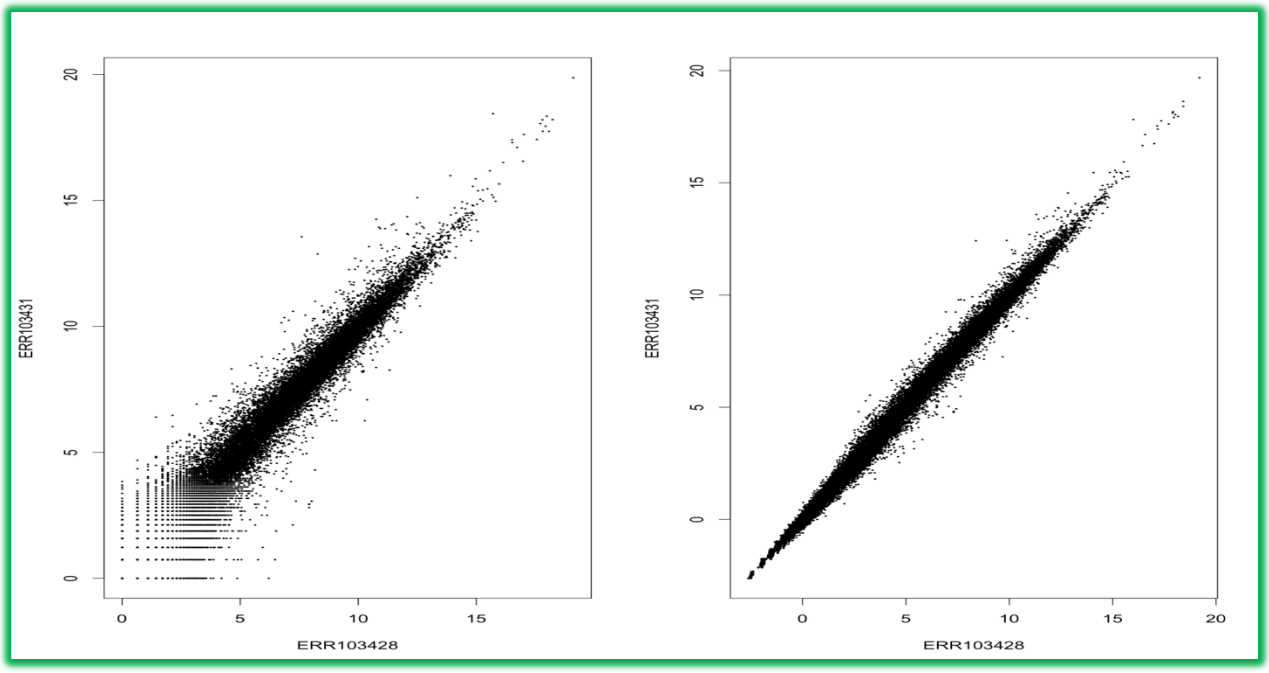


2) A Volcano plot showing differentially expressed genes across the two samples
Command line: `> csVolcano(genes(cuffdata), 'Healthy','Schizopherina')`

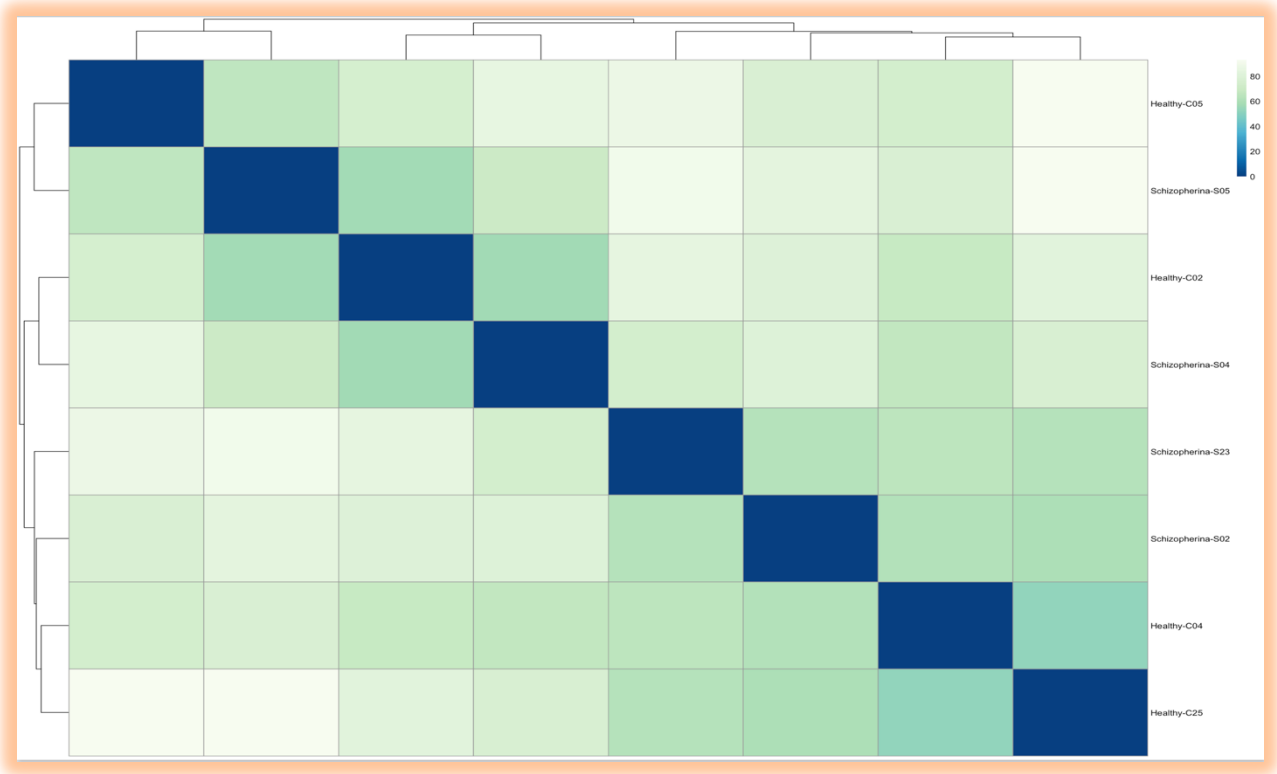


3) A scatter plot comparing the gene across the two samples Command Line:

