The CCBR RNA-Seq Pipeline

Fathi Elloumi, Ph.D

NCI CCBR

3/20/2017

Agenda

- Introduction
- Data analysis Workflow
  - Review main steps
- CCBR RNA-Seq pipeline
  - Workflow overview
  - Quality Control reports
  - Principal Component Analysis PCA and differential expressed reports reports
  - Downstream analysis after running the pipeline
- Running the CCBR pipeline
  - Use case and demo

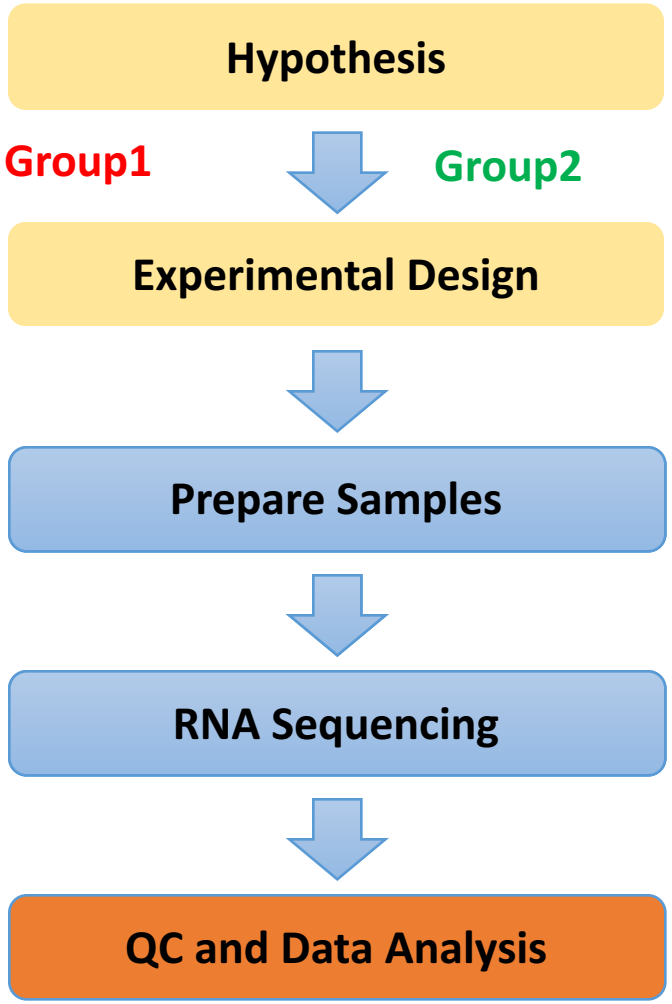
Agenda

- Introduction
- Data analysis Workflow
  - Review main steps
- CCBR RNA-Seq pipeline
  - Workflow overview
  - Quality Control reports
  - Principal Component Analysis PCA and differential expressed reports reports
  - Downstream analysis after running the pipeline
- Running the CCBR pipeline
  - Use case and demo

RNA-Seq Applications

- Differential Gene Expression
- Differential Transcript Expression
  - Still confined to known transcripts / isoforms
- Transcript Discovery / Whole Transcriptome Profiling
  - Interest is in looking for new isoforms or unannotated genes
- Others
  - SNP/Somatic Variant/Gene Fusion Detection

RNA-Seq project Overview



- RNA extraction protocol
- Depth
- Library type SE/PE
- Nb. Replicates
- …

Best Practices

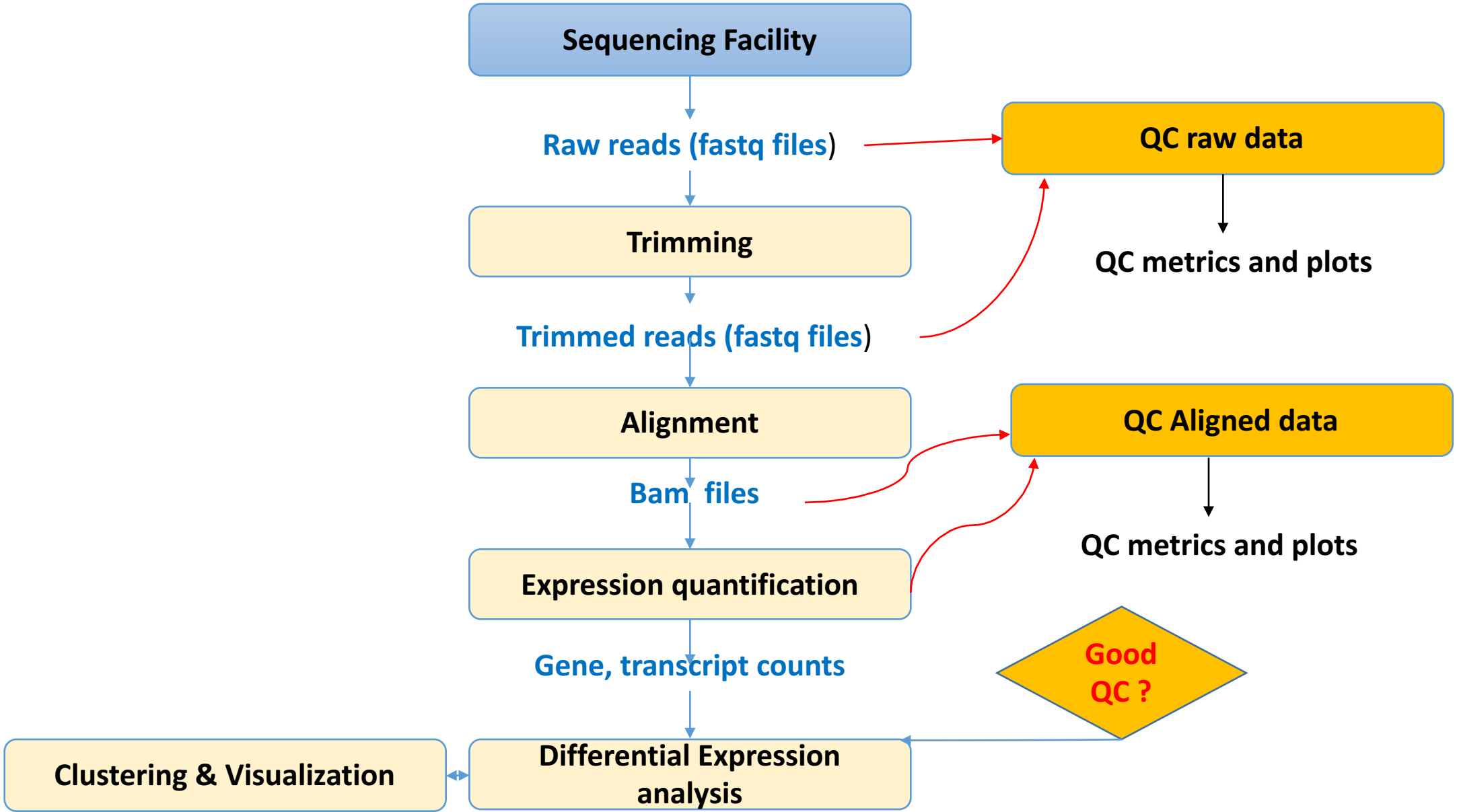
- Factor in at least 3 replicates (absolute minimum), but 4 if possible (optimum minimum). Biological replicates are recommended rather than technical replicates.
- Always process your RNA extractions at the same time. Extractions done at different times lead to unwanted batch effects.
- There are 2 major considerations for RNA-Seq libraries:
  - If you are interested in coding mRNA, you can select to use the mRNA library prep. The recommended sequencing depth is between 10-20M paired-end (PE) reads. Your RNA has to be high quality (RIN > 8).
  - If you are interested in long noncoding RNA as well, you can select the total RNA method, with sequencing depth ~25-60M PE reads. This is also an option if your RNA is degraded.
- Ideally to avoid lane batch effects, all samples would need to be multiplexed together and run on the same lane. This may require an initial MiSeq run for library balancing. Additional lanes can be run if more sequencing depth is needed.
- If you are unable to process all your RNA samples together and need to process them in batches, make sure that replicates for each condition are in each batch so that the batch effects can be measured and removed bioinformatically.

https://bioinformatics.cancer.gov/content/rna-seq

Agenda

- Introduction
- Data analysis Workflow
  - Review main steps
- CCBR RNA-Seq pipeline
  - Workflow overview
  - Quality Control reports
  - Principal Component Analysis PCA and differential expressed reports reports
  - Downstream analysis after running the pipeline
- Running the CCBR pipeline
  - Use case and demo

Typical RNA-Seq analysis workflow

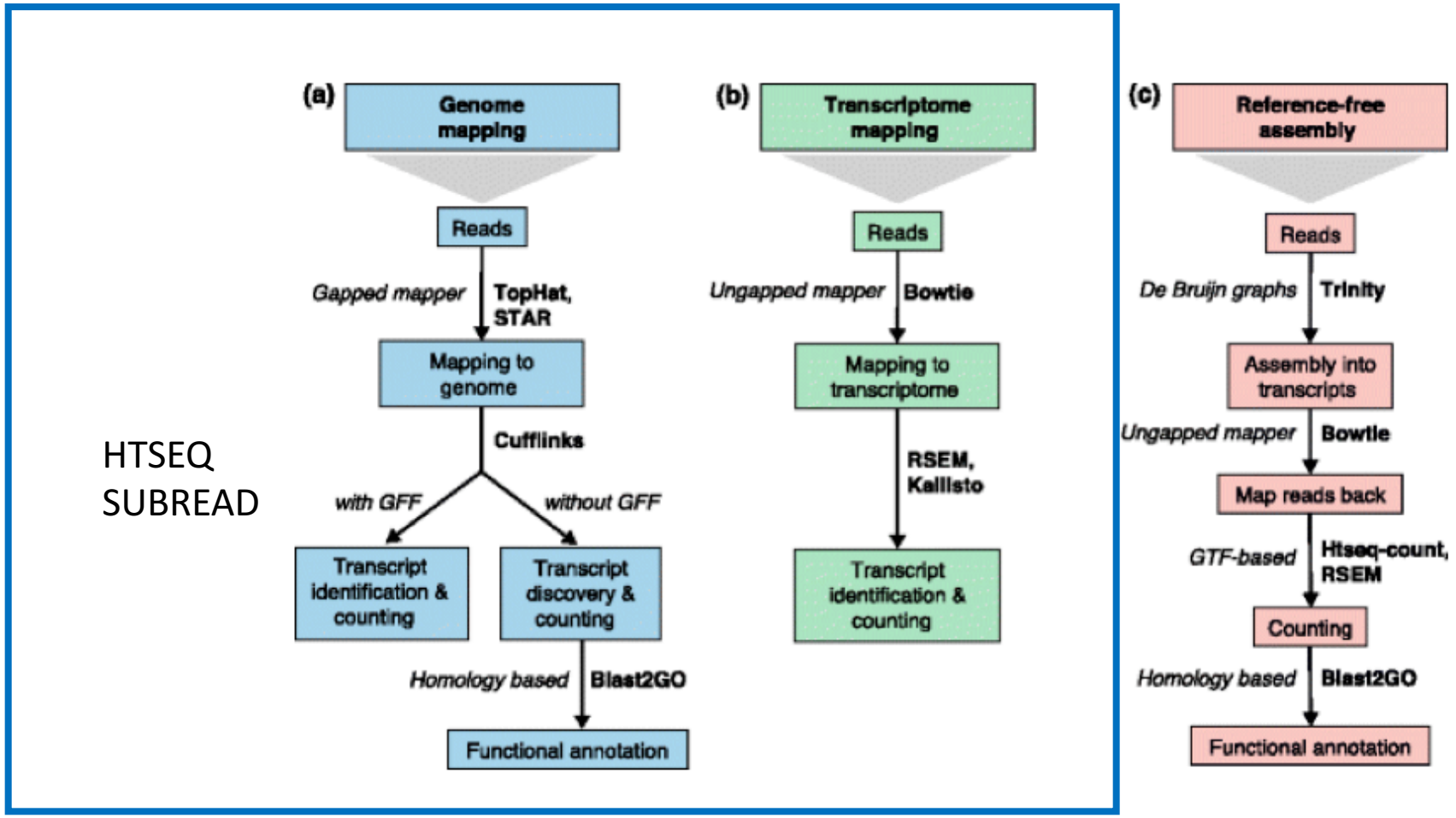


Quality control (QC) of raw data

- Detect issues related to Sample Collection, Library preparation or Sequencing
- Need to check
  - Base quality score
  - sequence quality
  - Sequence duplication level
  - GC content level
  - Presence of contaminants
    - bacteria or virus
    - Adaptor presence

Alignment & quantification

HTSEQ
SUBREAD



Post-alignment QC

- % mapped and uniquely mapped reads: 70-90%
- uniformity of read coverage over gene body
- Read distribution
- Check for read strandedness
- Biotype composition (check for rRNA)

Differential expression analysis

- What are the genes or transcripts that are differentially expressed between two or more groups?
  - do statistical test:
    - T-test
    - Empirical Bayes (moderated t-test)
    - Anova (> 2 groups)
    - …
  - adjust for multiple testing (FDR….)

Known differentially expression detection methods

Table I: Software packages for detecting differential expression

Method	Version	Reference	Normalization[a]	Read count distribution assumption	Differential expression test
edgeR	3.0.8	[4]	TMM/Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	[5]	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	[6]	Scaling factors (quantile/TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	[7]	RPKM/TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	[8]	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	[9]	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff 2 (Cufflinks)	2.0.2-beta	[10]	Geometric (DESeq-like)/quartile/classic-fpkm	Beta negative binomial distribution	*t*-test
EBSeq	1.1.7	[11]	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods

[a]In case of availability of several normalization methods, the default one is underlined.

Comparison of software packages for detecting differential expression in RNA-seq studies

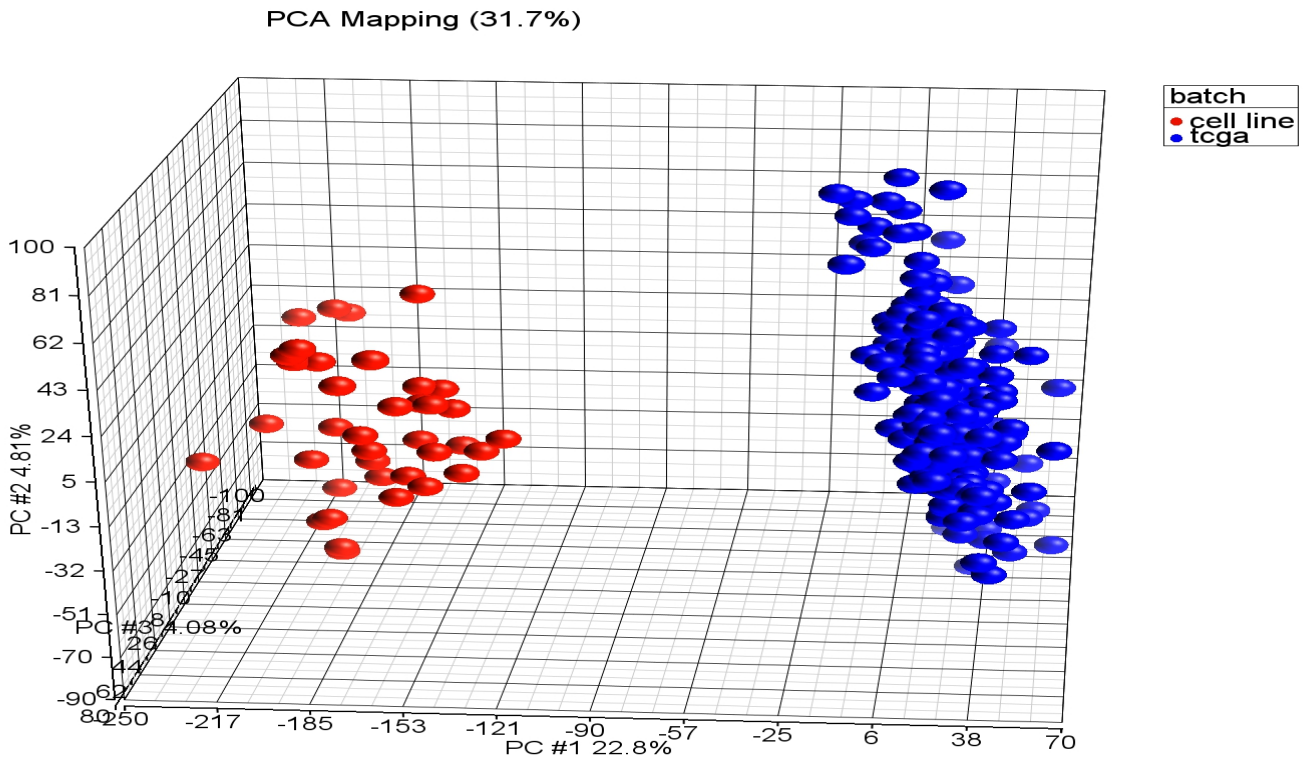
Briefings in Bioinformatics vol 16 N0I. 59-70

Normalization using scaling methods: overall gene expression is same across all samples

Method	Description
Total count (TC):	Gene counts are divided by the total number of mapped reads (or library size) associated with their sample and multiplied by the mean total count across all the samples of the dataset
Upper Quartile (UQ):	Very similar in principle to TC, the total counts are replaced by the upper quartile of counts different from 0 in the computation of the normalization factors
Median (Med):	Also similar to TC, the total counts are replaced by the median counts different from 0 in the computation of the normalization factors
DESeq	A scaling factor for a given sample is the median of the ratio, for each gene, of its read count over its geometric mean across all samples
Trimmed Mean of M-values (TMM)	A scaling factor is computed as the weighted mean of log ratios between the sample and the reference, after exclusion of the most expressed genes and the genes with the largest log ratios

Principal Component Analysis

- Method for dimension reduction to identify patterns
  (thousands of genes = thousands of dimensions)

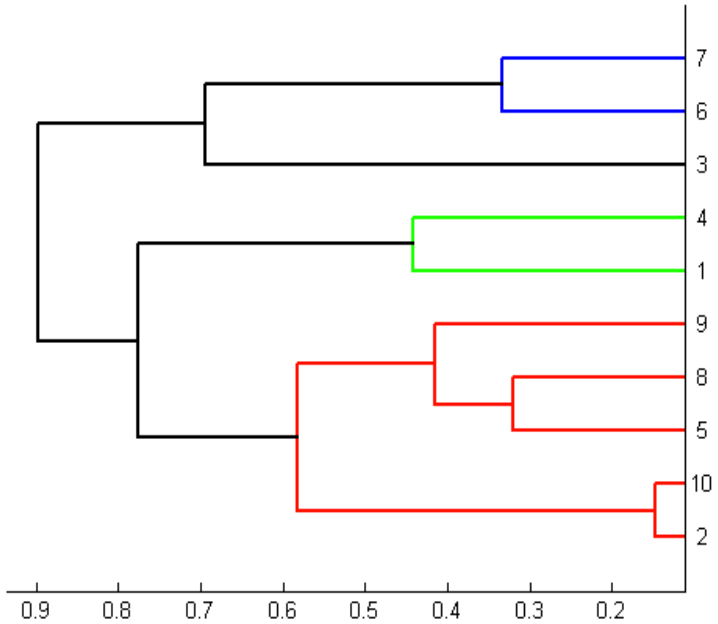


<u>The eigenvector with the largest eigenvalue (total variance)</u> is the first principal component.
The second largest eigenvalue will be the direction of the second largest variance.

