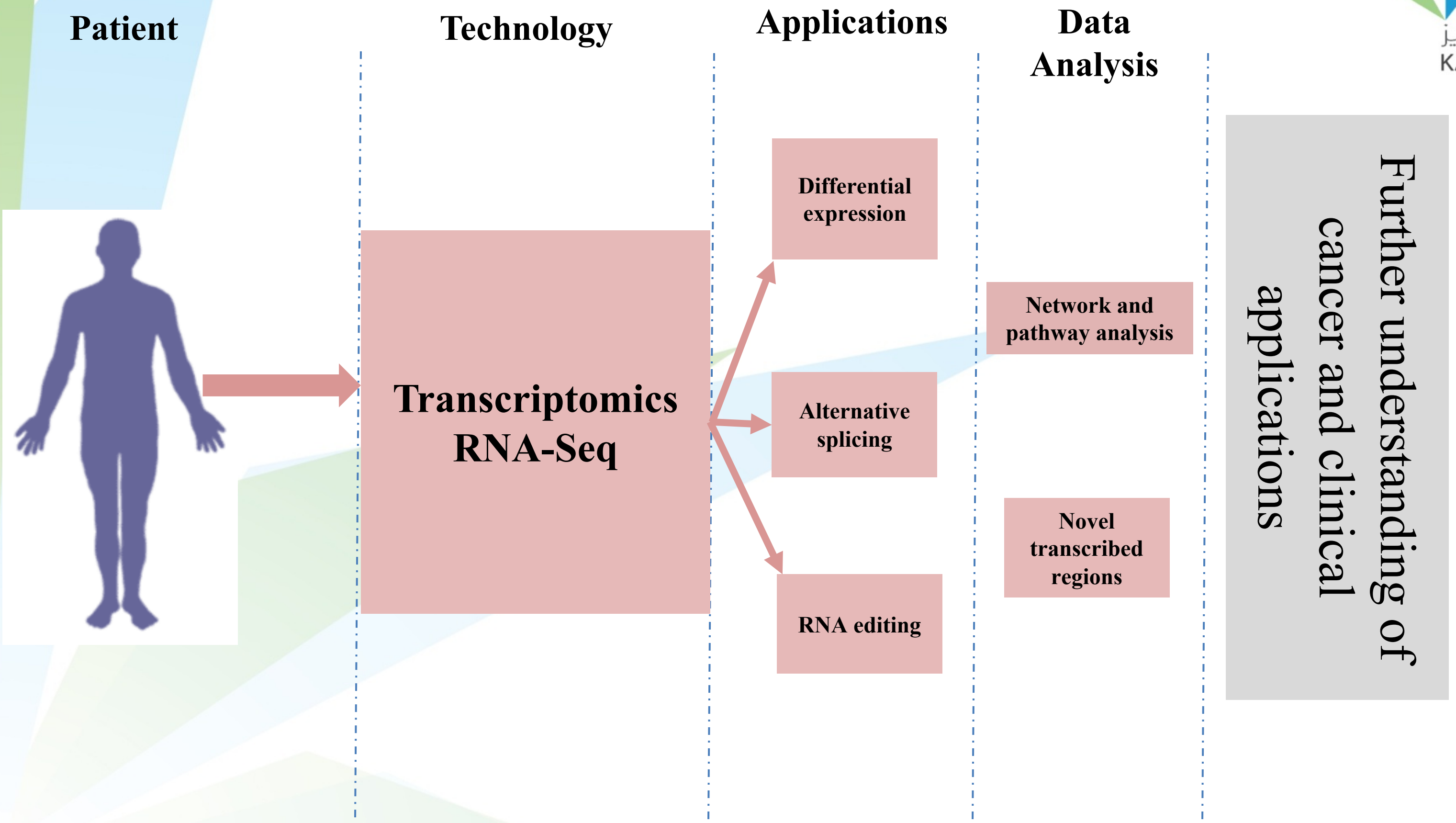
مدينة الملك عبدالعزيز
للعلوم والتقنية KACST

RNA-Seq data analysis

Bioinformatics Group

2 May 2017



Patient

Technology

Applications

Data Analysis

Transcriptomics RNA-Seq

Differential expression

Alternative splicing

RNA editing

Network and pathway analysis

Novel transcribed regions

Further understanding of cancer and clinical applications

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.