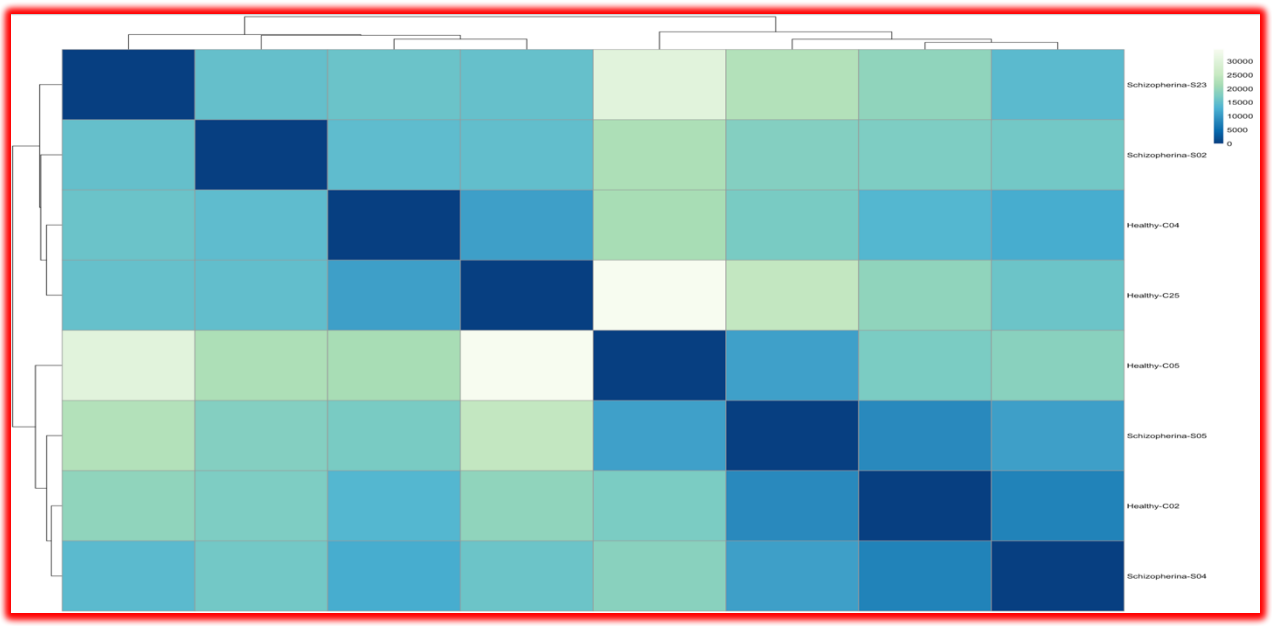
```
>csScatter(genes(cuffdata), 'Healthy', 'Schizopherina', smooth = T, logMode= TRUE, colorByStatus= T, hexbin=T)
```



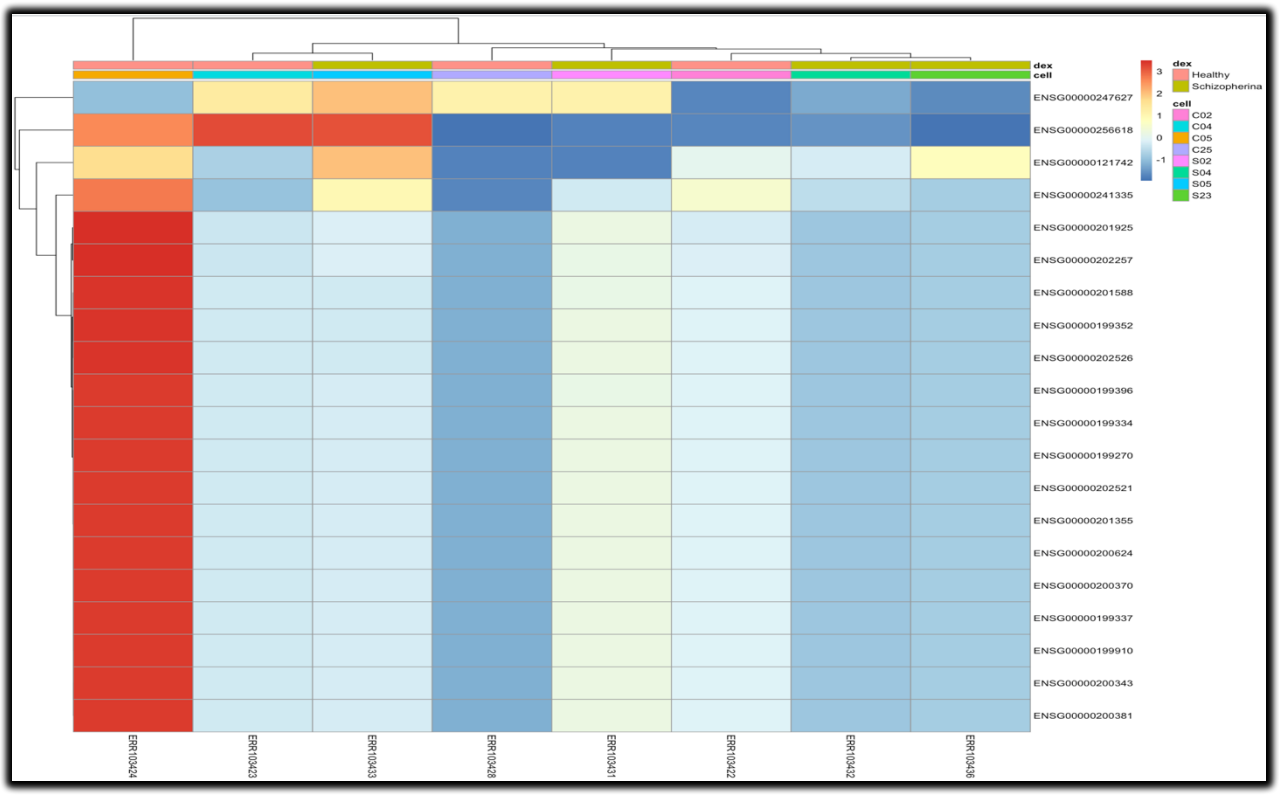
4) Scatterplot of transformed counts from two samples. Shown are scatterplots using the log2 transform of normalized counts (lift side) and using the rlog (right side).