Hierarchical Clustering

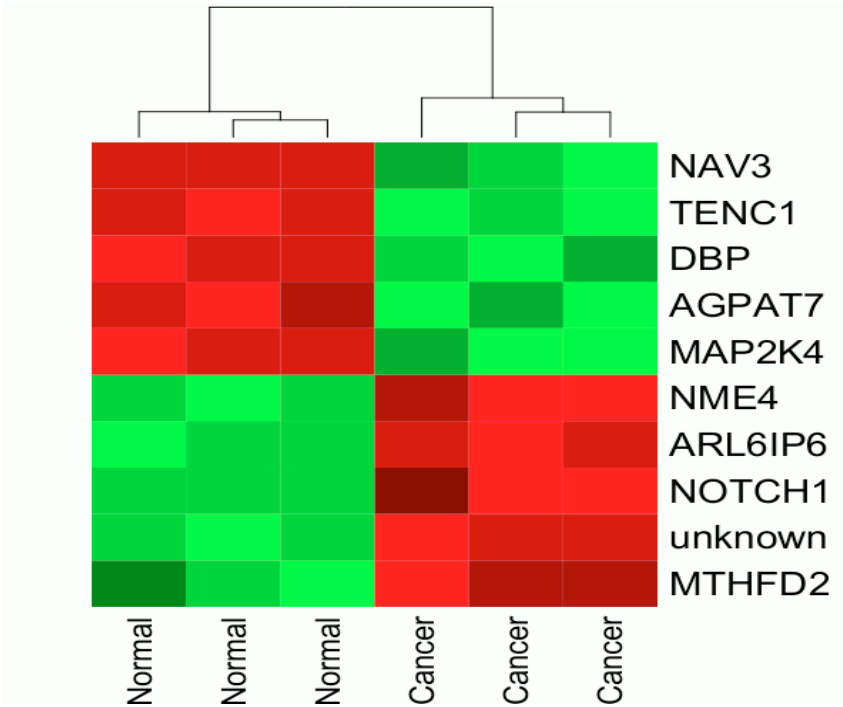
Dendrogram/tree

- branching diagram representing a hierarchy of categories based on degree of similarity



root branches leaves

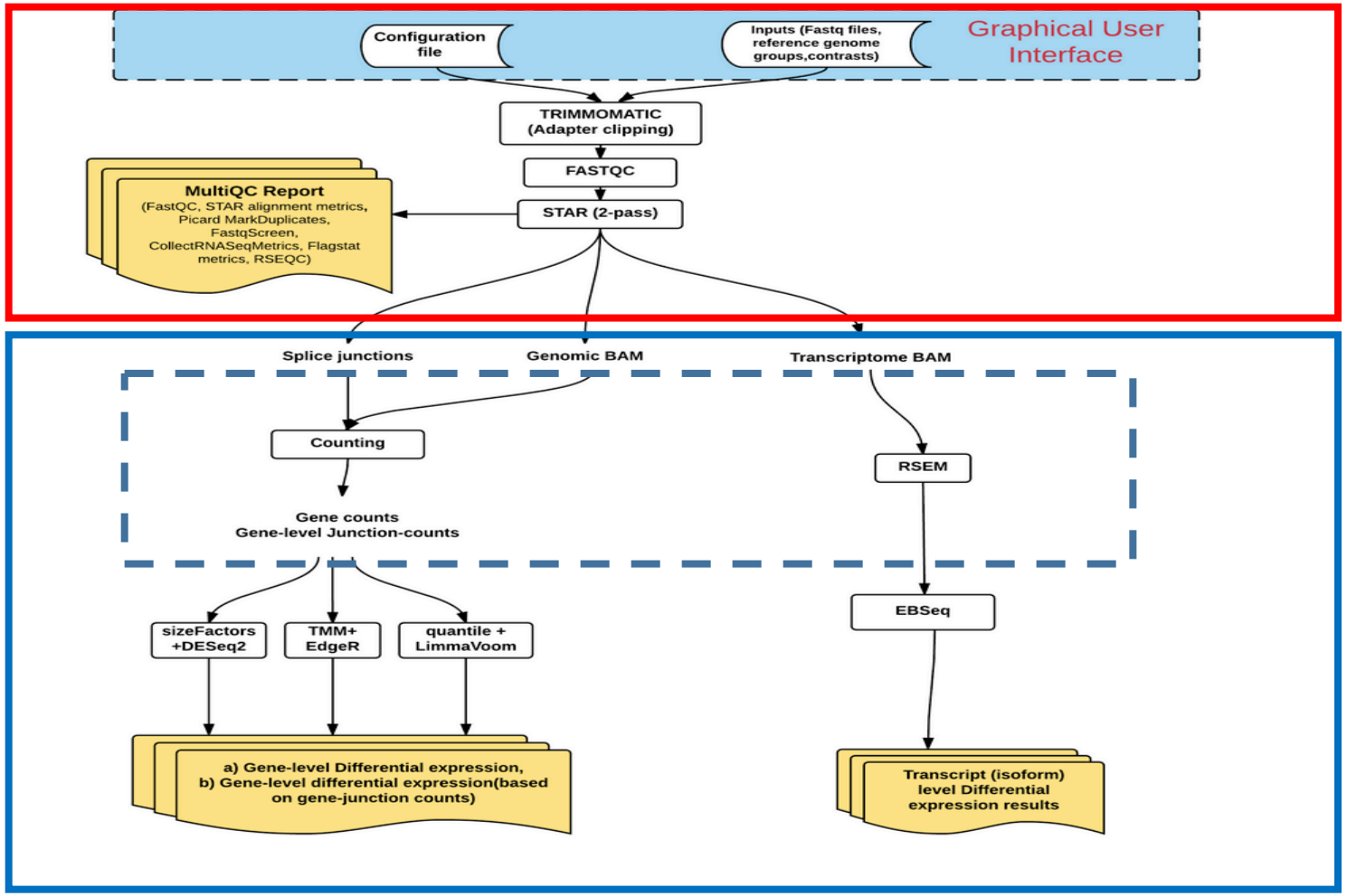
Heatmap



Agenda

- Introduction
- Data analysis Workflow
  - Review main steps
- CCBR RNA-Seq pipeline
  - Workflow overview
  - Quality Control reports
  - Principal Component Analysis PCA and differential expressed reports reports
  - Downstream analysis after running the pipeline
- Running the CCBR pipeline
  - Use case and demo

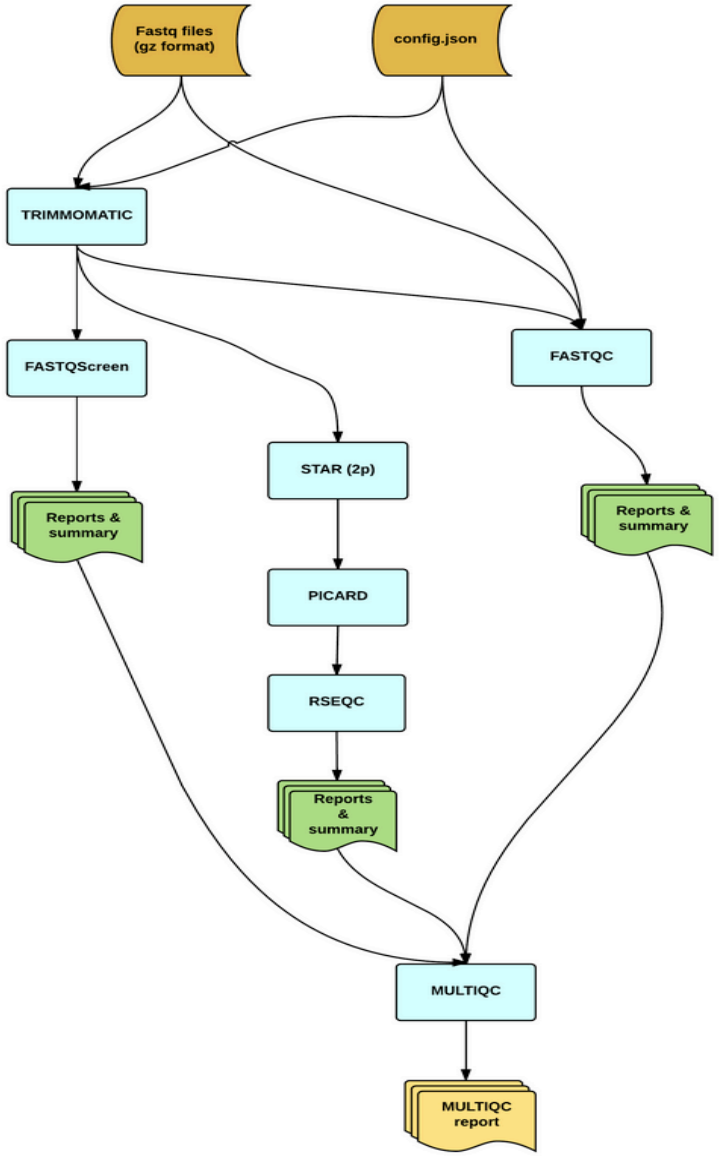
RNA-Seq Pipeline workflow



STEP1: INITIAL QC

STEP2: COUNTING & DEG

RNA-Seq: Initial QC workflow



- Trimmonatic: just adaptor clipping
- STAR2 pass mode: for most sensitive novel junctions discovery

Use case: 4 samples from SEQC study

- Mixture of biological sources and a set of synthetic RNAs from the External Rna Control Consortium (ERCC)

➢2 samples from group A : Strategene Universal Human Reference RNA (UHRR) – from 10 human cell lines-

➢2 samples from group B: Ambion Human Brain Reference RNA (HBRR)

➢Illumina HiSeq2000. -100 bp-

Base quality (Q score)

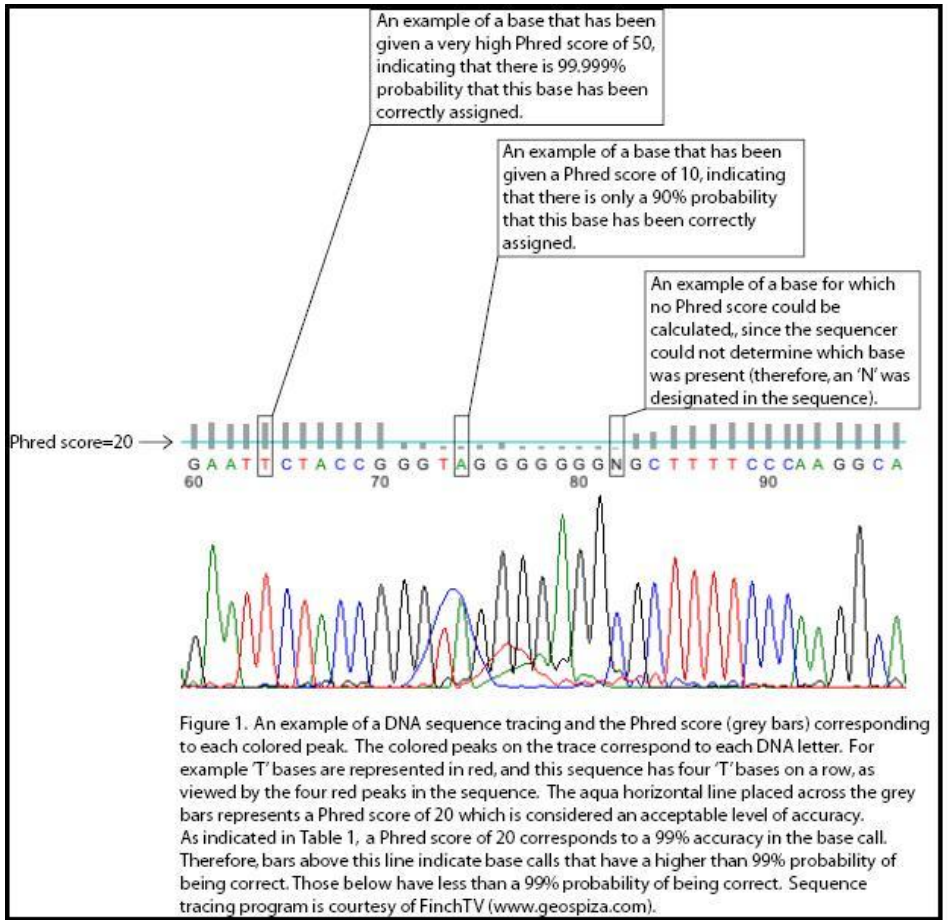


Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

$Q = -10 \log_{10} P$, where P is the base-calling error probability

Sample QC report

Base quality distribution

FastQC Report

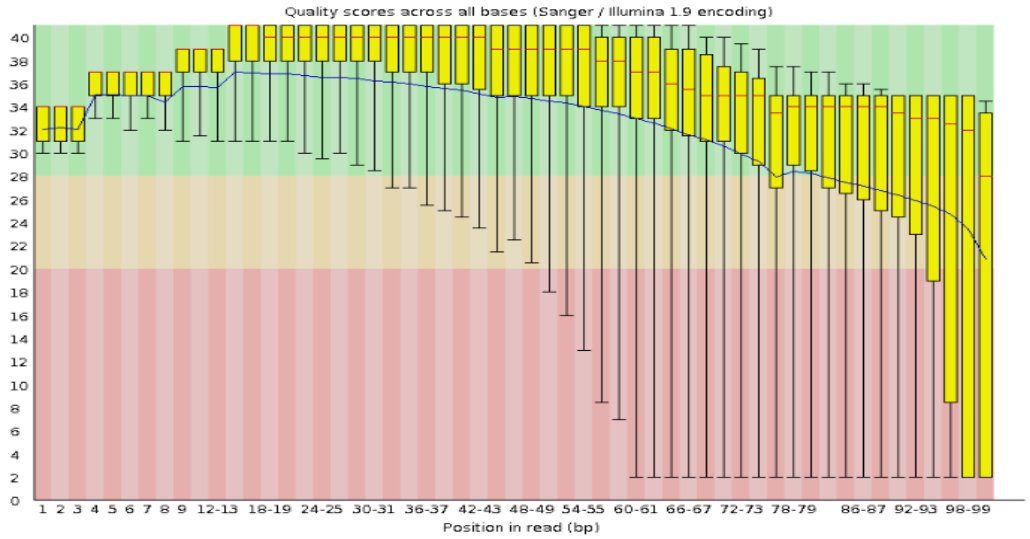
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	SRR950084.R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	125083194
Sequences flagged as poor quality	0
Sequence length	101
%GC	47

Per base sequence quality

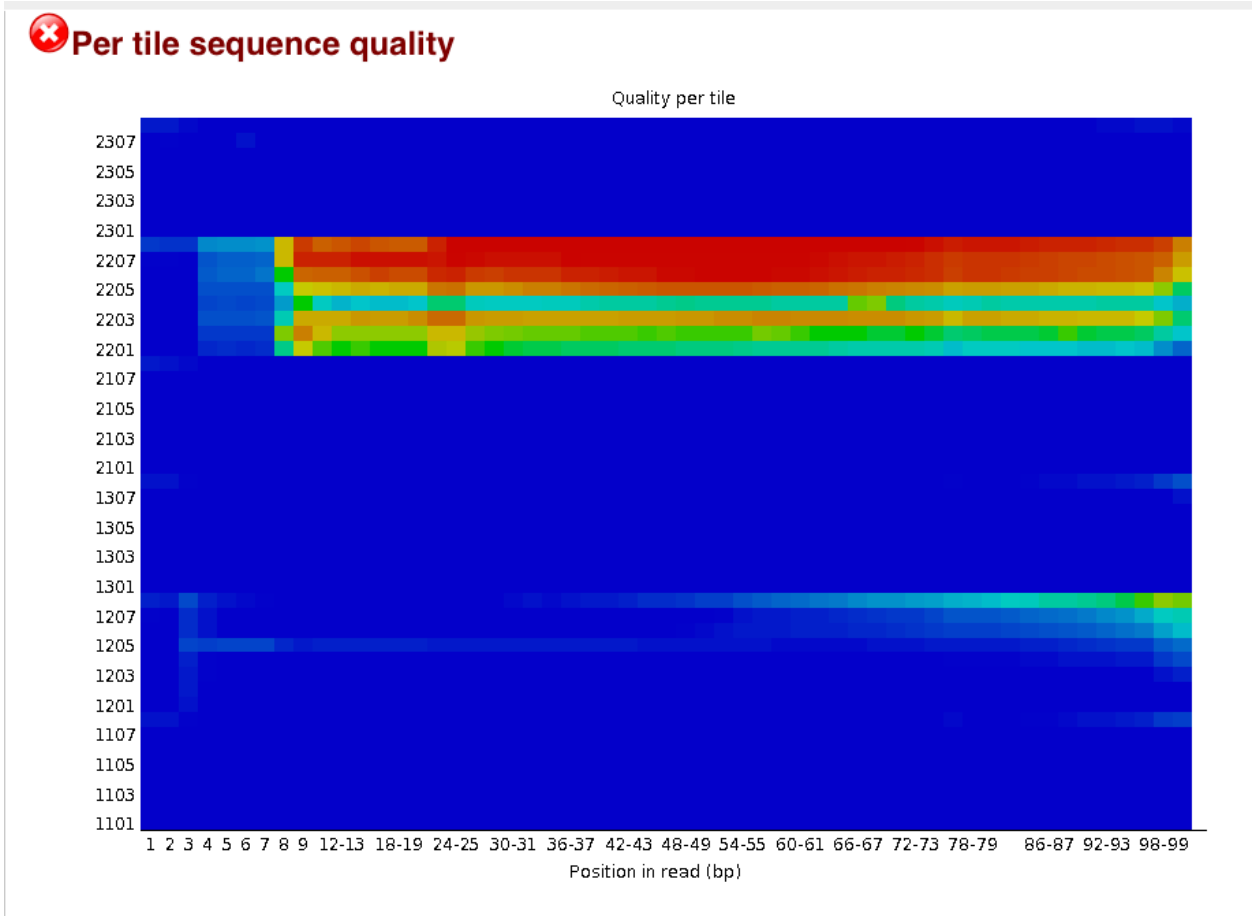


Common reasons for warnings

- General degradation of quality over the duration of long runs
- Loss quality earlier in the run (bubbles in flowcell)
- Reads of different length

Warning if the lower quartile for any base is less than 10, or if the median for any base is less than 25.
Failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

Tiles issues (bubble , smudge or debris in lane)

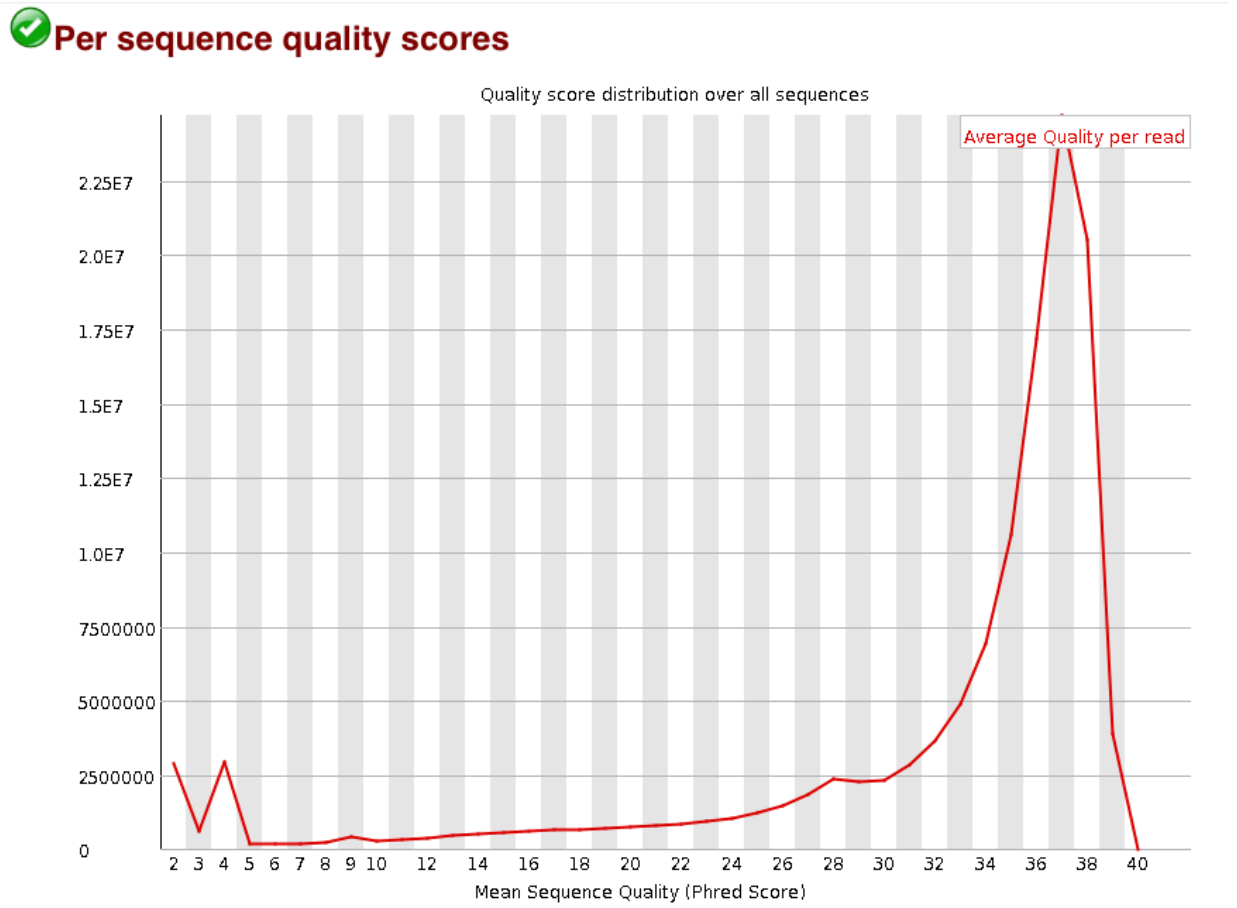


A good plot should be all blue !

Flowcell tile heatmap showing deviation from the average quality for each tile

Failure if any tile shows a mean Phred score more than 5 less than the mean for that base across all tiles

Check proportion of sequences with low quality values

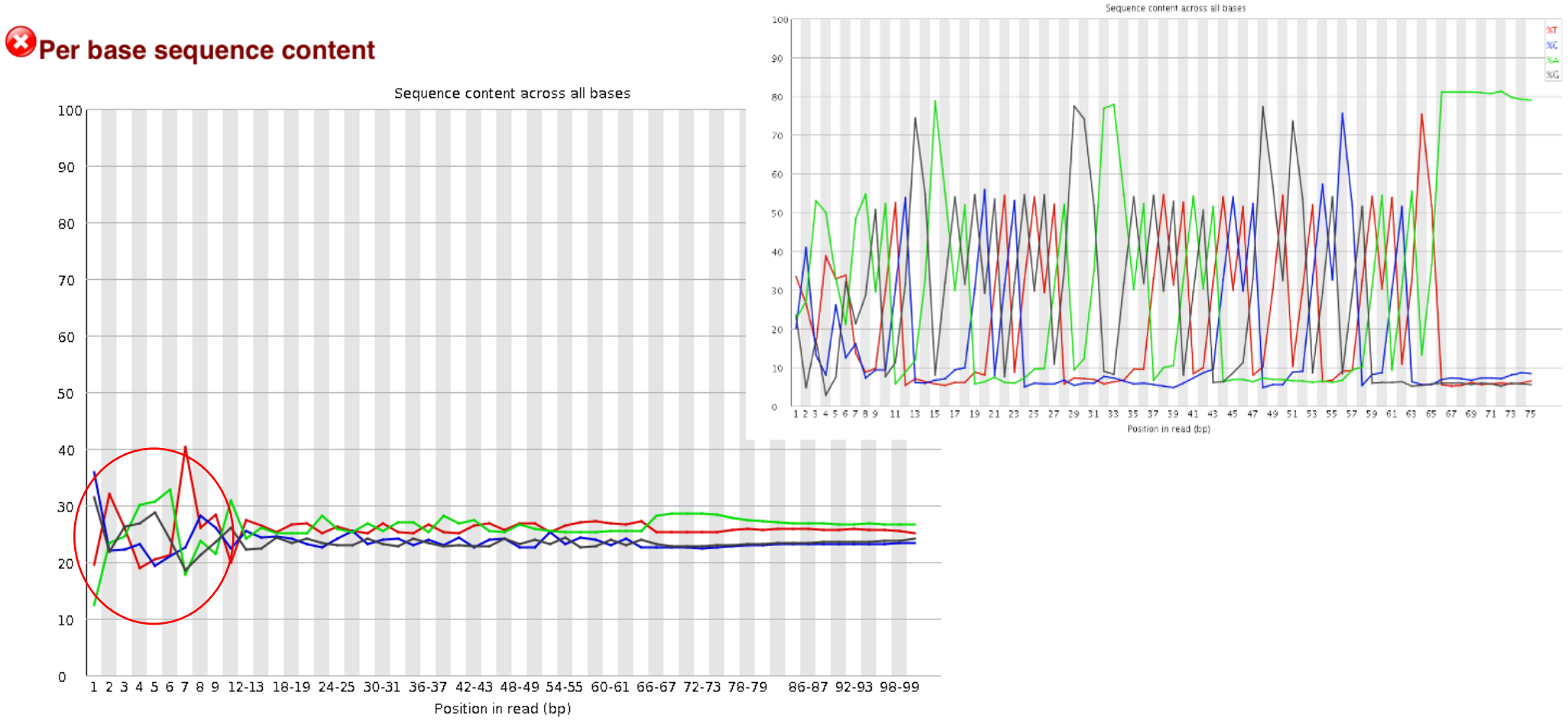


For bi-modal or complex distribution, should check with per tile qualities

Failure if the most frequently observed mean quality is below 20

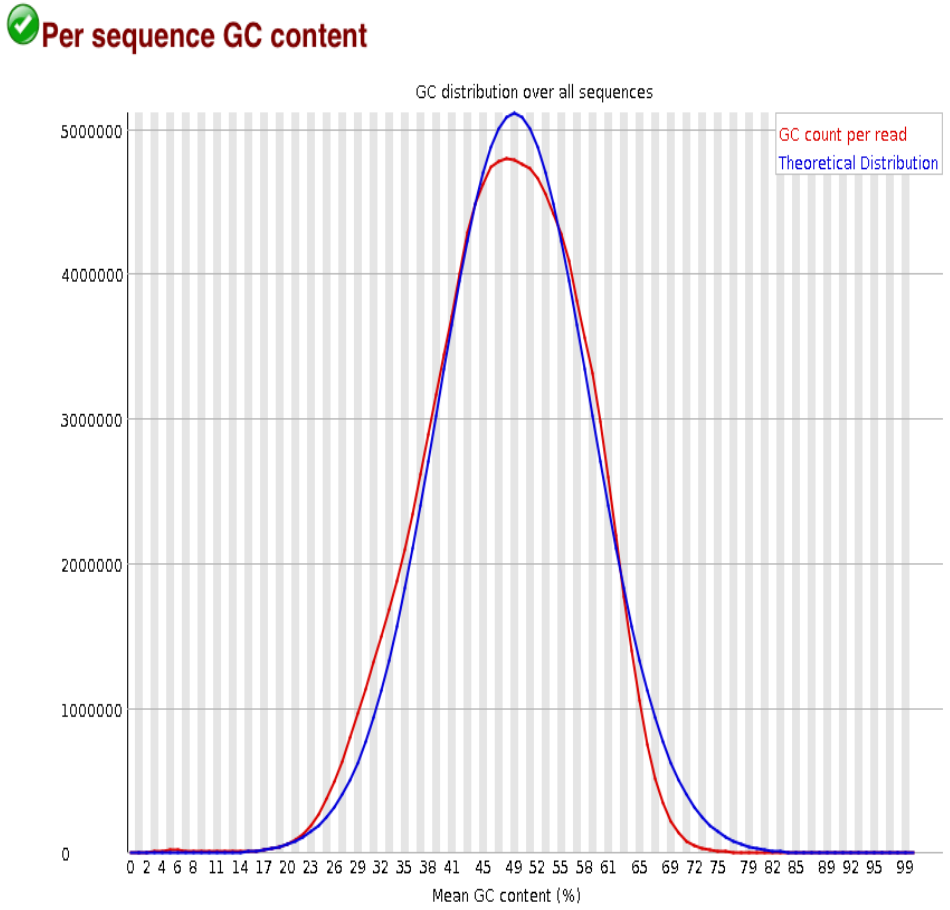
Per base sequence content should be uniform

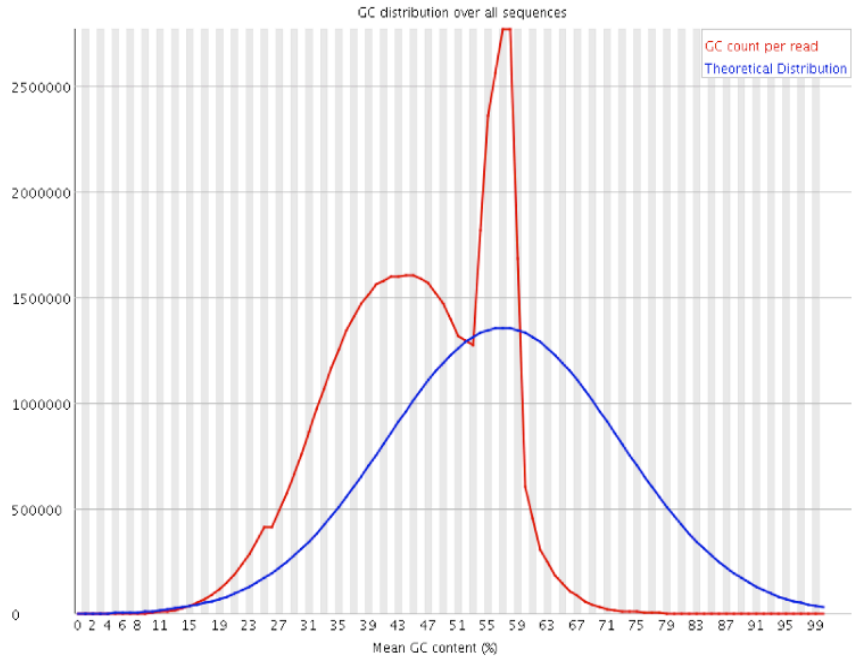
Biased fragmentation



RNA-Seq libraries produce biased sequence composition at start of the read (10-12 bp) / does not affect downstream analysis

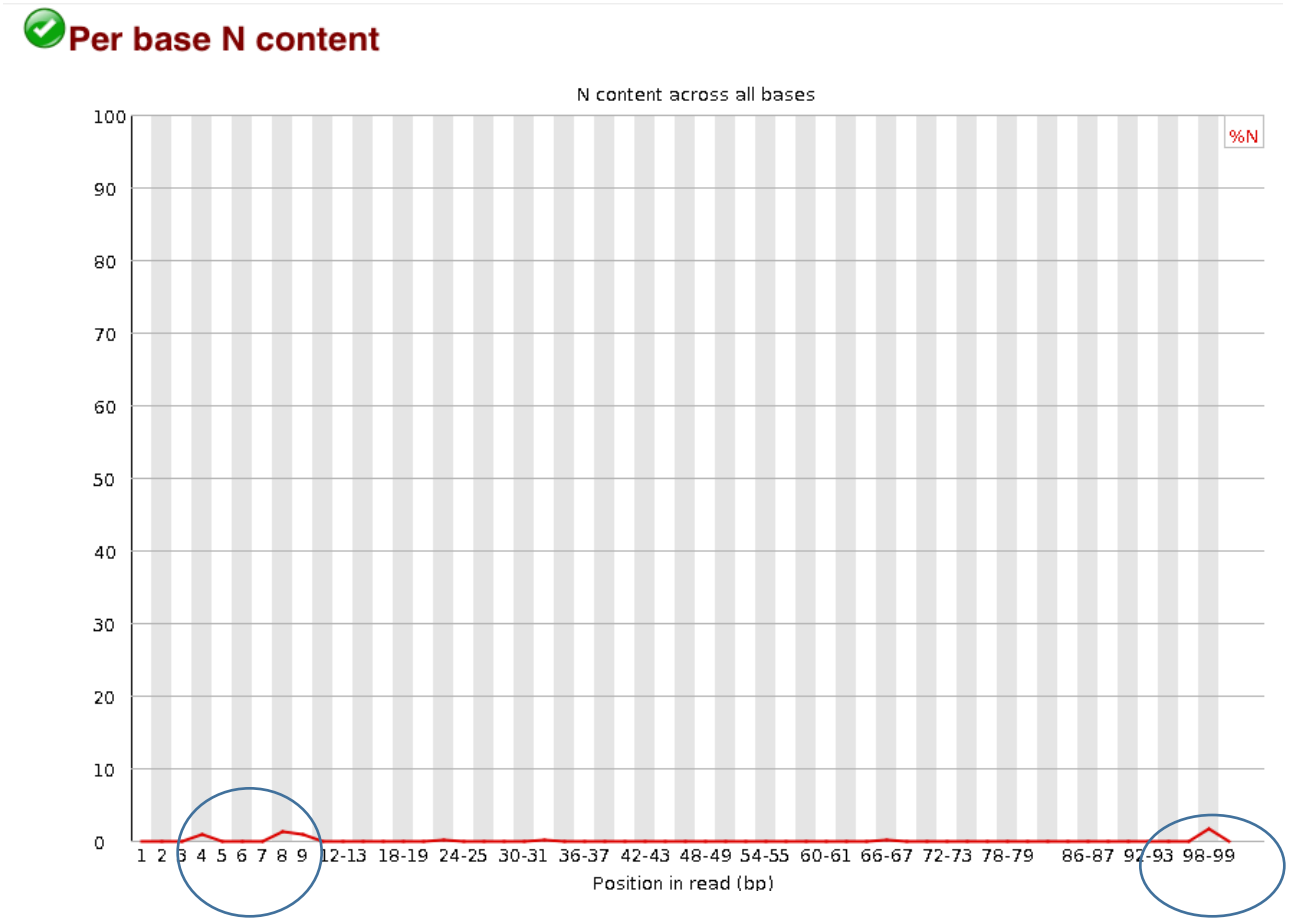
GC content should be a normal distribution





Contaminant issue (adapter dimers= paired of ligated adapters with no insert sequence)
Need to check overrepresented sequences