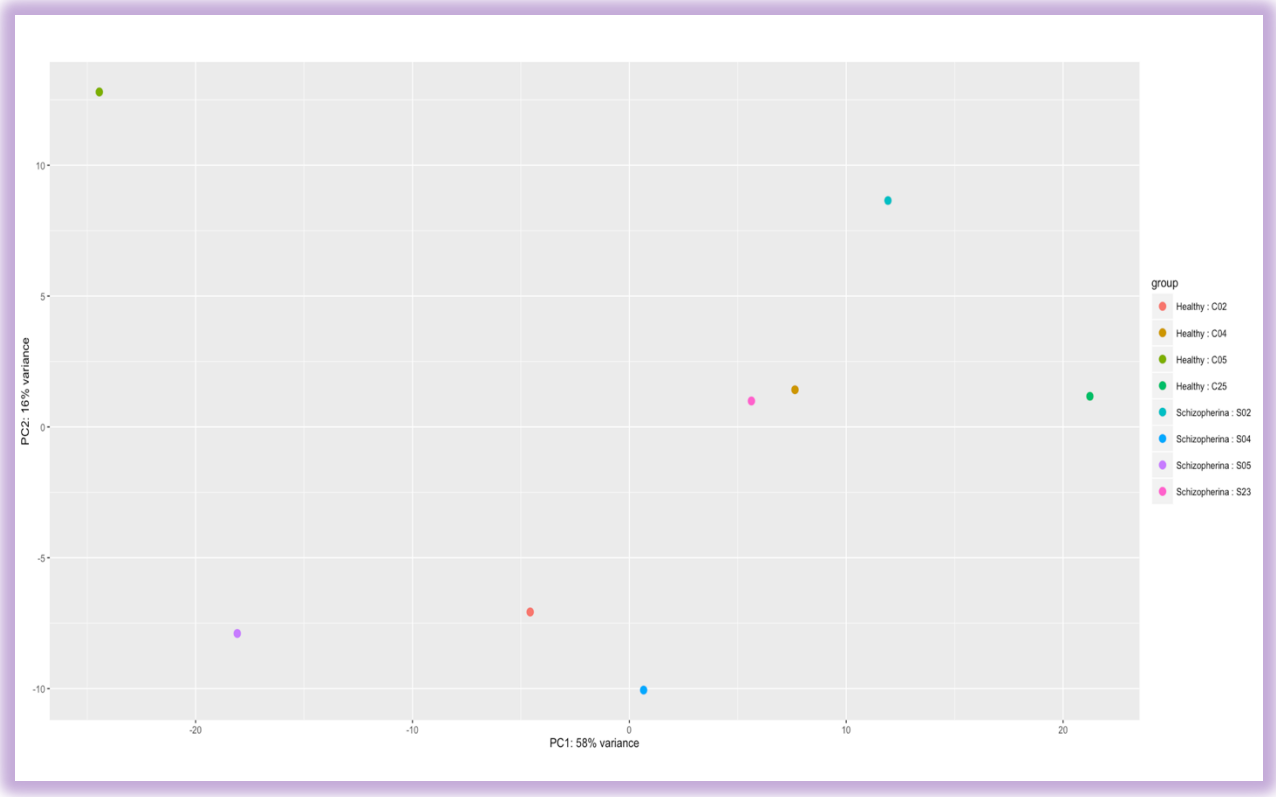
5) Heatmap of sample-to-sample distances using the rlog-transformed values



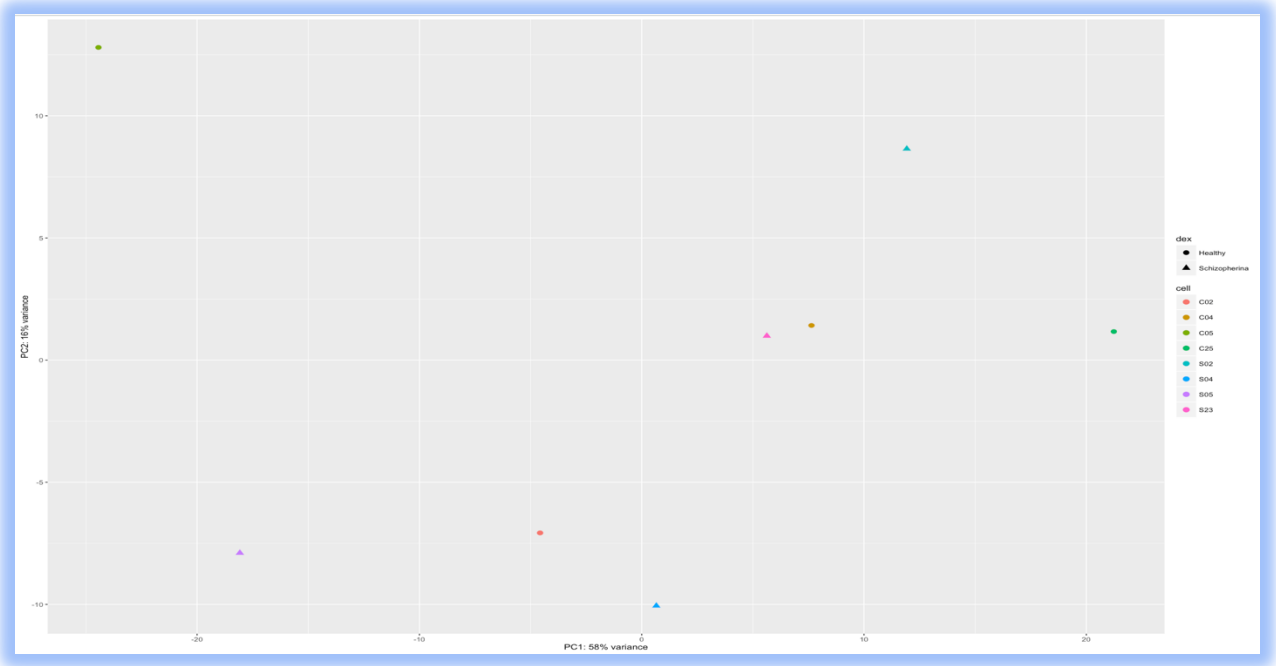
6) Heatmap of sample-to-sample distances using the Poisson Distance.