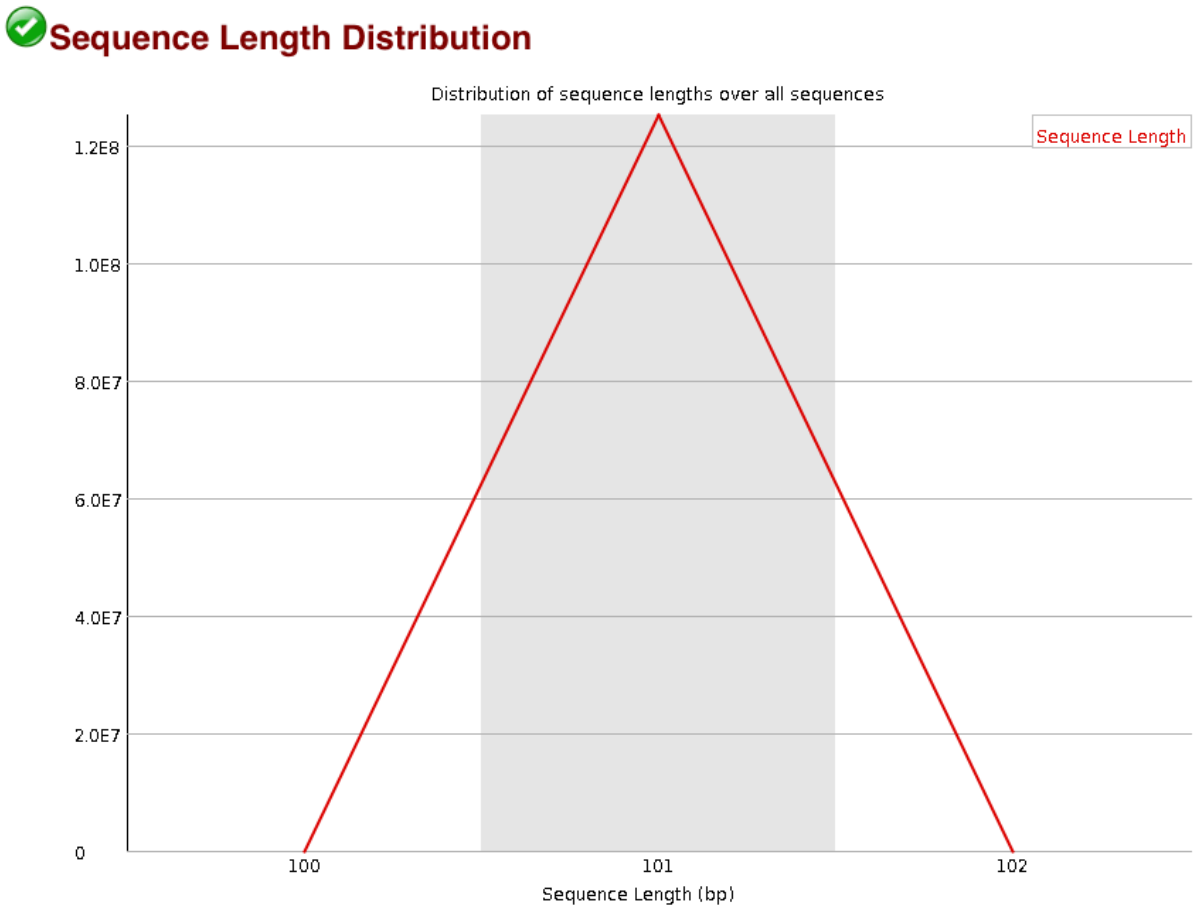
No call distribution



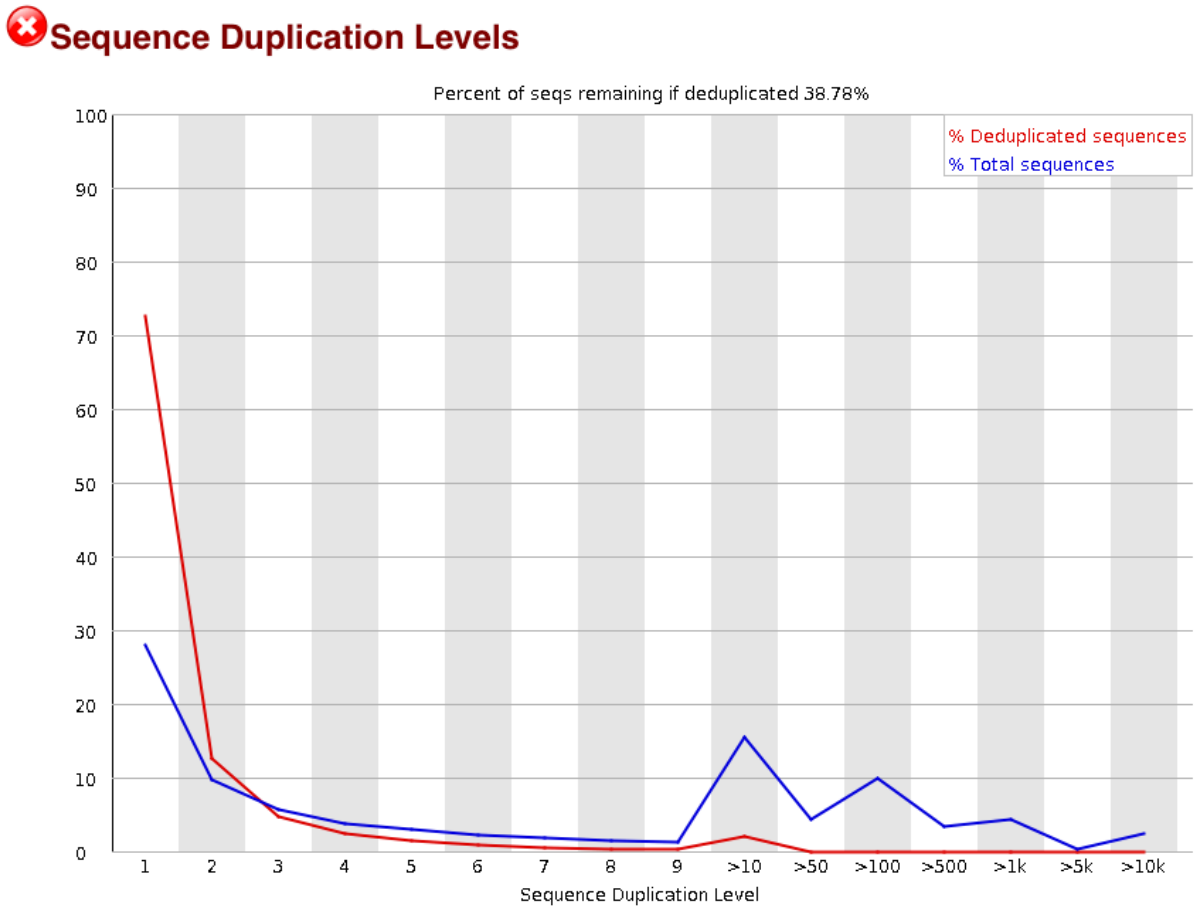
Biased sequence composition

Expected/ check with base quality

All sequences should have the same length

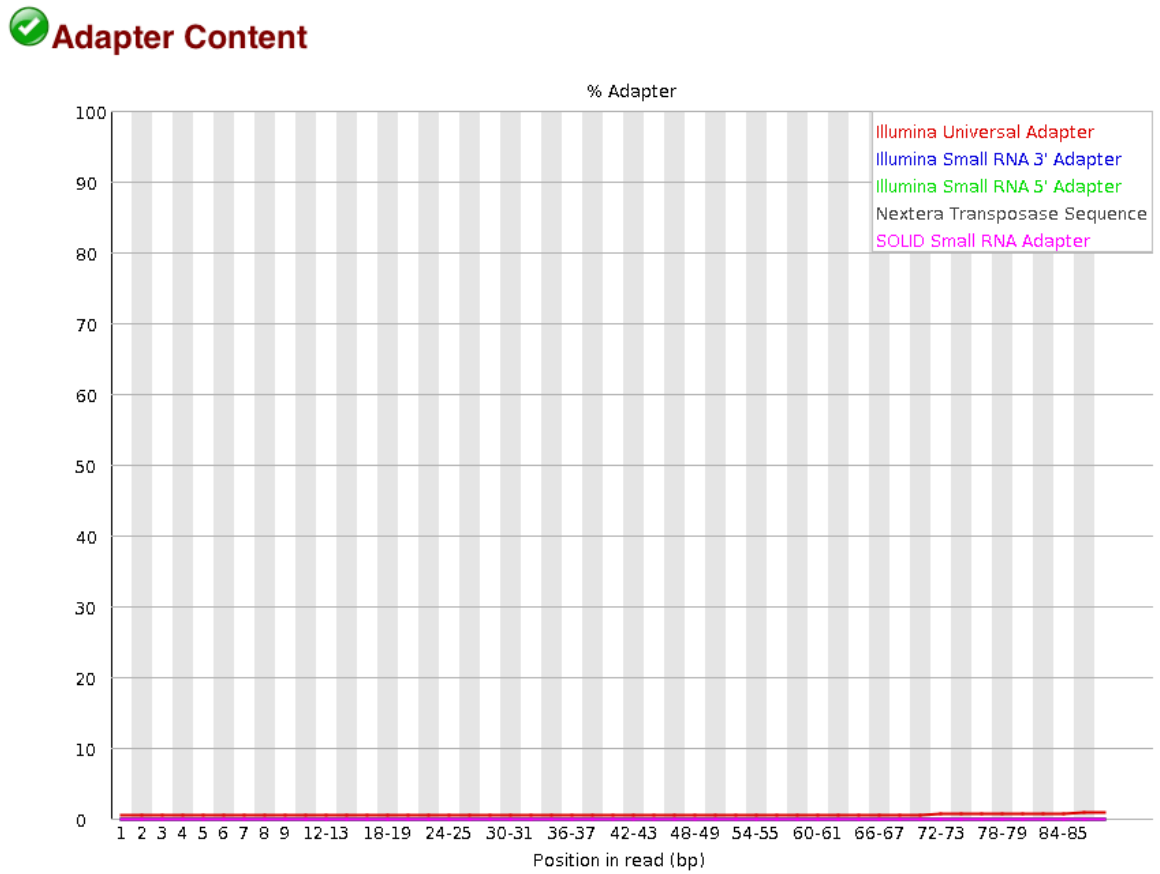


High duplication level should be carefully assessed



- Technical duplicates (PCR over amplification)
- Biological duplicates
  - Small RNA library
  - Over-sequence High expressed transcripts to observe low-expressed ones

Check for adapter sequence



If insert sizes are shorter than the read length -> need to remove adapter sequence

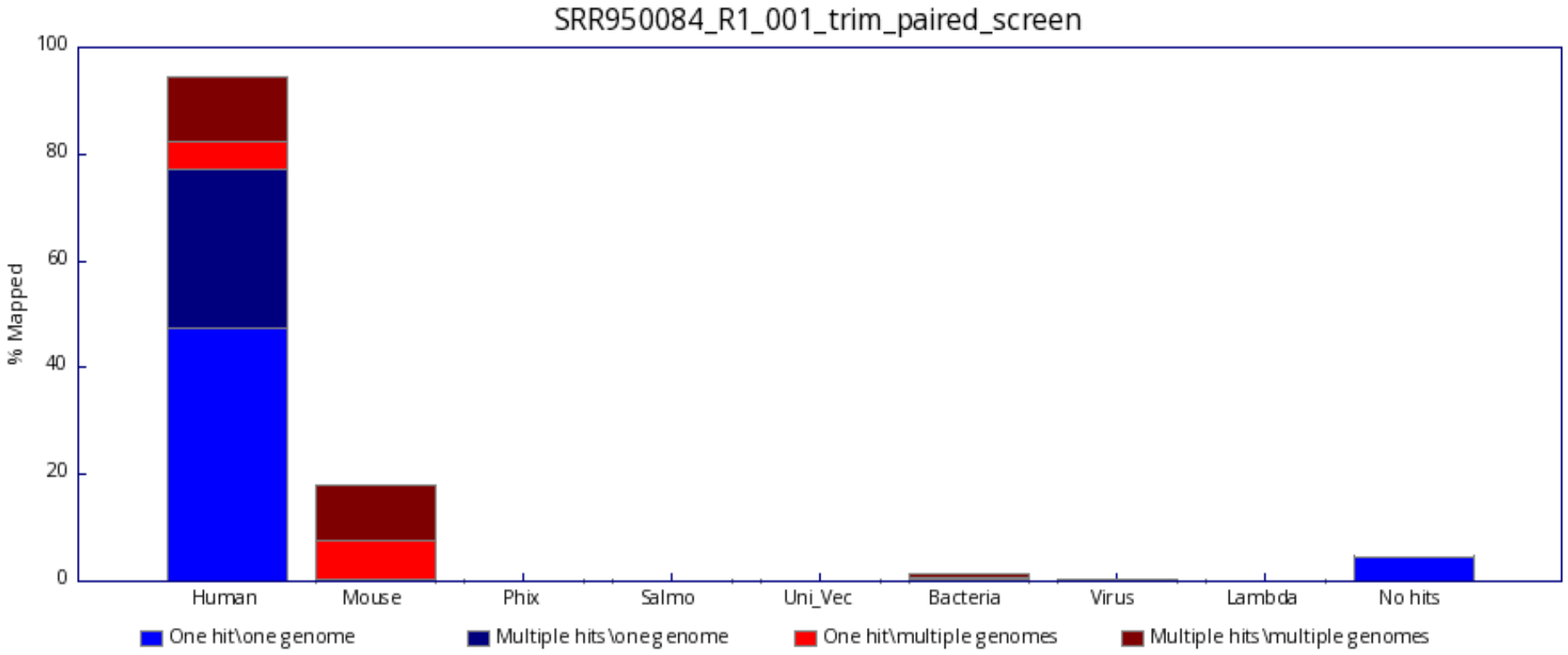
Check for contamination in Over-represented sequences:

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTCAACAATCTCGTAT	2328372	1.8614587024376752	TruSeq Adapter, Index 13 (97% over 40bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTCAACAATCTCGTA	676505	0.5408440401673785	TruSeq Adapter, Index 13 (97% over 40bp)

error if any sequence is found to represent more than 1% of the total

FastqScreen: look for Bacteria/ virus contamination



MultiQC report

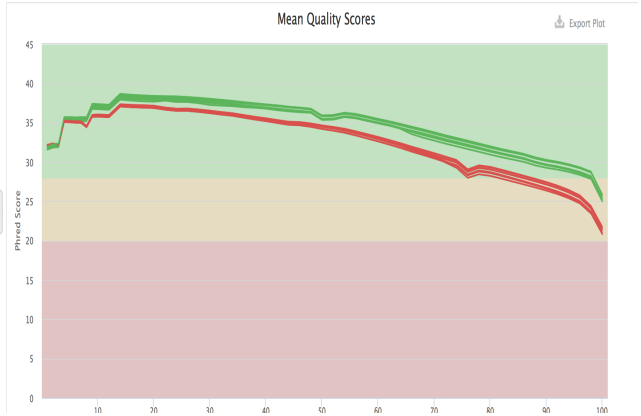
MultiQC: Multiple samples report

FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Quality Histograms 8 8

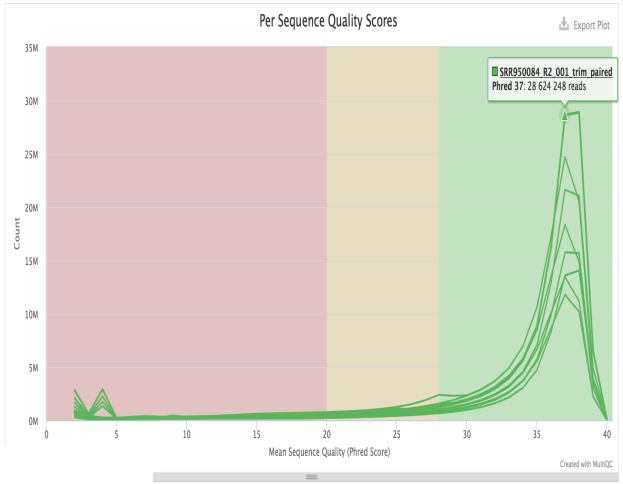
The mean quality value across each base position in the read. See the FastQC help.



Per Sequence Quality Scores 16

The number of reads with average quality scores. Shows if a subset of reads has poor quality. See the FastQC help.

Y-Limits: on

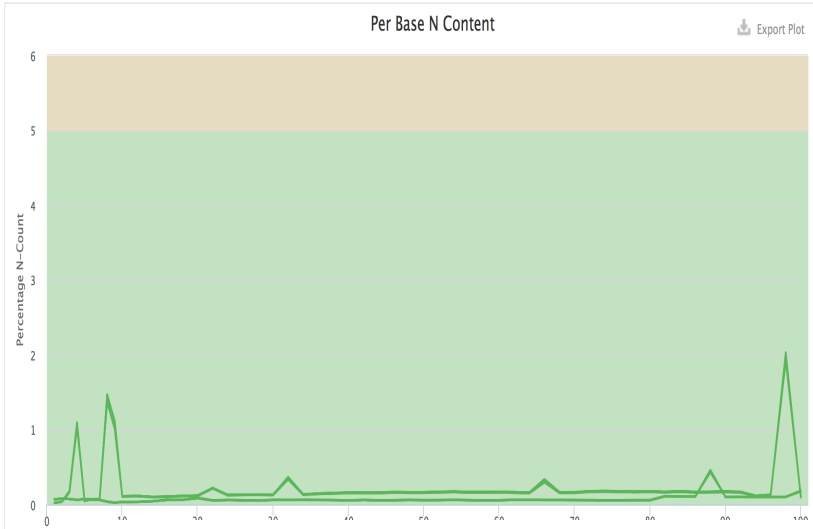


Y-Limits: on

Per Base N Content 16

The percentage of base calls at each position for which an N was called. See the FastQC help.

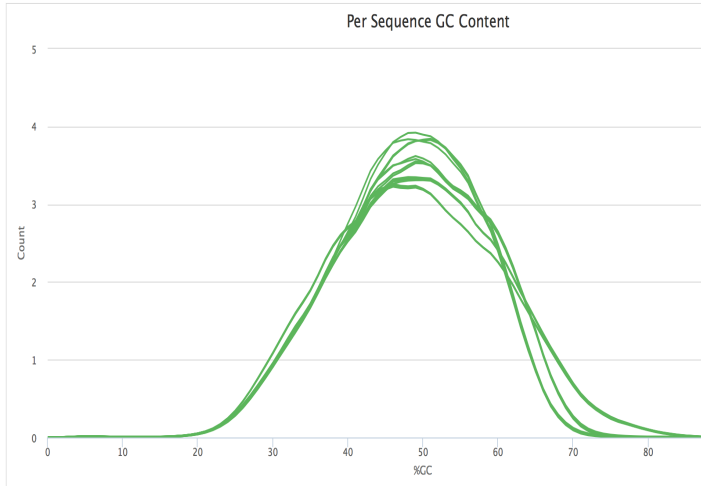
Y-Limits: on



Per Sequence GC Content 16

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content. See the FastQC help.

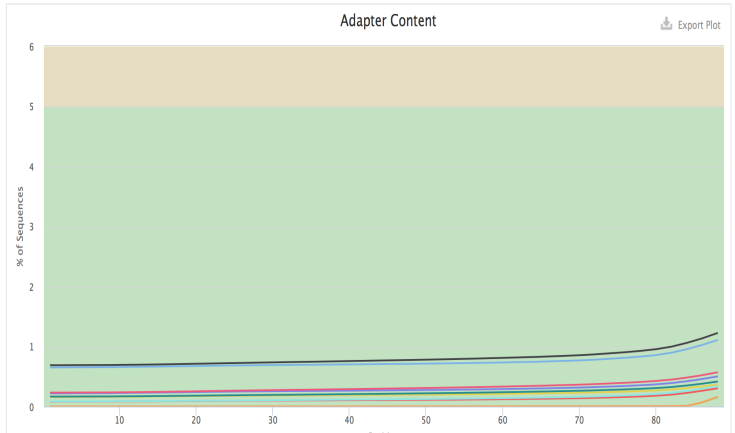
Percentages | Counts



Adapter Content 16

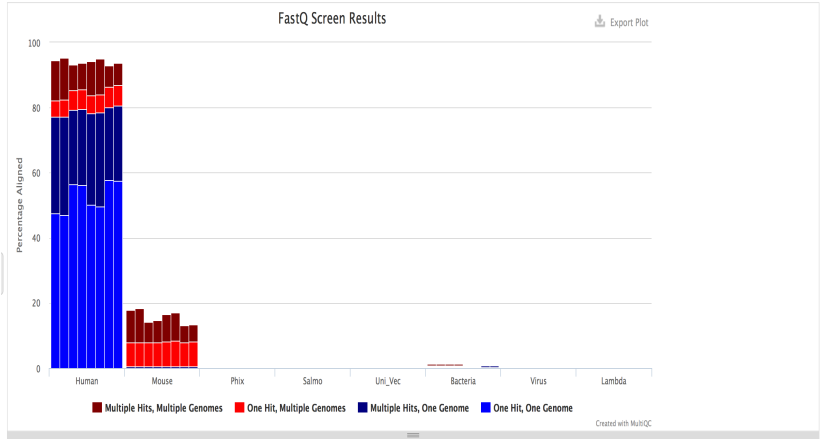
The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. See the FastQC help. Only samples with ≥ 0.1% adapter contamination are shown.

Y-Limits: on



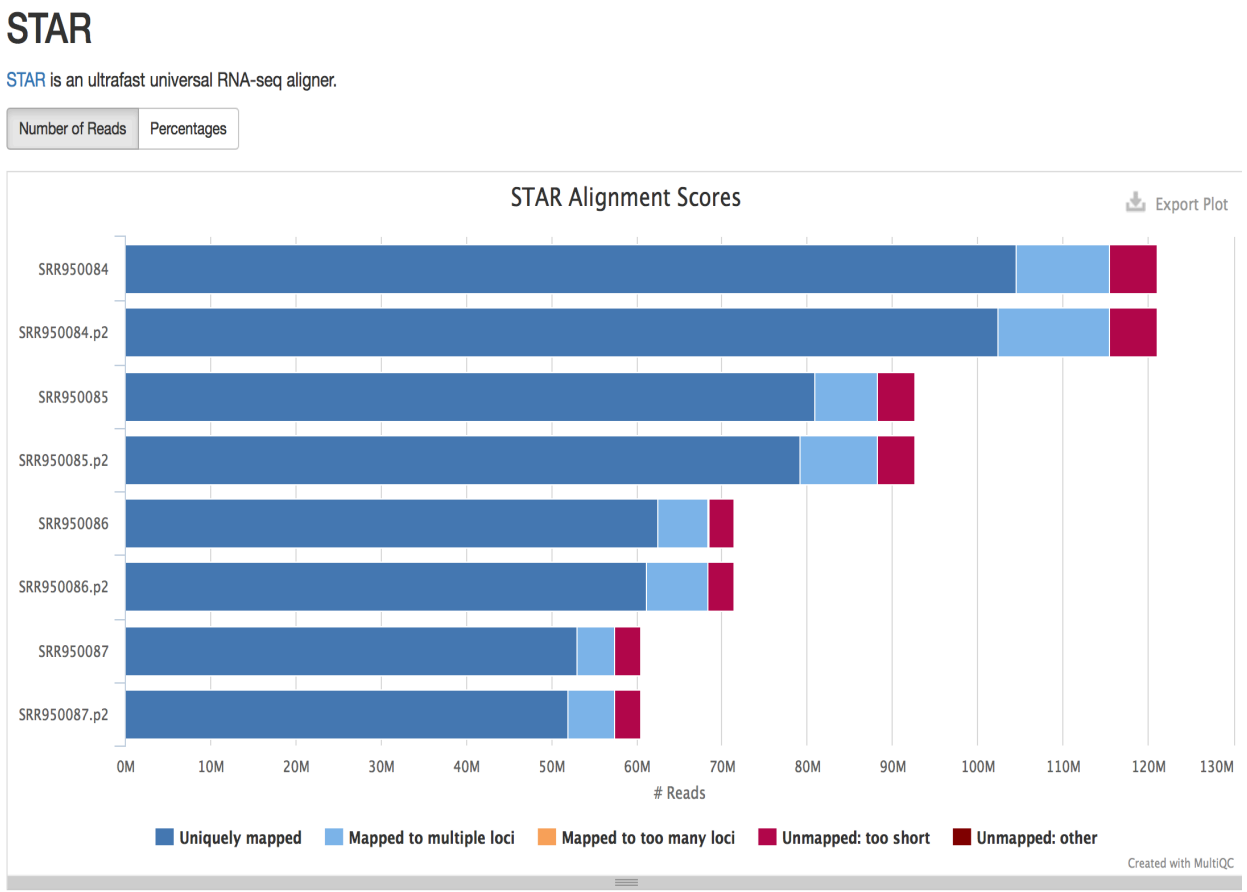
FastQ Screen

FastQ Screen allows you to screen a library of sequences in FastQ format against a set of sequence databases so you can see if the composition of the library matches with what you expect.

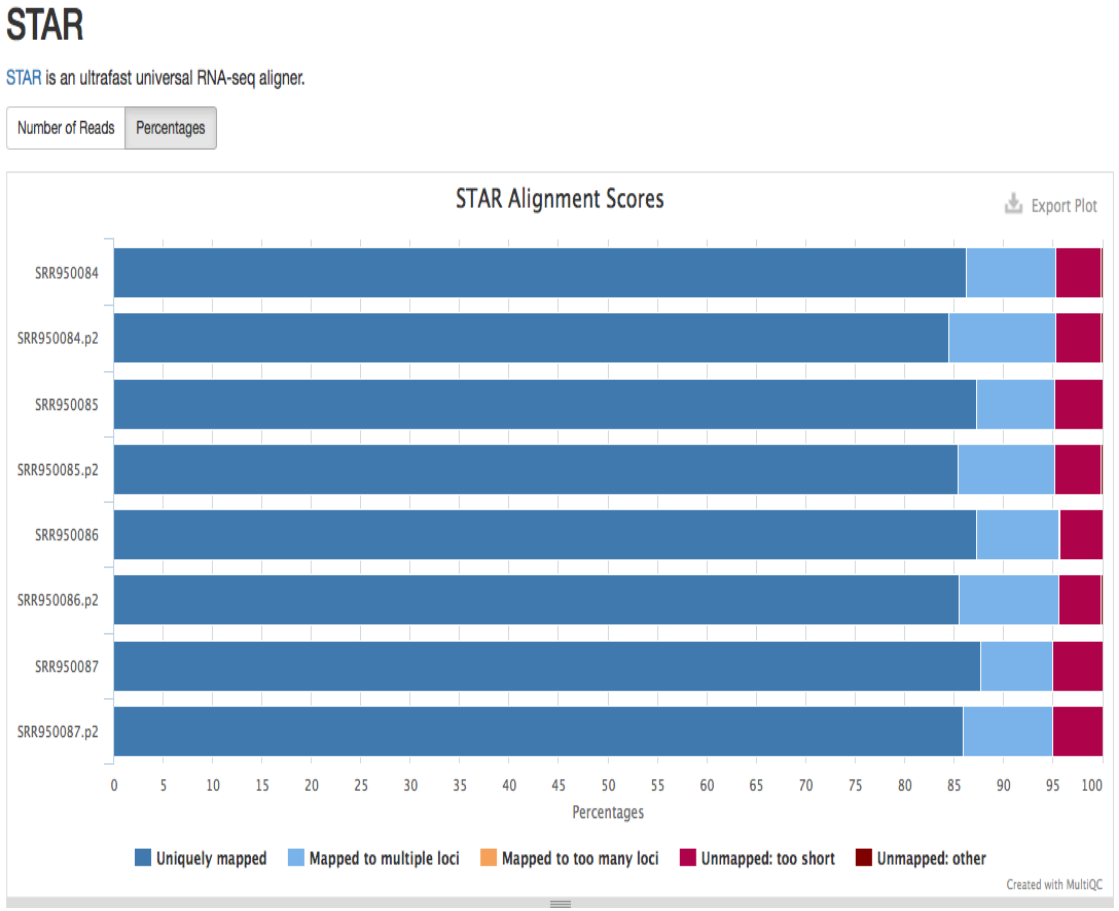


multiQC report: Mapping stats

nb.of mapped Reads



Mapping rate 70-90%



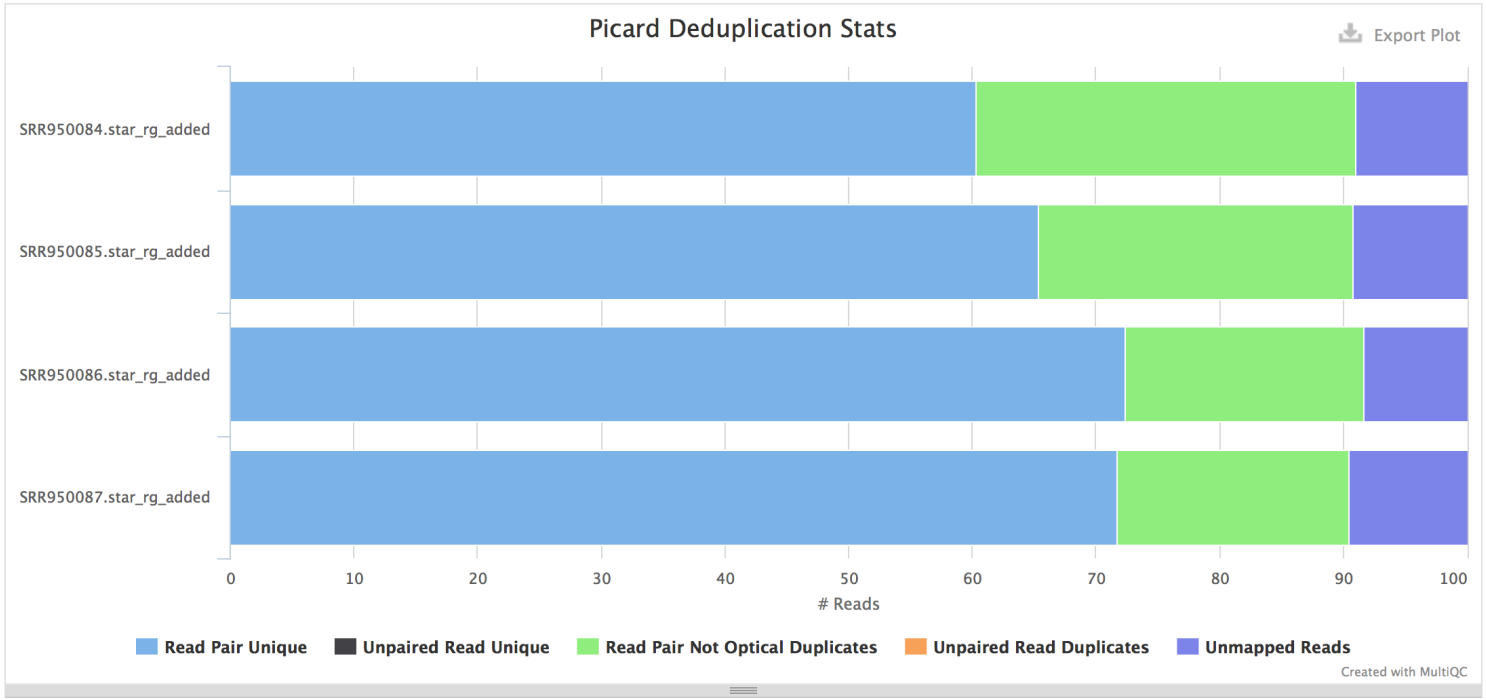
multiQC report: Picard duplication rate by paired reads

Picard

Picard is a set of Java command line tools for manipulating high-throughput sequencing data.

Mark Duplicates

Number of Reads | Percentages



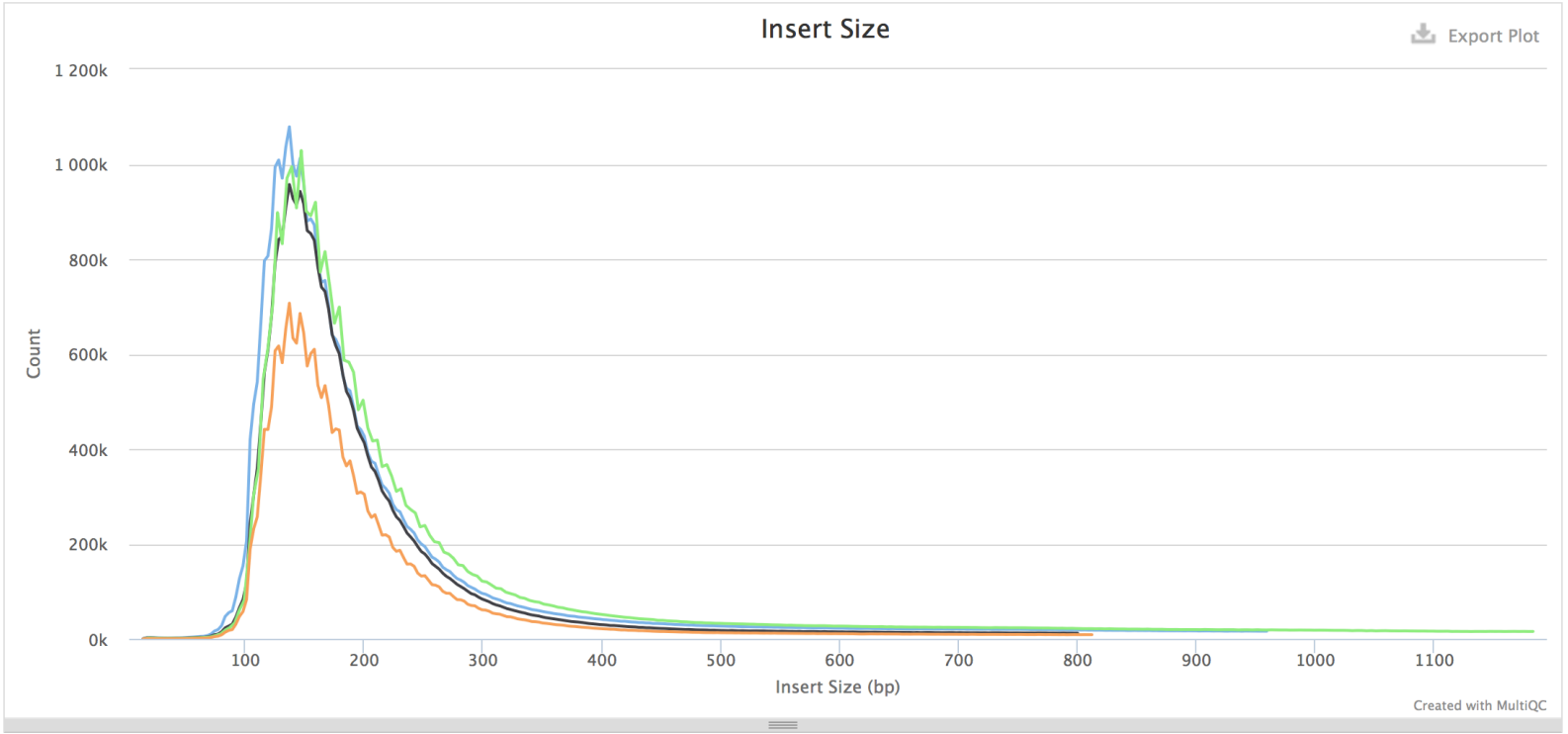
multiQC report: Picard

Insert Size

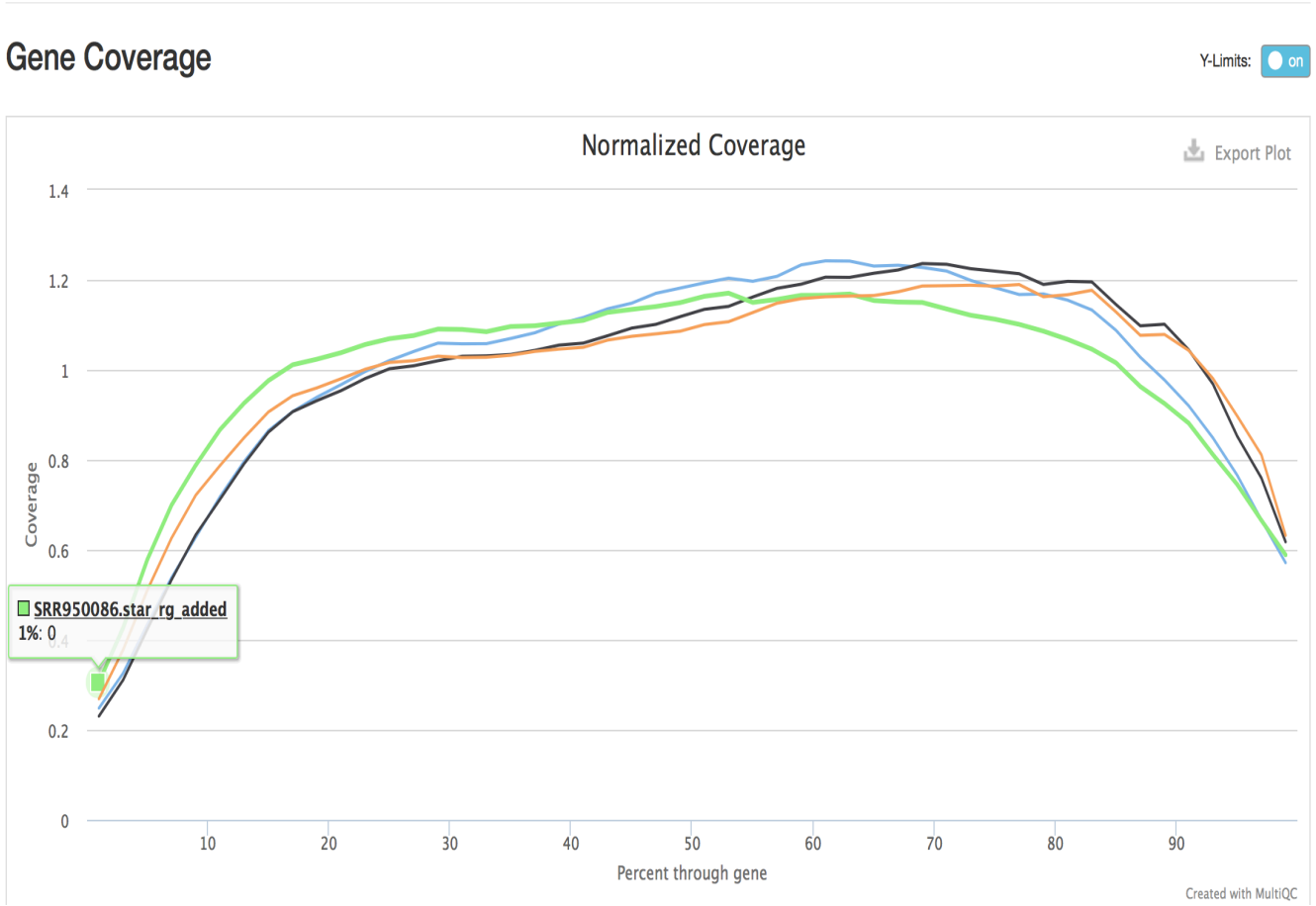
Plot shows the number of reads at a given insert size. Reads with different orientations are summed.

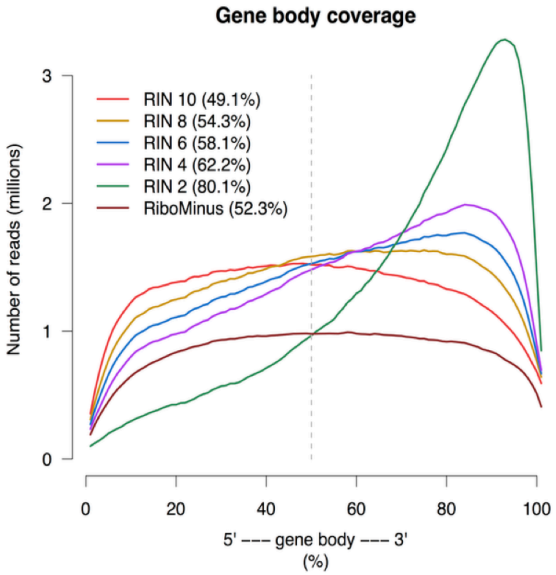
Y-Limits: on

Counts | Percentages



multiQC report: RNA quality check





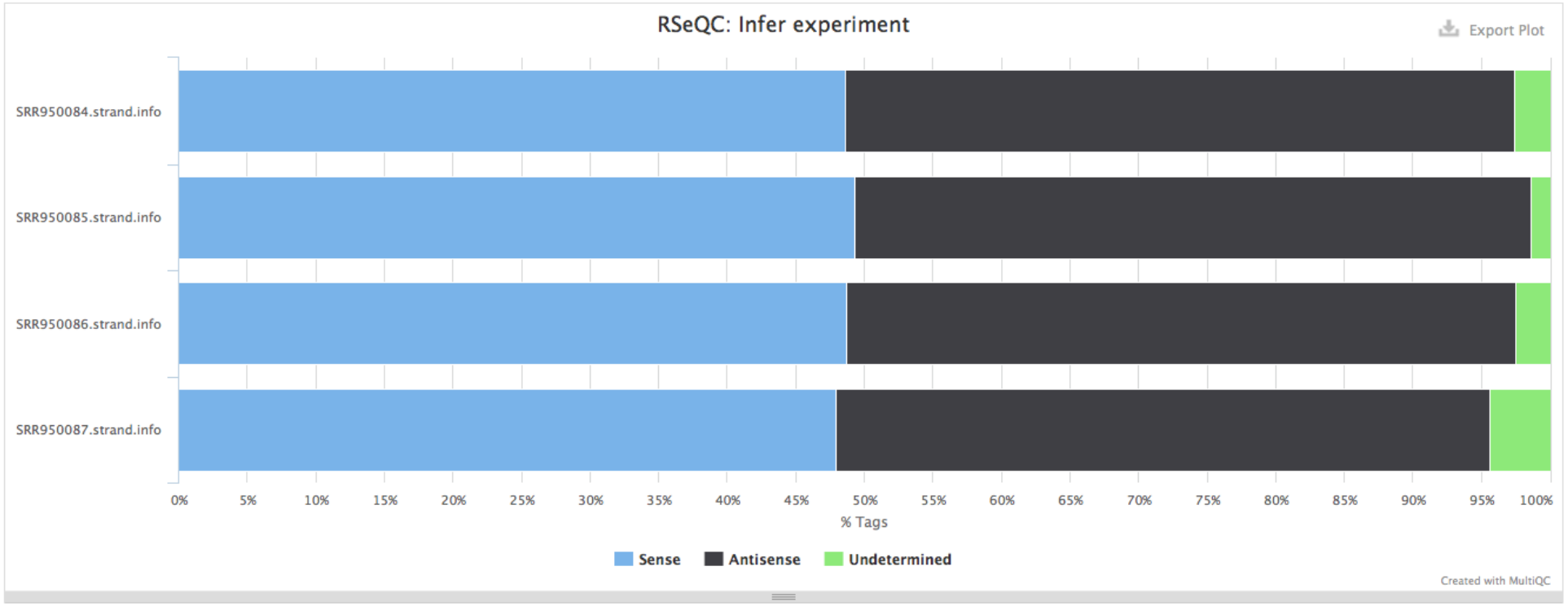
Sigurgeirsson B, Emanuelsson O, Lundeberg J (2014) Sequencing Degraded RNA Addressed by 3' Tag Counting. PLOS ONE 9(3): e91851. doi:10.1371/journal.pone.0091851
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0091851

Degraded RNA showing 3’ bias in coverage

multiQC report: RSEQC

Infer experiment

Infer experiment counts the percentage of reads and read pairs that match the strandedness of overlapping transcripts. It can be used to infer whether RNA-seq library preps are stranded (sense or antisense) .



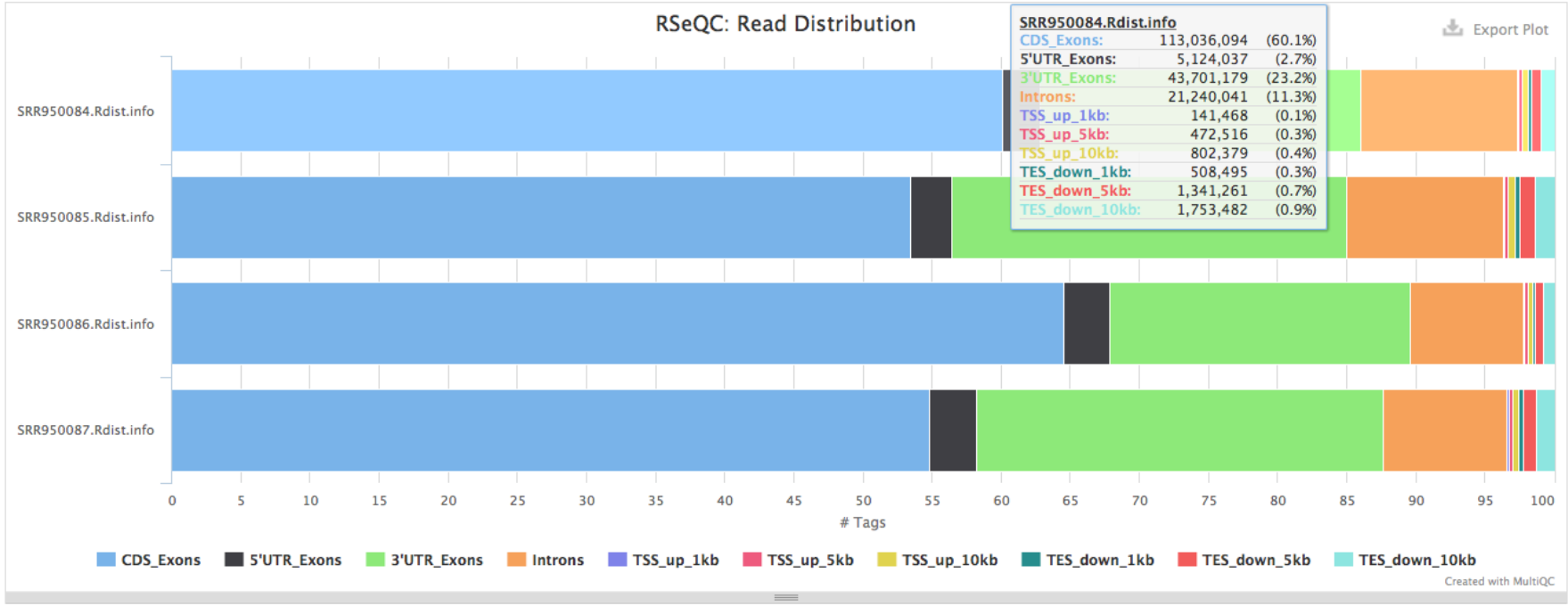
multiQC report: Exons coverage

RSeQC package provides a number of useful modules that can comprehensively evaluate high throughput RNA-seq data.