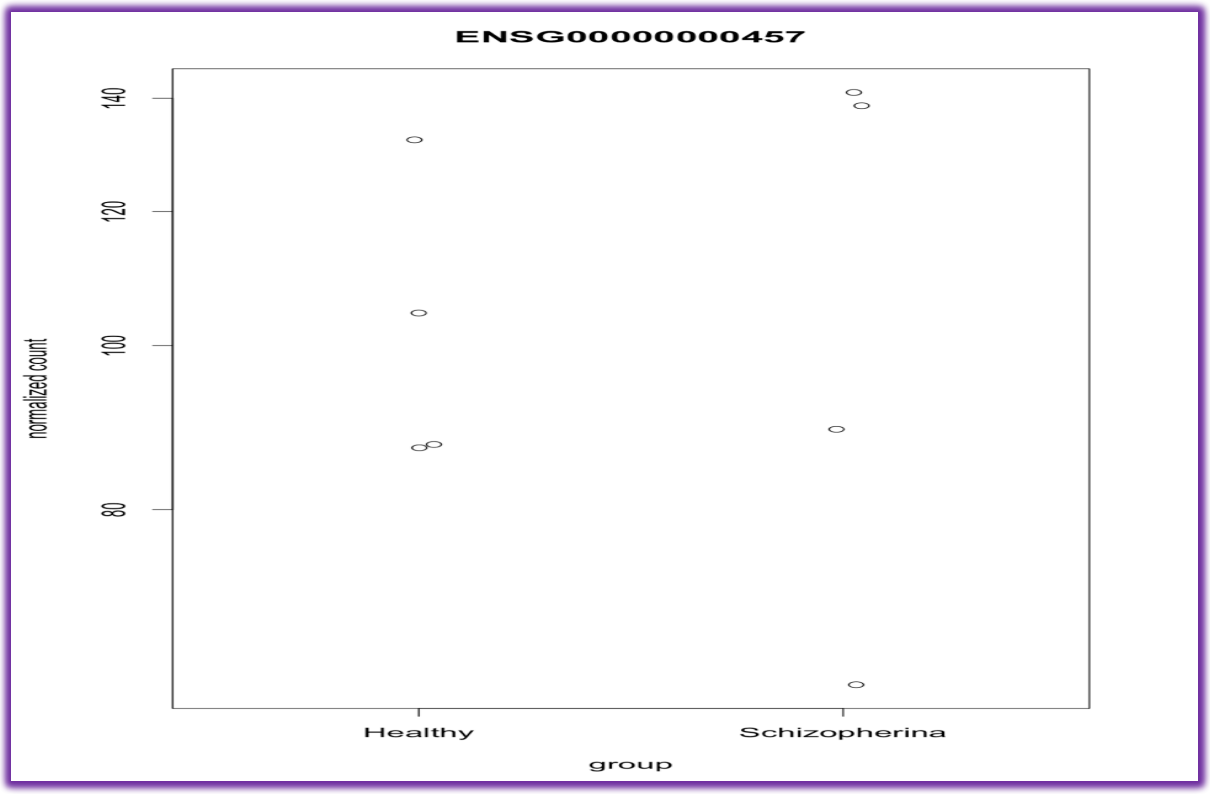
7) Heatmap of relative rlog-transformed vaules across samples.



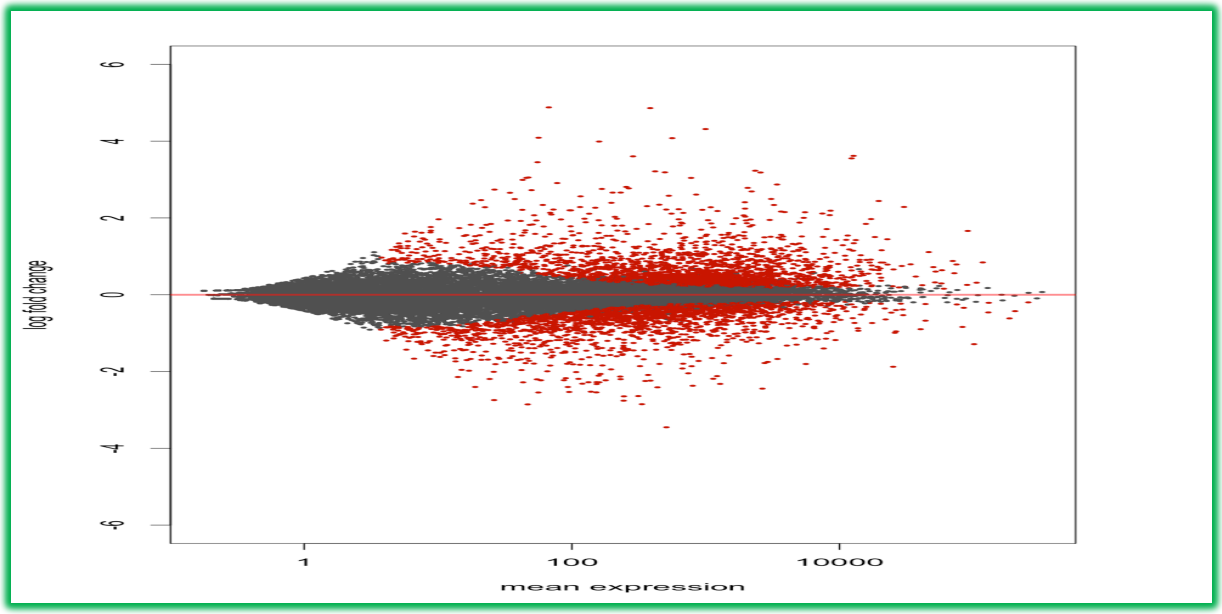
8) PCA plot using the rlog-transformed values. Each unique combination of Healthy and Schizophrenia sample is given its own colour.