Read Distribution

Read Distribution calculates how mapped reads are distributed over genome features.

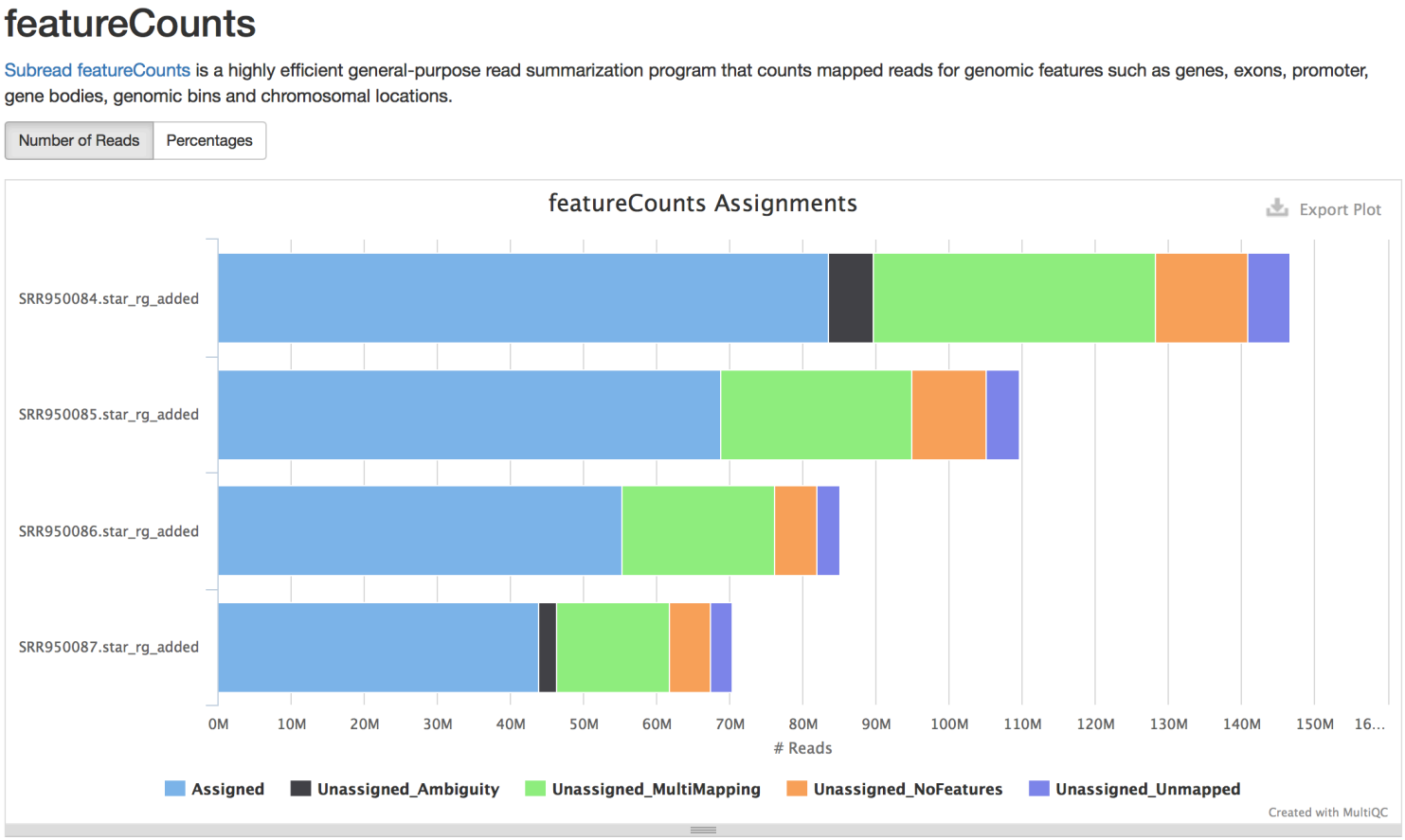
Number of Tags | Percentages



SRR950084.Rdist.info

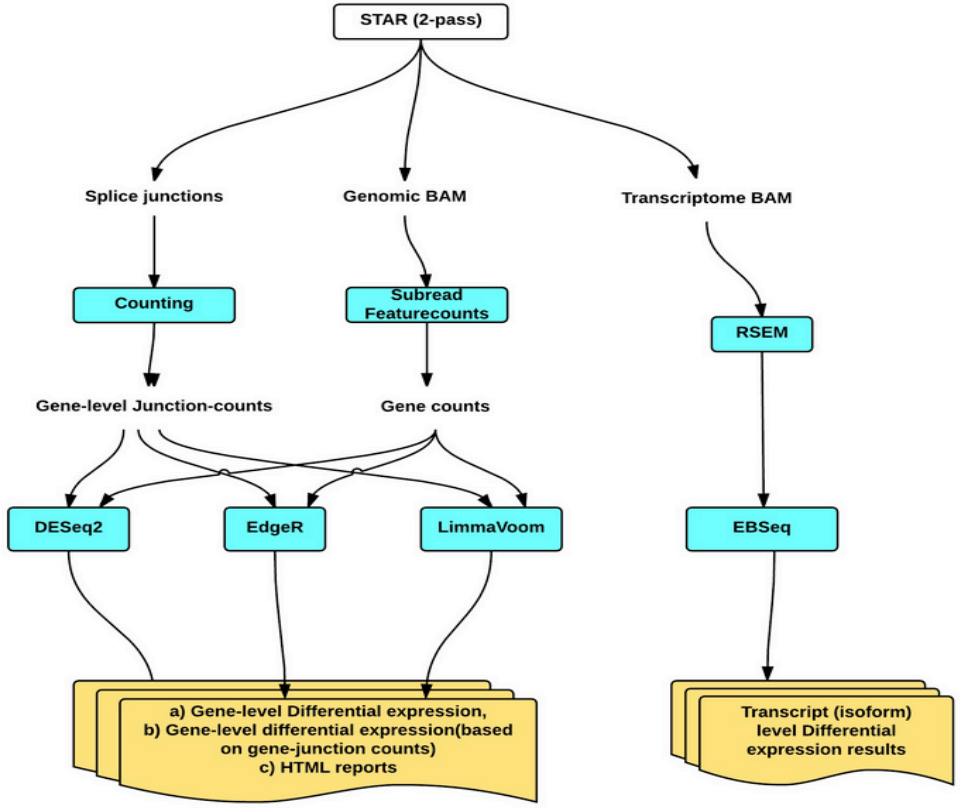
CDS_Exons:	113,036,094	(60.1%)
5'UTR_Exons:	5,124,037	(2.7%)
3'UTR_Exons:	43,701,179	(23.2%)
Introns:	21,240,041	(11.3%)
TSS_up_1kb:	141,468	(0.1%)
TSS_up_5kb:	472,516	(0.3%)
TSS_up_10kb:	802,379	(0.4%)
TES_down_1kb:	508,495	(0.3%)
TES_down_5kb:	1,341,261	(0.7%)
TES_down_10kb:	1,753,482	(0.9%)

multiQC report: Count check

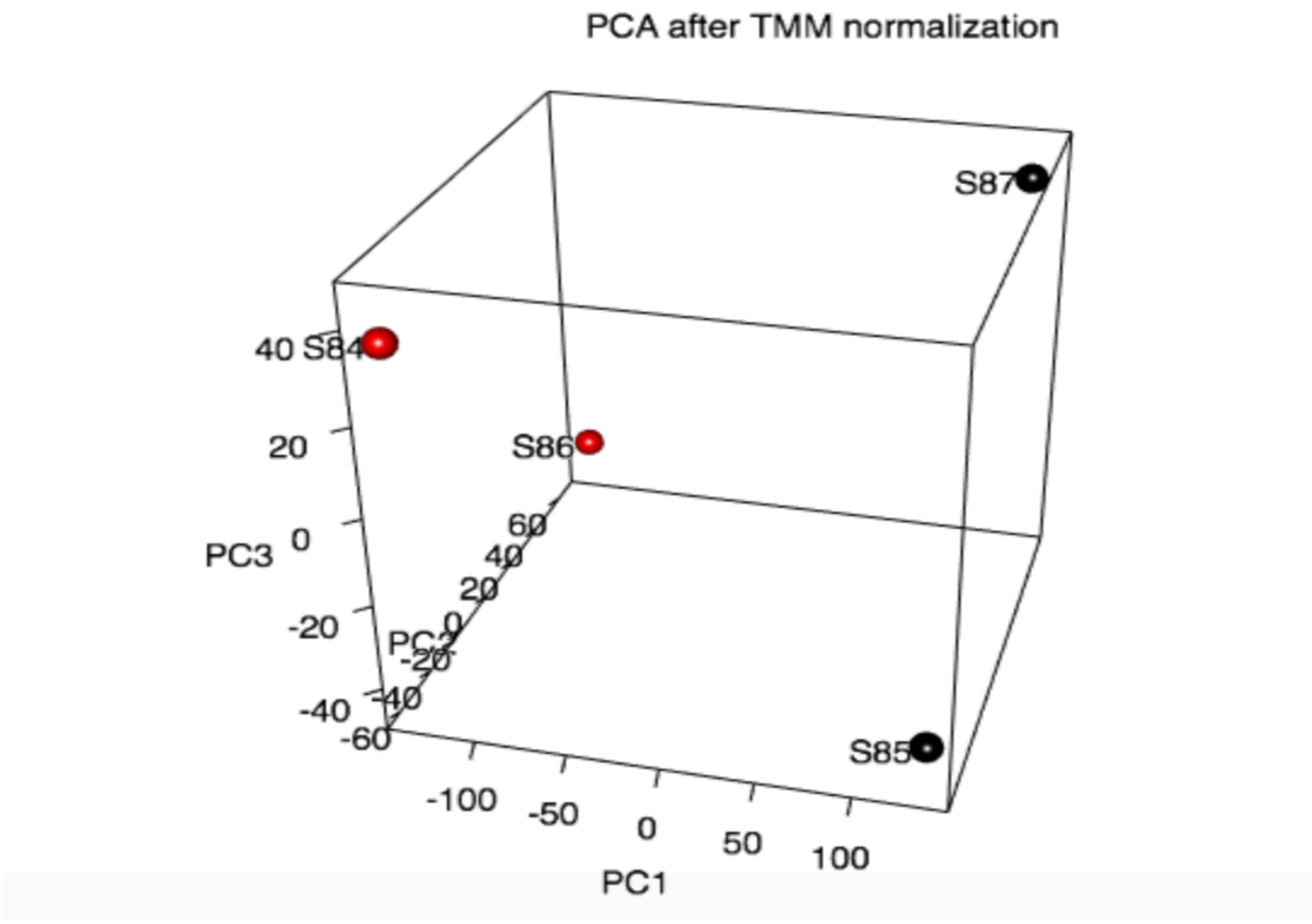


Checking unassigned rate for overlapping regions and multi-mapping reads

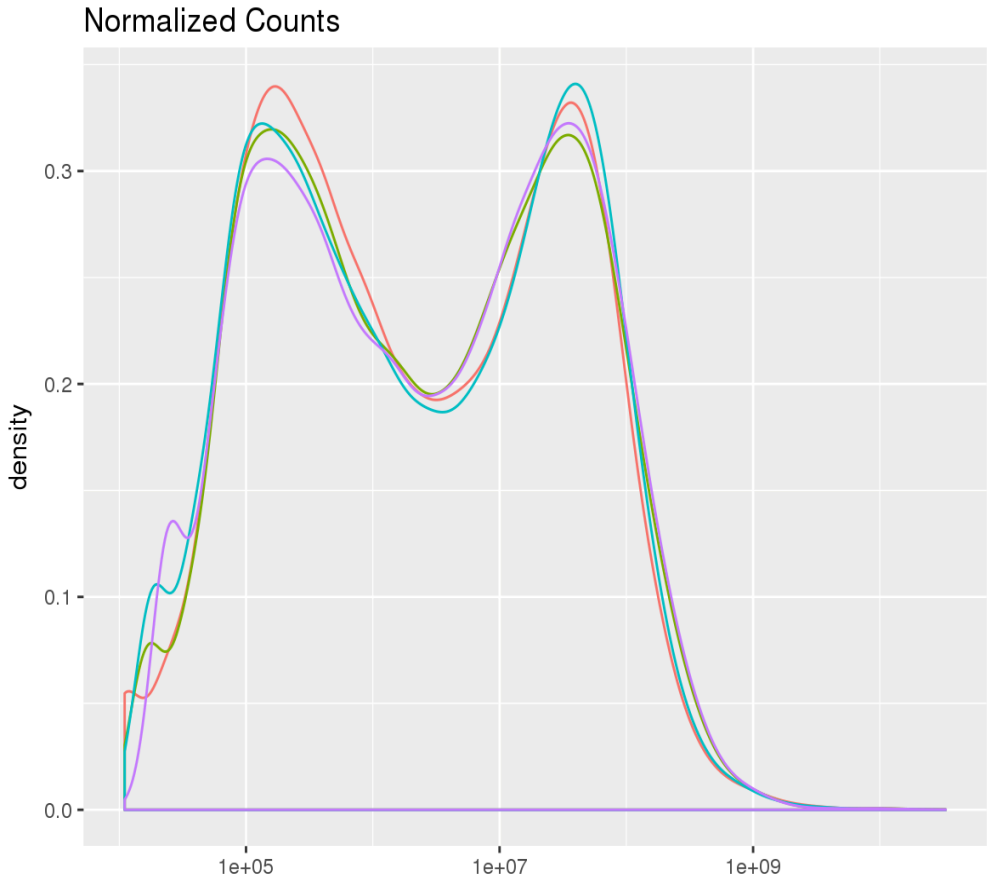
RNA-Seq: Differential expression workflow

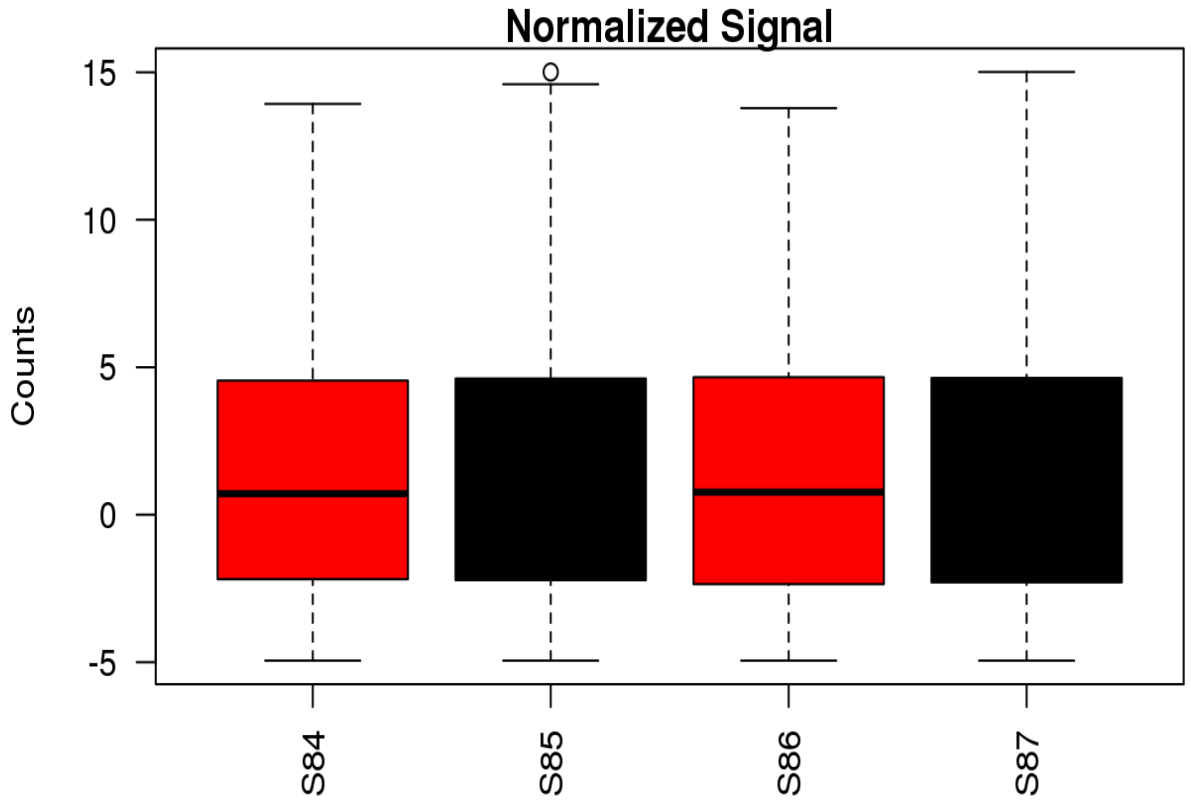


RNA-Seq: PCA report

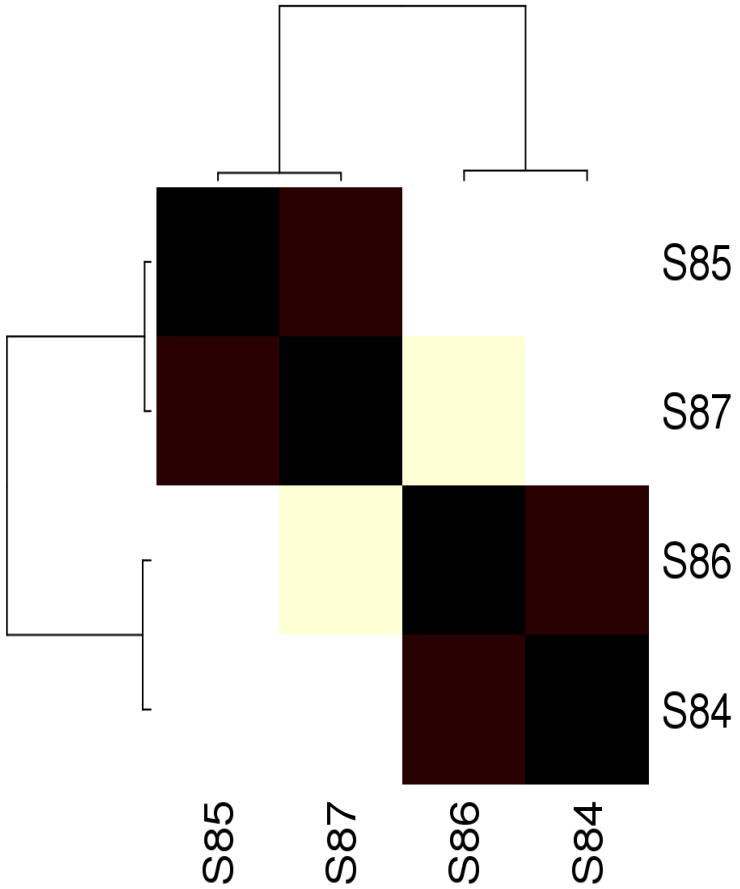


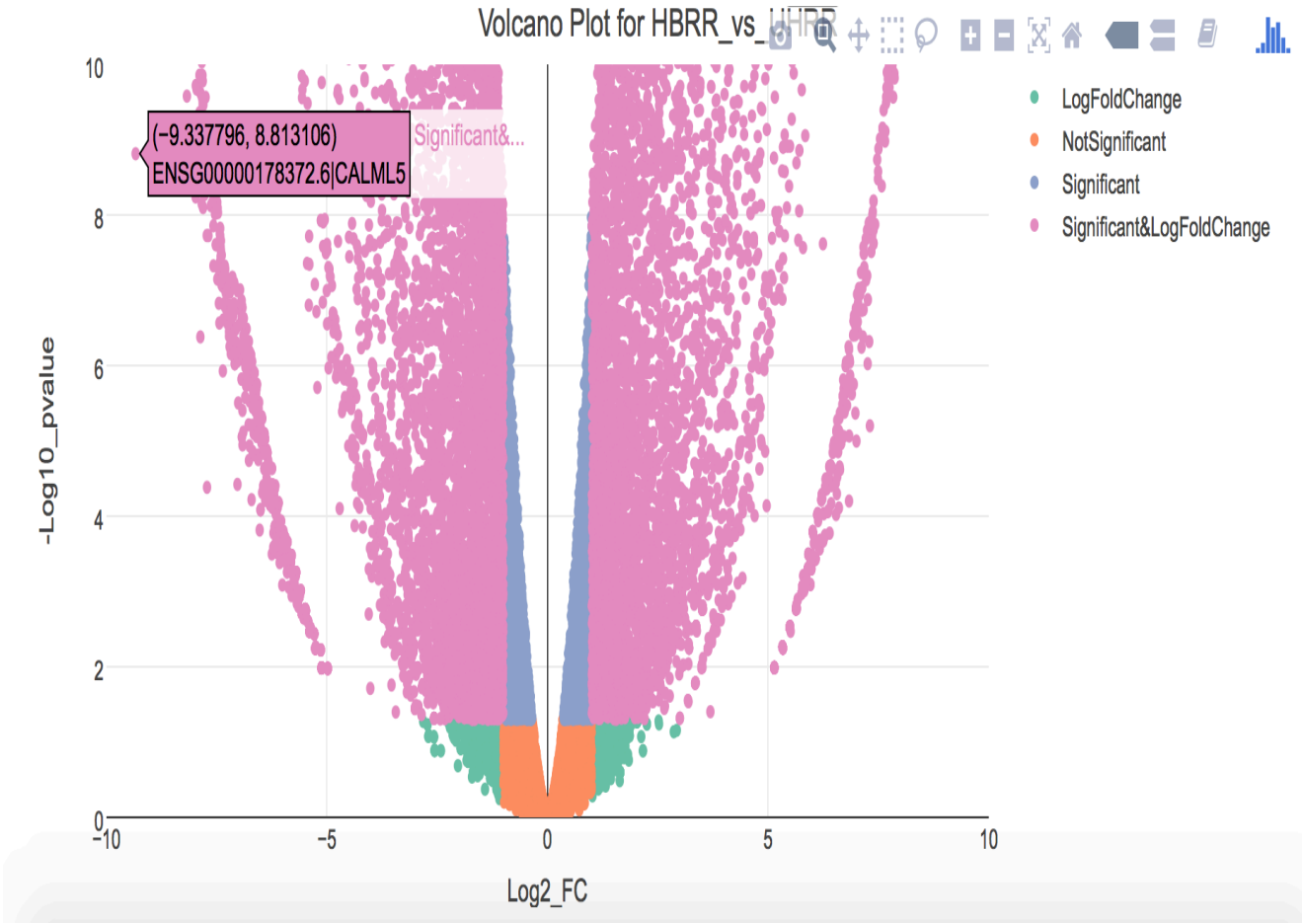
RNA-Seq: EdgeR DEG report (Limma, and Deseq2 also available)





RNA-Seq: EdgeR DEG report





DEG

Show 10 entries

Search:

	HBRR-UHRR_logFC	HBRR-UHRR_pval
ENSG00000000003.10|TSPAN6	-2.50928615777818	1.3380101042162e-28
ENSG00000000005.5|TNMD	-3.6495628070402	6.175042214714e-11
ENSG00000000419.8|DPM1	-1.46952506670122	1.49147150717859e-12
ENSG00000000457.9|SCYL3	-0.0751178228148991	0.776691564617663
ENSG00000000460.12|C1orf112	-2.7512929546514	2.07376241097899e-23
ENSG00000000938.8|FGR	5.77257447142532	8.64069023783316e-42
ENSG00000000971.11|CFH	0.367498252758908	0.120055574266859
ENSG00000001036.9|FUCA2	-2.50835963213615	7.88387959188995e-34
ENSG00000001084.6|GCLC	0.227863384547624	0.231618171572459
ENSG00000001167.10|NFYA	-0.765530062980488	0.000127763247166451

Showing 1 to 10 of 30,122 entries

EdgeR_deg_HBRR_vs_UHRR.txt

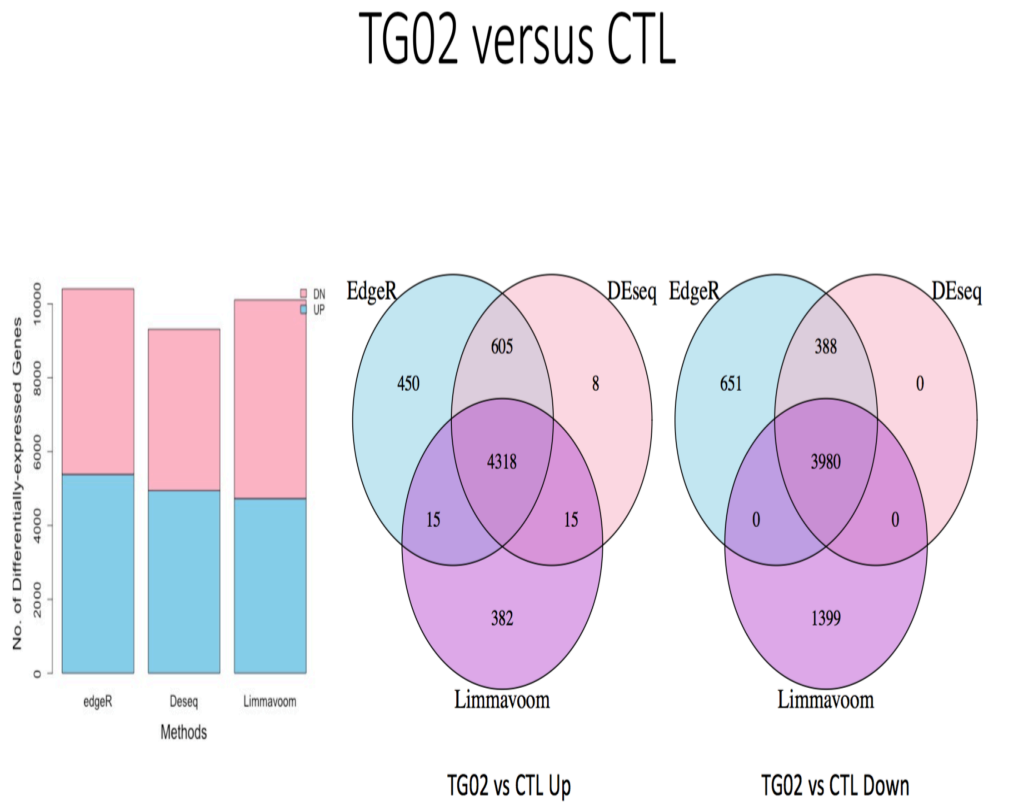
Id	ensID	gene	logFC	logCPM	PValue	FDR	FC
ENSG00000000003.10|TSPAN6	ENSG00000000003.10	TSPAN6	-2.509286158	4.944362674	1.34E-28	1.34E-27	-5.693383012
ENSG00000000005.5|TNMD	ENSG00000000005.5	TNMD	-3.649562807	-0.672287271	6.18E-11	2.61E-10	-12.54954199
ENSG00000000419.8|DPM1	ENSG00000000419.8	DPM1	-1.469525067	5.344403608	1.49E-12	7.04E-12	-2.769307134
ENSG00000000457.9|SCYL3	ENSG00000000457.9	SCYL3	-0.075117823	3.592577781	0.776691565	0.827749197	-1.053447066
ENSG00000000460.12|C1orf112	ENSG00000000460.12	C1orf112	-2.751292955	3.857511111	2.07E-23	1.70E-22	-6.733202968
ENSG00000000938.8|FGR	ENSG00000000938.8	FGR	5.772574471	1.710361124	8.64E-42	1.32E-40	54.66609706
ENSG00000000971.11|CFH	ENSG00000000971.11	CFH	0.367498253	4.312593606	0.120055574	0.164057252	1.290113731
ENSG00000001036.9|FUCA2	ENSG00000001036.9	FUCA2	-2.508359632	5.662233909	7.88E-34	9.44E-33	-5.68972779
ENSG00000001084.6|GCLC	ENSG00000001084.6	GCLC	0.227863385	5.654432914	0.231618172	0.293797219	1.171099279
ENSG00000001167.10|NFYA	ENSG00000001167.10	NFYA	-0.765530063	5.437495321	0.000127763	0.000297089	-1.699994481
ENSG00000001460.13|STPG1	ENSG00000001460.13	STPG1	1.457008799	4.439054736	1.16E-09	4.46E-09	2.745385607
ENSG00000001461.12|NIPAL3	ENSG00000001461.12	NIPAL3	2.759319171	7.353825876	1.14E-48	2.10E-47	6.770766522
ENSG00000001497.12|LAS1L	ENSG00000001497.12	LAS1L	-0.222818625	6.354111211	0.24308328	0.306533033	-1.167011376
ENSG00000001561.6|ENPP4	ENSG00000001561.6	ENPP4	3.557250132	5.450703915	2.46E-57	5.68E-56	11.77169475
ENSG00000001617.7|SEMA3F	ENSG00000001617.7	SEMA3F	-0.94998227	4.107566858	0.000351012	0.000761977	-1.931848916
ENSG00000001626.10|CFTR	ENSG00000001626.10	CFTR	1.363661906	0.807413586	0.000139806	0.000323271	2.573375357
ENSG00000001629.5|ANKIB1	ENSG00000001629.5	ANKIB1	0.058685985	6.421272185	0.756280072	0.80934623	1.04151671
ENSG00000001630.11|CYP51A1	ENSG00000001630.11	CYP51A1	-0.290194885	2.269595297	0.670414156	0.732924384	-1.222805448
ENSG00000001631.10|KRIT1	ENSG00000001631.10	KRIT1	-0.210802525	5.404130871	0.292101282	0.359584568	-1.157331792
ENSG00000002016.12|RAD52	ENSG00000002016.12	RAD52	-0.075564881	4.123075561	0.758993715	0.811817651	-1.053773555
ENSG00000002079.8|MYH16	ENSG00000002079.8	MYH16	-1.63888835	-0.611449531	0.001090409	0.002194368	-3.114257744
ENSG00000002330.9|BAD	ENSG00000002330.9	BAD	0.729585279	3.317604969	0.394161428	0.4666849	1.658162363
ENSG00000002549.8|LAP3	ENSG00000002549.8	LAP3	-0.714689805	6.043843772	0.000333134	0.000726081	-1.641130319
ENSG00000002587.5|HS3ST1	ENSG00000002587.5	HS3ST1	5.247363704	3.974820467	3.29E-54	7.00E-53	37.98515238
ENSG00000002726.15|AOC1	ENSG00000002726.15	AOC1	-4.329901882	-1.834716476	9.70E-08	3.16E-07	-20.11084621
ENSG00000002745.8|WNT16	ENSG00000002745.8	WNT16	2.365264313	0.04674071	1.08E-09	4.18E-09	5.152470403

What is the method to use?

No clear answer!

Compare results:

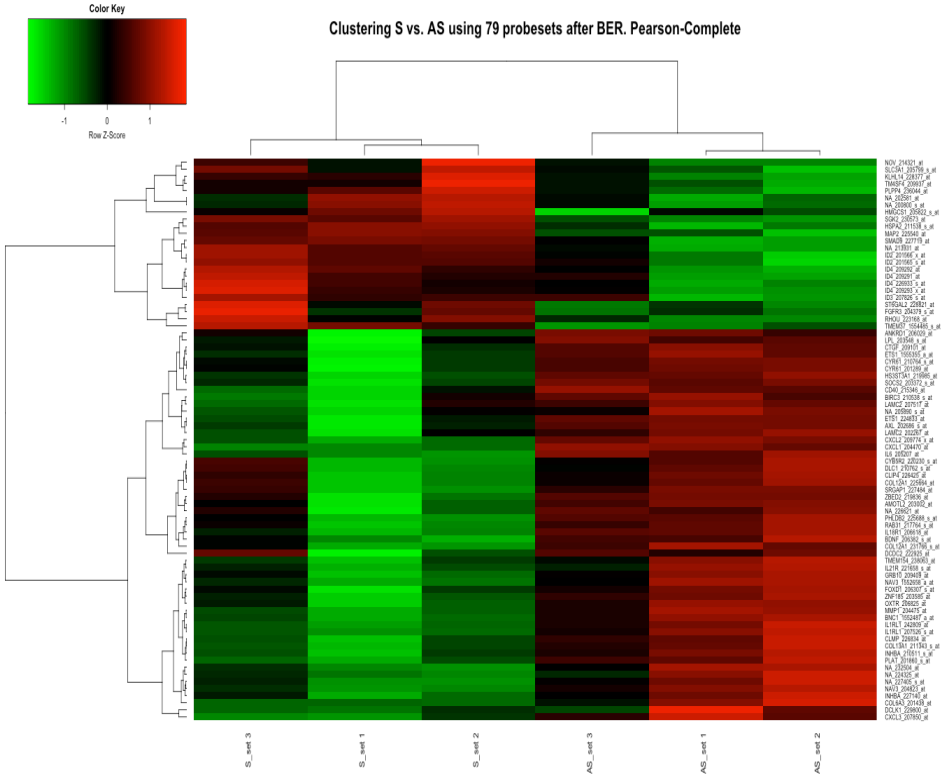
- PCA
- Sample clustering
- DEG results

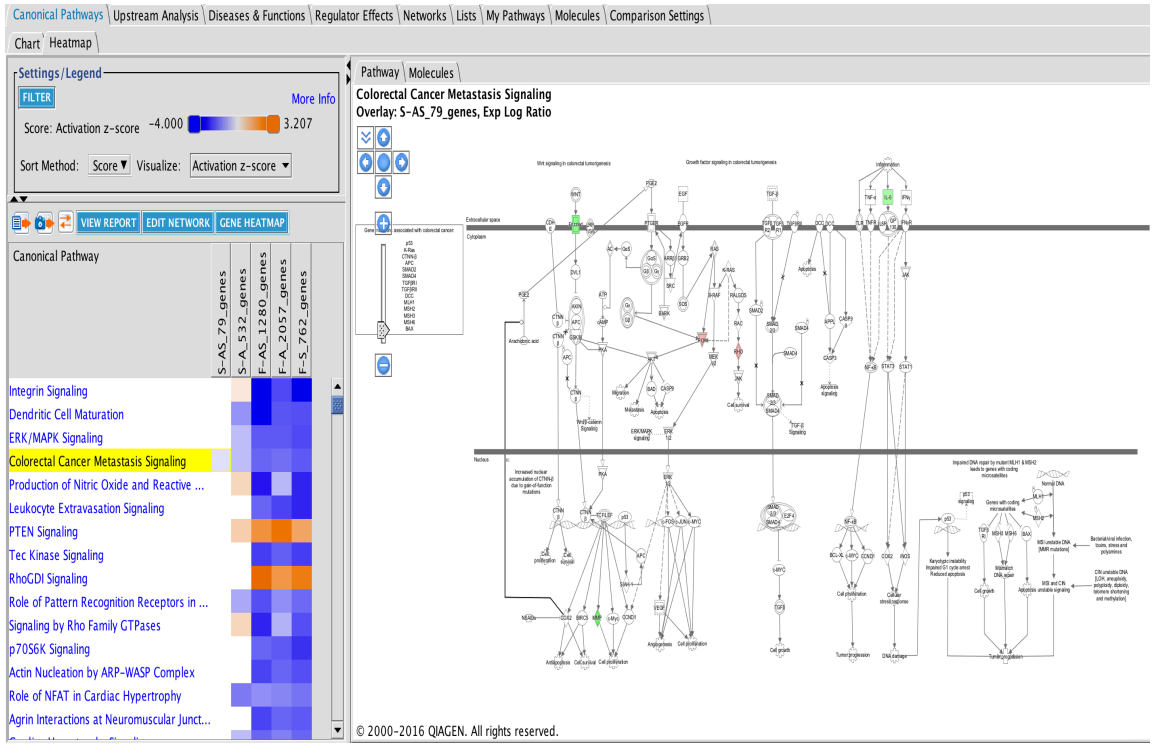


DEG Venn diagram

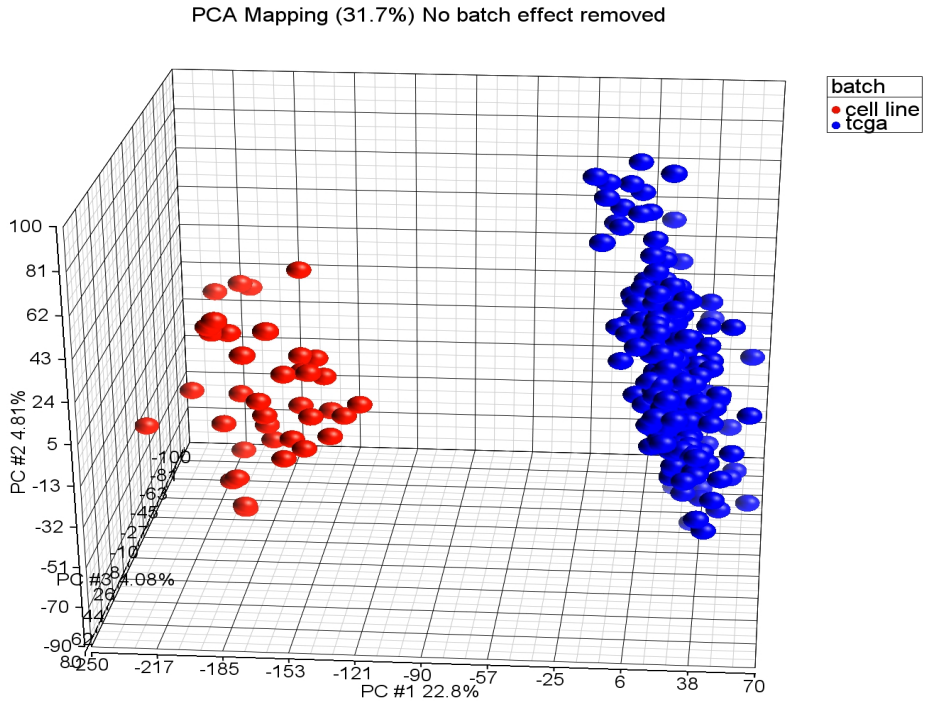
Visualization and enrichment analysis

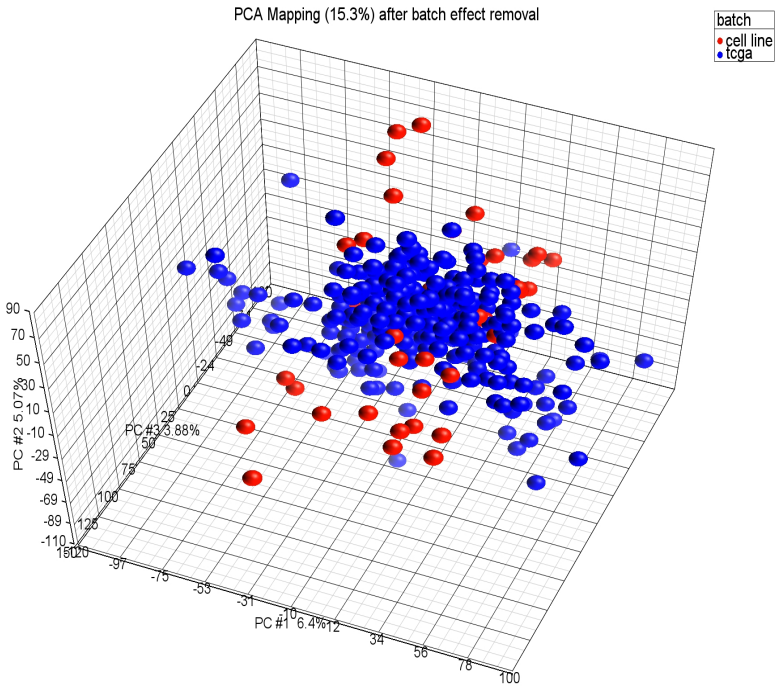
- Cluster the samples based on the top ranked genes (sd, mad, IQR..)
- Pathway enrichment (GSEA, IPA, …)
  - Easy use of DEG files