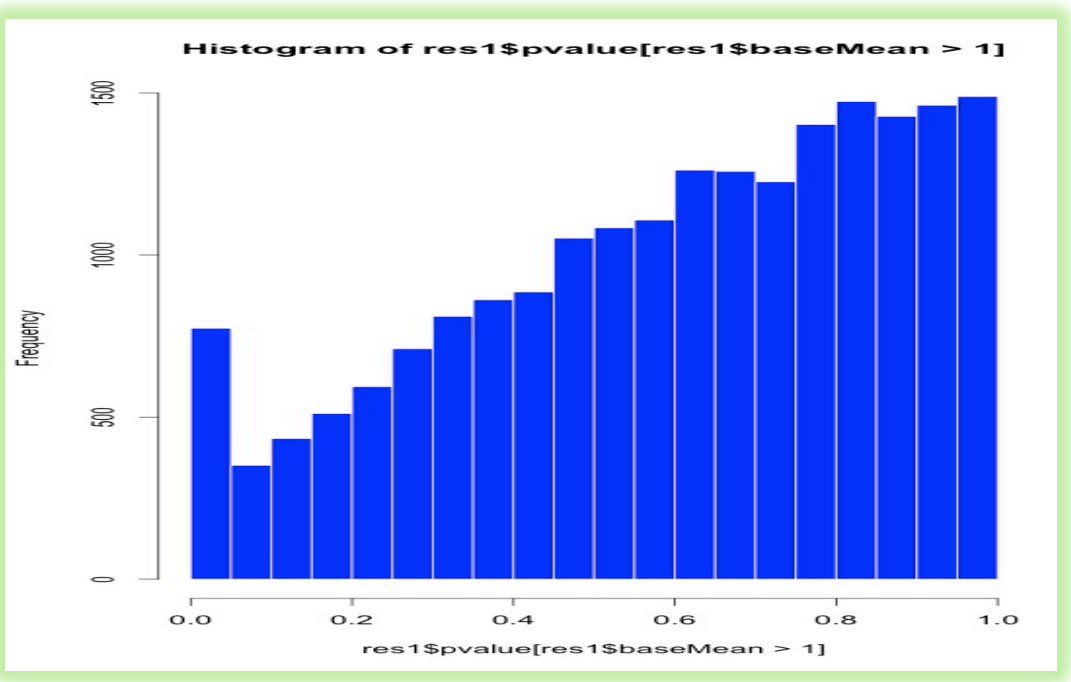
9) PCA plot (Plotting symbol for sample condition).



10) Normalized counts for a single gene over Schizophrenia group.



11) An MA-plot of a test for large log2 fold changes. The red points indicate genes for which the log2 fold change was significantly higher than 1 or less than -1



12) Histogram of p values for genes with mean normalized count larger than 1.

Second Pipeline

A) Build the reference index

```
./bowtie2-build <reference_in> <bt2_index_base>
```

```
sultanalharbi@SULTANs-iMac:~/Desktop/RNA_seq_Analysis/SULTAN_BIN$ ./bowtie2-build /Users/sultanalharbi/Desktop/Arabidopsis_thaliana.TAIR10.Genome.fa/Arabidopsis_thaliana.TAIR10.31.dna_rm.genome.fa Arabidopsis_thaliana.TAIR10.31
Settings:
  Output files: "Arabidopsis_thaliana.TAIR10.31.*.bt2"
  Line rate: 6 (line is 64 bytes)
  Lines per side: 1 (side is 64 bytes)
  Offset rate: 4 (one in 16)
  FTable chars: 10
  Strings: unpacked
  Max bucket size: default
  Max bucket size, sqrt multiplier: default
  Max bucket size, len divisor: 4
  Difference-cover sample period: 1024
  Endianness: little
  Actual local endianness: little
  Sanity checking: disabled
  Assertions: disabled
  Random seed: 0
```

B)Using Tophat (Version 2.0.10)

Mapping RNA-seq Reads to a reference

```
./tophat2 -o basename_Th2output/ -p 8 <index_genome> reads1 reads2
```

```
sultanalharbi@SULTANs-iMac:~/Desktop/RNA_seq_Analysis/SULTAN_BIN$ ./tophat2 -p 32 -o Tophat2output1SRR671949_Root_polyA_KNO3_Treatment_rep2 /Users/sultanalharbi/Desktop/Arabidopsis_thaliana_Ensembl_TAIR10/Arabidopsis_thaliana/Ensembl/TAIR10/Sequence/Bowtie2Index/genome /Users/sultanalharbi/Desktop/RNA_seq_Reads.SRA/SRR671949_Root_polyA_KNO3_Treatment_rep2.fastq

[2016-07-01 06:25:11] Beginning TopHat run (v2.0.10)
-----------------------------------------------
[2016-07-01 06:25:11] Checking for Bowtie
                Bowtie version:         2.1.0.0
[2016-07-01 06:25:11] Checking for Samtools
              Samtools version:         0.1.7.0
[2016-07-01 06:25:11] Checking for Bowtie index files (genome)..
[2016-07-01 06:25:11] Checking for reference FASTA file
[2016-07-01 06:25:11] Generating SAM header for /Users/sultanalharbi/Desktop/Arabidopsis_thaliana_Ensembl_TAIR10/Arabidopsis_thaliana/Ensembl/TAIR10/Sequence/Bowtie2Index/genome
[2016-07-01 06:25:11] Preparing reads
         left reads: min. length=50, max. length=50, 5177274 kept reads (5143 discarded)
[2016-07-01 06:29:01] Mapping left_kept_reads to genome genome with Bowtie2
[2016-07-01 06:37:33] Mapping left_kept_reads_seg1 to genome genome with Bowtie2 (1/2)
[2016-07-01 06:39:24] Mapping left_kept_reads_seg2 to genome genome with Bowtie2 (2/2)
[2016-07-01 06:40:32] Searching for junctions via segment mapping
        Coverage-search algorithm is turned on, making this step very slow
        Please try running TopHat again with the option (--no-coverage-search) if this step takes too much time or memory.
[2016-07-01 06:44:38] Retrieving sequences for splices
[2016-07-01 06:44:42] Indexing splices
[2016-07-01 06:44:56] Mapping left_kept_reads_seg1 to genome segment_juncs with Bowtie2 (1/2)
[2016-07-01 06:45:32] Mapping left_kept_reads_seg2 to genome segment_juncs with Bowtie2 (2/2)
[2016-07-01 06:46:08] Joining segment hits
[2016-07-01 06:46:56] Reporting output tracks
-----------------------------------------------
[2016-07-01 06:55:54] A summary of the alignment counts can be found in Tophat2output1SRR671949_Root_polyA_KNO3_Treatment_rep2/align_summary.txt
[2016-07-01 06:55:54] Run complete: 00:30:42 elapsed
sultanalharbi@SULTANs-iMac:~/Desktop/RNA_seq_Analysis/SULTAN_BIN$
```