Dealing with Batch effect

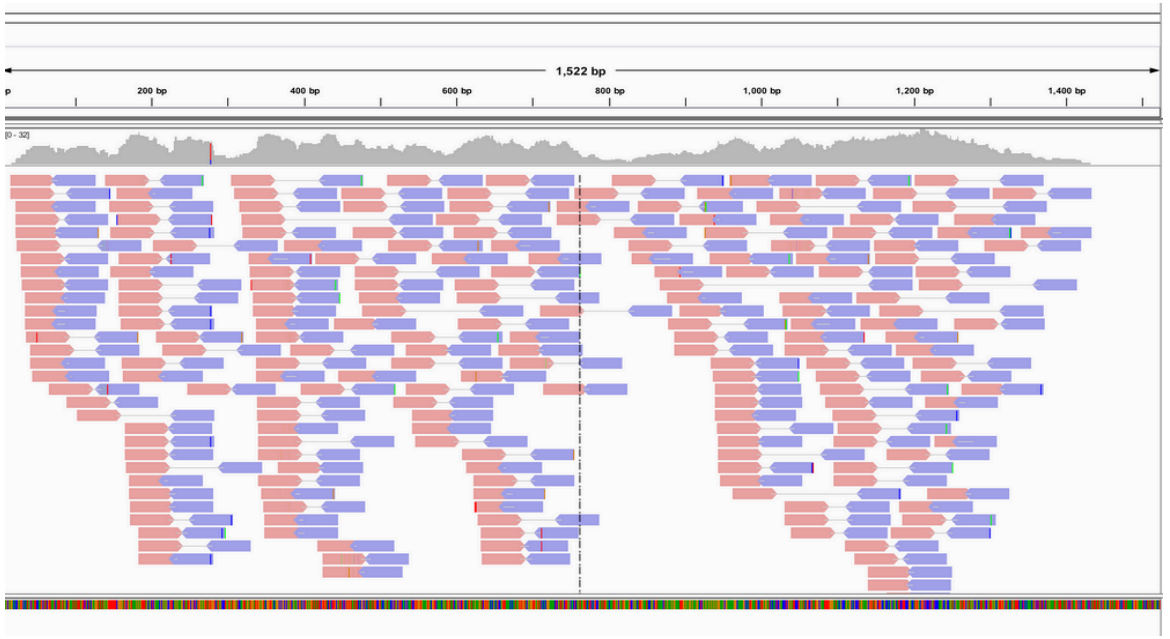




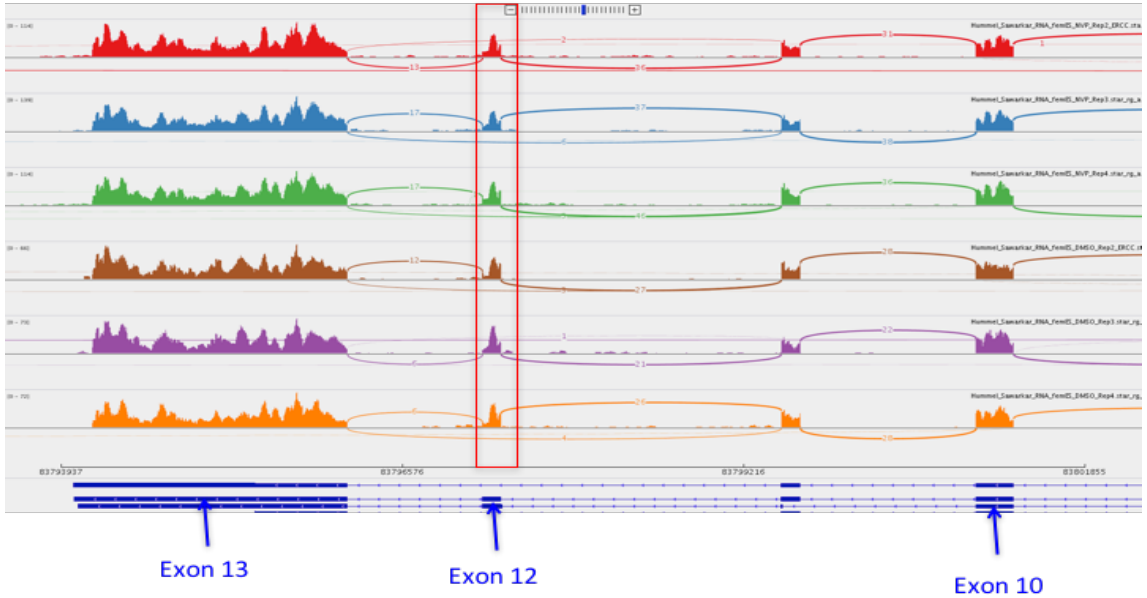
- incorporate batch effect as co-variate in the model)

Viewing RNA-Seq data

- Integrative Genomics Viewer (IGV)
  - Read alignments
  - Splices junctions



Sashimi plot



Agenda

- Introduction
- Data analysis Workflow
  - Review main steps
- CCBR RNA-Seq pipeline
  - Workflow overview
  - Quality Control reports
  - Principal Component Analysis PCA and differential expressed reports reports
  - Downstream analysis after running the pipeline
- Running the CCBR pipeline
  - Use case and demo

CCBR Pipeliner

- Offers for now 3 NGS data workflow: RnaSeq, ExomeSEq and GenomeSeq.
- Each workflow:
  - ✓ is version-aware
  - ✓ is modular and extensible
    - Multiple options/programs can be selected for a task.
  - ✓ is reproducible
    - uses a config file
  - ✓ maintains an audit trail (as a log file)
  - ✓ runs on NIH cluster and use Queue system
  - ✓ informs user, via email, once run is complete

Data preparation/ Input

- Pipeliner takes in raw paired-end NGS data: fastq.gz files
- Fastq naming convention:
  - <samplename>.R1.fastq.gz,
  - <samplename>.R2.fastq.gz

- Pipeliner can convert filenames to the desired naming convention
  - labels.txt: two-column text file
    - SampleA_R1_001.fastq.gz TumR1_Batch1.R1.fastq.gz
- For DEG, you need to know the phenotype/group for the samples and the contrasts for differential analysis

“groups.tab” file

Mandatory Fields (without labels)

Sample Name	group	Sample label
sample1	treat	treat1
sample2	treat	treat2
sample3	treat	treat3
sample4	control	ctrl1
sample5	control	ctrl2
sample6	control	ctrl3
…	…	…

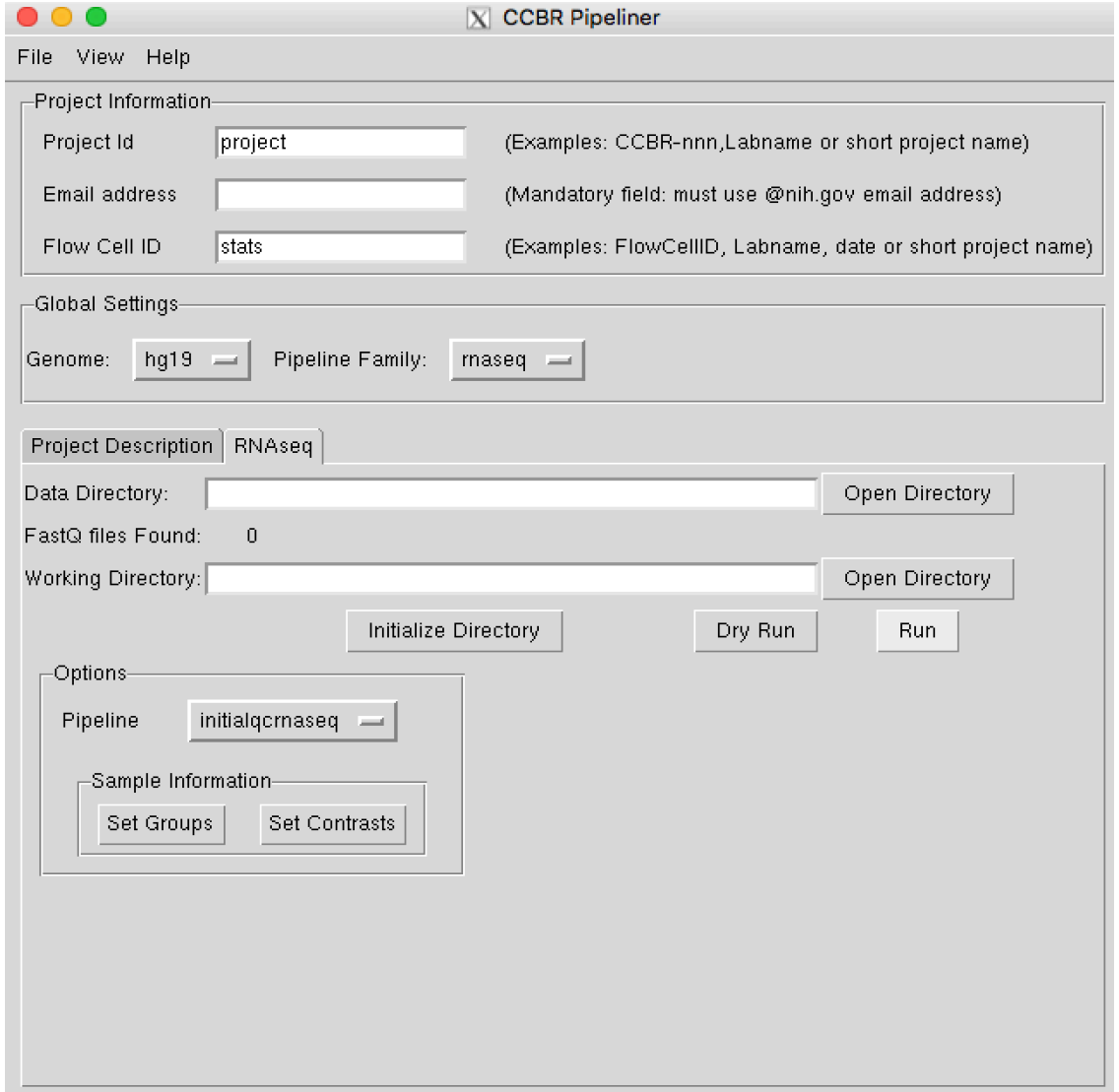
Only one factor

(you can simulate multifactor variable)

“contrasts.tab” file

Group1	Vs. group2
treat	control
…	…

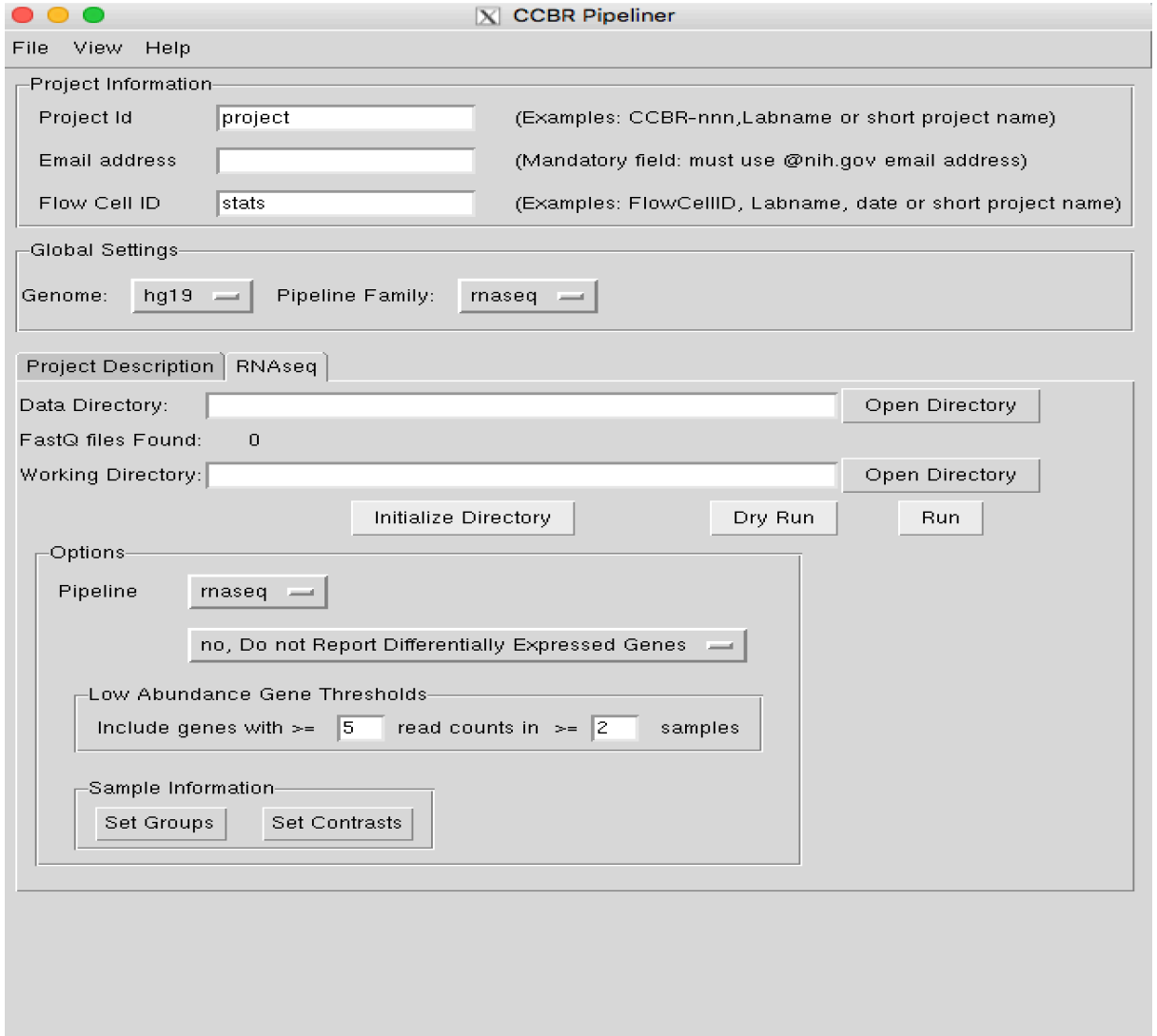
CCBR RNASEQ Pipeline (InitialQC)



Data directory:
/scratch/elloumif/SEQC4/

Working directory:
/data/<user>/…

CCBR RNASEQ Pipeline (DEG Analysis)



Data directory:
/scratch/elloumif/SEQC4/

Working directory:
/data/<user>/…

RNA-Seq Output: Main directories

- **rawQC**: Fastqc results on raw data
- **Trim**: trimmed data (adaptor cut)
- **QC**: Fastqc results on trimmed data
- **FQscreen**: FastqScreen results (trimmed data)
- **Reports**: contains Multiqc report and main log file of the pipeline (snakemake.log)
- **DEG_genes**: DEG results based on gene count + Html reports
- **DEG_genejunctions**: DEG results based on junction gene count + Html reports

DEG directory output files

- Limma* files (txt, png, html)
- Deseq2* files
- edgeR* files

RNA-Seq Output: Main files (main working directory)

- Bam files (*.bam)
- raw count data (3 methods):
  - Gene: RawCountFile_gene.txt and RawCountFile_genes_filtered.txt
- Gene Normalized data: CPM_TMM_counts.txt
- RSEM results:
  - <sample>.rsem.genes.results
  - <sample>.rsem.isoforms.results
- EBSEQ results:
  - **<sample>isoform..EBSeq**
  - <sample>.isoform.EBSeq.normalized_data_matrix
  - <sample>.isoform.EBSeq.counts.matrix
- Run.json: configuration file – run settings

Configuration file


```
"project": {
    "DEG": "yes",
    "MINCOUNTGENEJUNCTIONS": "5",
    "MINCOUNTGENES": "5",
    "MINCOUNTJUNCTIONS": "5",
    "MINSAMPLES": "2",
    "PICARDSTRAND": "NONE",
    "SJDBOVERHANG": "100",
    "STARDIR": "/fdb/STAR_current/GENCODE/Gencode_human/release_19/genes-100",
    "STARSTRANDCOL": "2",
    "STRANDED": "0",
    "TRIM": "yes",
    "analyst": "",
    "annotation": "hg19",
    "batchsize": "20",
    "binset": "standard-bin",
    "cluster": "cluster_medium.json",
    "contrasts": {
        "rcontrasts": [
            "HBRR",
            "UHRR"
        ]
    },
    "custom": [],
    "datapath": "/data/CCBR/dev/RNA-Seq-techdev/SEQC_dataset/FASTQfiles",
    "description": "Enter CCBR Project Description and Notes here.\n",
    "efiletype": "fastq",
    "filetype": "fastq.gz",
    "flowcellid": "stats",
    "groups": {
        "rgroups": [
            "UHRR",
            "HBRR",
            "UHRR",
            "HBRR"
        ],
        "rlabels": [
            "S84",
            "S85",
            "S86",
            "S87"
        ],
        "rsamps": [
            "SRR950084",
            "SRR950085",
            "SRR950086",
            "SRR950087"
        ]
    },
```



```
"rnaseq": {
    "CONFMULTIQC": "/data/CCBR_Pipeliner/db/PipeDB/Rnaseq/multiqc_config.yaml",
    "RSEM": "/usr/local/apps/rsem/1.3.0",
    "STARVER": "STAR/2.5.2b",
    "MULTIQC": "multiqc/0.9dev0",
    "PICARDVER": "picard/1.119",
    ...
```

Setup before running ccbrpipeliner

- Helix and Biowulf accounts
- X11 client (Windows: Putty, NoMachine; Mac: Xquartz, NoMachine)
- Space:
  - Biowulf home directories have default of 100GB allocation: not enough to run NGS pipelines.
  - Best option: have a lab-wide /data/labname storage allocation, with higher storage
- Basic knowledge of Unix commands (ssh, mkdir, vi)

CCBR pipeliner availability

- ✓ https://github.com/CCBR/Pipeliner
- ✓ via module “ccbrpipeliner” at Biowulf

CCBR pipeliner documentation

https://github.com/CCBR/Pipeliner/blob/master/PipelinerVer1.0_documentation.pdf

Demo

Use case: 4 samples from SEQC study

- Mixture of biological sources and a set of synthetic RNAs from the External Rna Control Consortium (ERCC)

➢2 samples from group A : Strategene Universal Human Reference RNA (UHRR) – from 10 human cell lines-

➢2 samples from group B: Ambion Human Brain Reference RNA (HBRR)

➢Illumina HiSeq2000. -100 bp-

Input files

- Fastq files
- Labels.txt
- Groups.tab
- Contrasts.tab

Output files

- FASTQC report
- MultiQC report
- Pca report
- Edge R report
- Rawcount files
- Normalized data files

Q&A