C)Using cufflinks to assemble transcripts

```
./cufflinks -p 32 -o files to this directory <accepted bam file>
```

```
sultanalharbi@SULTANs-iMac:~/Desktop/RNA_seq_Analysis/SULTAN_BIN$ ./cufflinks -p 32 -o cufflinksoutput_SRR671949_Root_polyA_KNO3_Treatment_rep2 /Users/sultanalharbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/Tophat2outputSRR671949_Root_polyA_KNO3_Treatment_rep2/accepted_hits.bam
Warning: Your version of Cufflinks is not up-to-date. It is recommended that you upgrade to Cufflinks v2.2.1 to benefit from the most recent features and bug fixes (http://cufflinks.cbcb.umd.edu).
[05:25:01] Inspecting reads and determining fragment length distribution.
> Processed 54756 loci.                        [*************************] 100%
> Map Properties:
>       Normalized Map Mass: 1766793.00
>       Raw Map Mass: 1766793.00
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                Default Std Dev: 80
[05:26:02] Assembling transcripts and estimating abundances.
> Processed 54781 loci.                        [*************************] 100%
```

```
./cuffmerge -g genes.gtf -s genome.fa -p 2 assembled.txt
```

```
SULTAN_BIN — -bash — 145×51
...alysis/SULTAN_BIN — -bash | ...alysis/SULTAN_BIN — -bash | ...alysis/SULTAN_BIN — -bash | ...alysis/SULTAN_BIN — -bash | ...alysis/SULTAN_BIN — -bash
cuffmerge
Done
sultanalharbi@SULTANs-iMac:~/Desktop/RNA_seq_Analysis/SULTAN_BIN$ ./cuffmerge -g /Users/sultanalharbi/Desktop/Arabidopsis_thaliana_Ensembl_TAIR10
/Arabidopsis_thaliana/Ensembl/TAIR10/Annotation/Archives/archive-2015-07-17-14-28-46/Genes/genes.gtf -s /Users/sultanalharbi/Desktop/Arabidopsis_
thaliana_Ensembl_TAIR10/Arabidopsis_thaliana/Ensembl/TAIR10/Sequence/WholeGenomeFasta/genome.fa -p 2 /Users/sultanalharbi/Desktop/assembled.txt

[Fri Jul  1 05:46:02 2016] Beginning transcriptome assembly merge
-----------------------------------------------

[Fri Jul  1 05:46:02 2016] Preparing output location ./merged_asm/
[Fri Jul  1 05:46:06 2016] Converting GTF files to SAM
[05:46:06] Loading reference annotation.
[05:46:06] Loading reference annotation.
[05:46:07] Loading reference annotation.
[05:46:07] Loading reference annotation.
[Fri Jul  1 05:46:07 2016] Quantitating transcripts
Warning: Your version of Cufflinks is not up-to-date. It is recommended that you upgrade to Cufflinks v2.2.1 to benefit from the most recent feat
ures and bug fixes (http://cufflinks.cbcb.umd.edu).
Command line:
cufflinks -o ./merged_asm/ -F 0.05 -g /Users/sultanalharbi/Desktop/Arabidopsis_thaliana_Ensembl_TAIR10/Arabidopsis_thaliana/Ensembl/TAIR10/Annota
tion/Archives/archive-2015-07-17-14-28-46/Genes/genes.gtf -q --overhang-tolerance 200 --library-type=transfrags -A 0.0 --min-frags-per-transfrag
0 --no-5-extend -p 2 ./merged_asm/tmp/mergeSam_tmp.4.mEujmd
[bam_header_read] EOF marker is absent. The input is probably truncated.
[bam_header_read] invalid BAM binary header (this is not a BAM file).
File ./merged_asm/tmp/mergeSam_tmp.4.mEujmd doesn't appear to be a valid BAM file, trying SAM...
[05:46:08] Loading reference annotation.
[05:46:28] Inspecting reads and determining fragment length distribution.
Processed 31328 loci.
> Map Properties:
>       Normalized Map Mass: 17176.00
>       Raw Map Mass: 17176.00
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[05:46:29] Assembling transcripts and estimating abundances.
Processed 31328 loci.
[Fri Jul  1 05:47:32 2016] Comparing against reference file /Users/sultanalharbi/Desktop/Arabidopsis_thaliana_Ensembl_TAIR10/Arabidopsis_thaliana
/Ensembl/TAIR10/Annotation/Archives/archive-2015-07-17-14-28-46/Genes/genes.gtf
Warning: Your version of Cufflinks is not up-to-date. It is recommended that you upgrade to Cufflinks v2.2.1 to benefit from the most recent feat
ures and bug fixes (http://cufflinks.cbcb.umd.edu).
[Fri Jul  1 05:48:12 2016] Comparing against reference file /Users/sultanalharbi/Desktop/Arabidopsis_thaliana_Ensembl_TAIR10/Arabidopsis_thaliana
/Ensembl/TAIR10/Annotation/Archives/archive-2015-07-17-14-28-46/Genes/genes.gtf
Warning: Your version of Cufflinks is not up-to-date. It is recommended that you upgrade to Cufflinks v2.2.1 to benefit from the most recent feat
ures and bug fixes (http://cufflinks.cbcb.umd.edu).
```

```
./cuffdiff -o cuffdiff_result -b genome.fa -p 32 -L control,treatment -u genes.gtf  list of accepted bam files
```

```
SULTAN_BIN — -bash — 260×52
...esktop/RNA_seq_Analysis/SULTAN_BIN — -bash | ...esktop/RNA_seq_Analysis/SULTAN_BIN — -bash | ...esktop/RNA_seq_Analysis/SULTAN_BIN — -bash | ...esktop/RNA_seq_Analysis/SULTAN_BIN — -bash | ...esktop/RNA_seq_Analysis/SULTAN_BIN — -bash
sultanalharbi@SULTANs-iMac:~/Desktop/RNA_seq_Analysis/SULTAN_BIN$ ./cuffdiff -o diff_result -b /Users/sultanalharbi/Desktop/Arabidopsis_thaliana_Ensembl_TAIR10/Arabidopsis_thaliana/Ensembl/TAIR10/Sequence/WholeGenomeFasta/genome.fa -p 32 -L Root_Kcl_control,Ro
ot_KNO3_treatment -u /Users/sultanalharbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/merged_asm/merged.gtf /Users/sultanalharbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/Tophat2outputSRR671946_Root_polyA_KCl_Control_rep1/accepted_hits.bam,/Users/sultanalharbi/Desktop/RNA_s
eq_Analysis/SULTAN_BIN/Tophat2outputSRR671947_Root_polyA_KCl_Control_rep2/accepted_hits.bam /Users/sultanalharbi/Desktop/RNA_seq_Analysis/SULTAN_BIN/Tophat2outputSRR671948_Root_polyA_KNO3_Treatment_rep1/accepted_hits.bam,/Users/sultanalharbi/Desktop/RNA_seq_An
alysis/SULTAN_BIN/Tophat2outputSRR671949_Root_polyA_KNO3_Treatment_rep2/accepted_hits.bam
Warning: Your version of Cufflinks is not up-to-date. It is recommended that you upgrade to Cufflinks v2.2.1 to benefit from the most recent features and bug fixes (http://cufflinks.cbcb.umd.edu).
[05:58:05] Loading reference annotation and sequence.
[05:58:10] Inspecting maps and determining fragment length distributions.
[05:59:49] Modeling fragment count overdispersion.
[06:00:31] Modeling fragment count overdispersion.
> Map Properties:
>       Normalized Map Mass: 935260.43
>       Raw Map Mass: 843861.05
>       Number of Multi-Reads: 471527 (with 951569 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
> Map Properties:
>       Normalized Map Mass: 935260.43
>       Raw Map Mass: 1193204.67
>       Number of Multi-Reads: 699596 (with 1401690 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
> Map Properties:
>       Normalized Map Mass: 935260.43
>       Raw Map Mass: 974566.32
>       Number of Multi-Reads: 587928 (with 1177192 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
> Map Properties:
>       Normalized Map Mass: 935260.43
>       Raw Map Mass: 810503.27
>       Number of Multi-Reads: 472883 (with 946607 total hits)
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[06:01:08] Calculating preliminary abundance estimates
> Processed 31313 loci.                         [*************************] 100%
[06:01:56] Learning bias parameters.
[06:02:44] Testing for differential expression and regulation in locus.
> Processed 31313 loci.                         [*************************] 100%
Performed 4941 isoform-level transcription difference tests
Performed 4780 tss-level transcription difference tests
Performed 4746 gene-level transcription difference tests
Performed 4655 CDS-level transcription difference tests
Performed 0 splicing tests
Performed 0 promoter preference tests
Performing 0 relative CDS output tests
Writing isoform-level FPKM tracking
Writing TSS group-level FPKM tracking
Writing gene-level FPKM tracking
Writing CDS-level FPKM tracking
Writing isoform-level count tracking
Writing TSS group-level count tracking
```

Using STAR (version 020201)

STAR (Spliced Transcripts Alignment to a reference) is a relatively new aligner and faster than tophat

1) Generating genome indexes

```
./star --runMode genomeGenerate --genomeDir storedDirectory --genomeFastaFiles genome.fa --
sjdbGTFfile annotation.gtf -   sjdbOverhang 74 --runThreadN 8
```

```
sultanalharbi@SULTANs-iMac:/Volumes/SULTAN2016/STAR-2.5.2a/bin/MacOSX_x86_64$ ./star --runThreadN 32 --runMode genomeGenerate --genomeDir /Volumes/SULTAN2016/Drosophila_melano
gaster_Ensembl_BDGP5.25/Drosophila_melanogaster/Ensembl/BDGP5.25/Sequence/WholeGenomeFasta --genomeFastaFiles /Volumes/SULTAN2016/Drosophila_melanogaster_Ensembl_BDGP5.25/Dros
ophila_melanogaster/Ensembl/BDGP5.25/Sequence/WholeGenomeFasta/genome.fa --sjdbGTFfile /Volumes/SULTAN2016/Drosophila_melanogaster_Ensembl_BDGP5.25/Drosophila_melanogaster/Ens
embl/BDGP5.25/Annotation/Archives/archive-2015-07-17-14-30-26/Genes/genes.gtf
Jun 26 17:28:35 ..... Started STAR run
Jun 26 17:28:35 ... Starting to generate Genome files
Jun 26 17:28:38 ... starting to sort  Suffix Array. This may take a long time...
Jun 26 17:28:38 ... sorting Suffix Array chunks and saving them to disk...
Jun 26 17:29:34 ... loading chunks from disk, packing SA...
Jun 26 17:29:39 ... Finished generating suffix array
Jun 26 17:29:39 ... Generating Suffix Array index
Jun 26 17:30:09 ... Completed Suffix Array index
Jun 26 17:30:09 ..... Processing annotations GTF
Jun 26 17:30:11 ..... Inserting junctions into the genome indices
Jun 26 17:30:33 ... writing Genome to disk ...
Jun 26 17:30:34 ... writing Suffix Array to disk ...
Jun 26 17:30:47 ... writing SAindex to disk
Jun 26 17:31:05 ..... Finished successfully
sultanalharbi@SULTANs-iMac:/Volumes/SULTAN2016/STAR-2.5.2a/bin/MacOSX_x86_64$
```

2) Mapping RNA-seq reads with a Reference

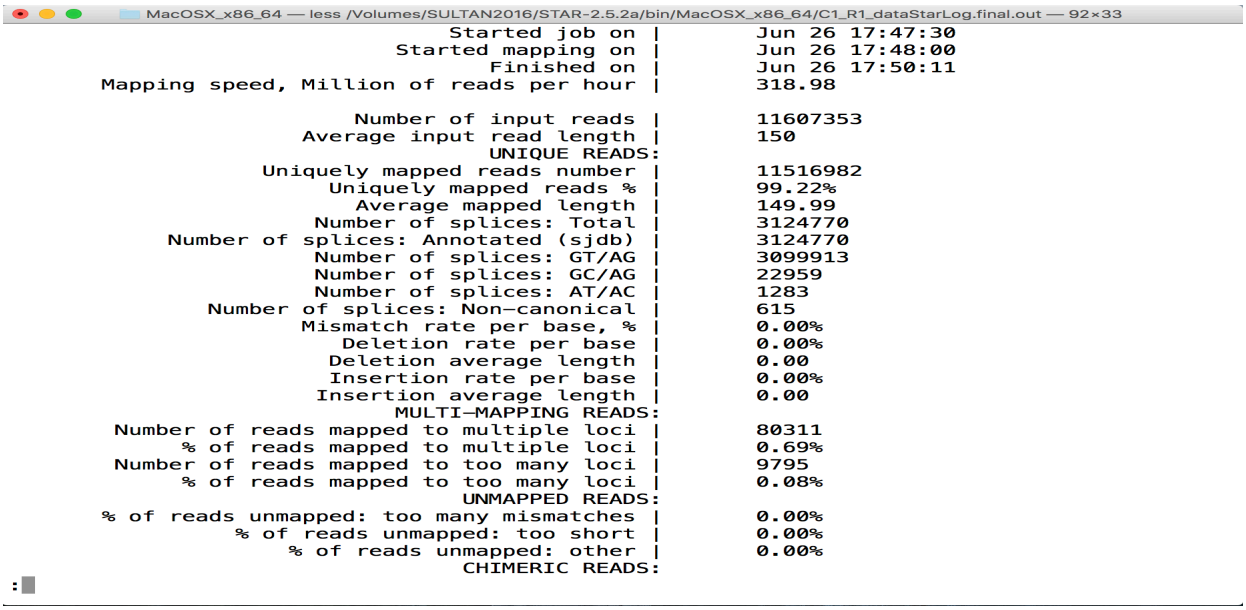
```
./star --genomeDir Directory to your STARGenomeIndex  --runThreadN 8 --readFilesIn RNA-seqFASTQ -
-outFileNamePrefix <OutputPrefix>
```

```
sultanalharbi@SULTANs-iMac:/Volumes/SULTAN2016/STAR-2.5.2a/bin/MacOSX_x86_64$ ./STAR --genomeDir /Volumes/SULTAN2016/Drosophila_melanogaster_Ensembl_BDGP5.25/Drosophila_melano
gaster/Ensembl/BDGP5.25/Sequence/WholeGenomeFasta --runThreadN 24 --readFilesIn /Volumes/SULTAN2016/GSE32038_simulated_fastq_files/GSM794483_C1_R1_1.fq /Volumes/SULTAN2016/GSE
32038_simulated_fastq_files/GSM794483_C1_R1_2.fq --outFileNamePrefix C1_R1_dataStar
Jun 26 17:47:30 ..... Started STAR run
Jun 26 17:47:30 ..... Loading genome
Jun 26 17:48:00 ..... Started mapping
Jun 26 17:50:11 ..... Finished successfully
sultanalharbi@SULTANs-iMac:/Volumes/SULTAN2016/STAR-2.5.2a/bin/MacOSX_x86_64$
```

The output of STAR aligner is five different files:

1) Basename_dataStarAligned.out.sam
2) Basename_dataStarLog.final.out
3) Basename_dataStarLog.out
4) Basename_dataStarLog.progress.out
5) Basename_dataStarSJ.out.tab

The Basename_dataStar.final.out file provides useful mapping statistics



Option	meaning	comment
--runMode	genomeGenerate	generate genome files
--genomeDir	path/to/your prefered/directory	string: path to the directory where genome files will be stored.
--genomeFastaFiles	The fasta files with genomic sequence	
--sjdbGTFfile	the GTF file with annotations	
--sjdbOverhang	int>0: length of the donor/acceptor sequence on each side of the junctions	ideally = (mate_length - 1)
--runThreadN	number of threads to run STAR	
--readFilesIn	RNA-seq reads (Fastq) file	
--outFileNamePrefix	output files name prefix	