Benefits & Challenges

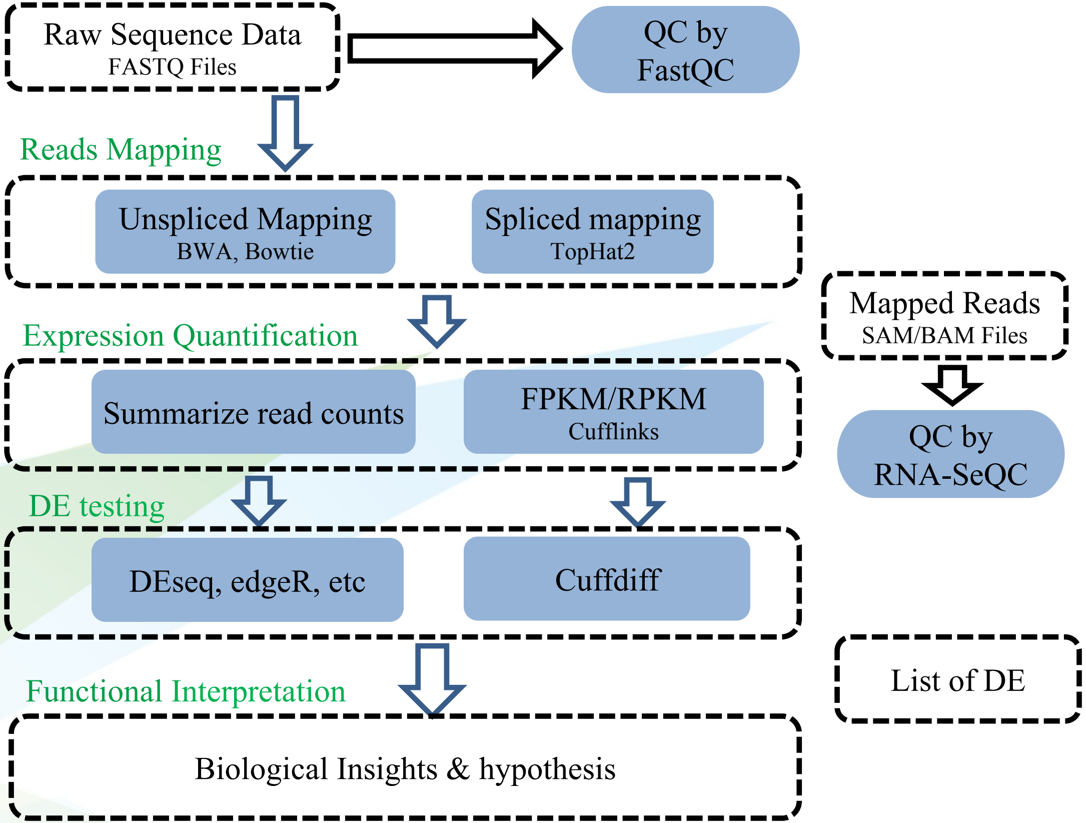
Benefits:

- Independence on prior knowledge
- High resolution, sensitivity and large dynamic range

Challenge:

- Interpretation is not straightforward
- Procedures continue to evolve

From reads to differential expression



Raw Sequence Data
FASTQ Files

QC by FastQC

Reads Mapping

Unspliced Mapping
BWA, Bowtie

Spliced mapping
TopHat2

Mapped Reads
SAM/BAM Files

QC by RNA-SeQC

Expression Quantification

Summarize read counts

FPKM/RPKM
Cufflinks

DE testing

DEseq, edgeR, etc

Cuffdiff

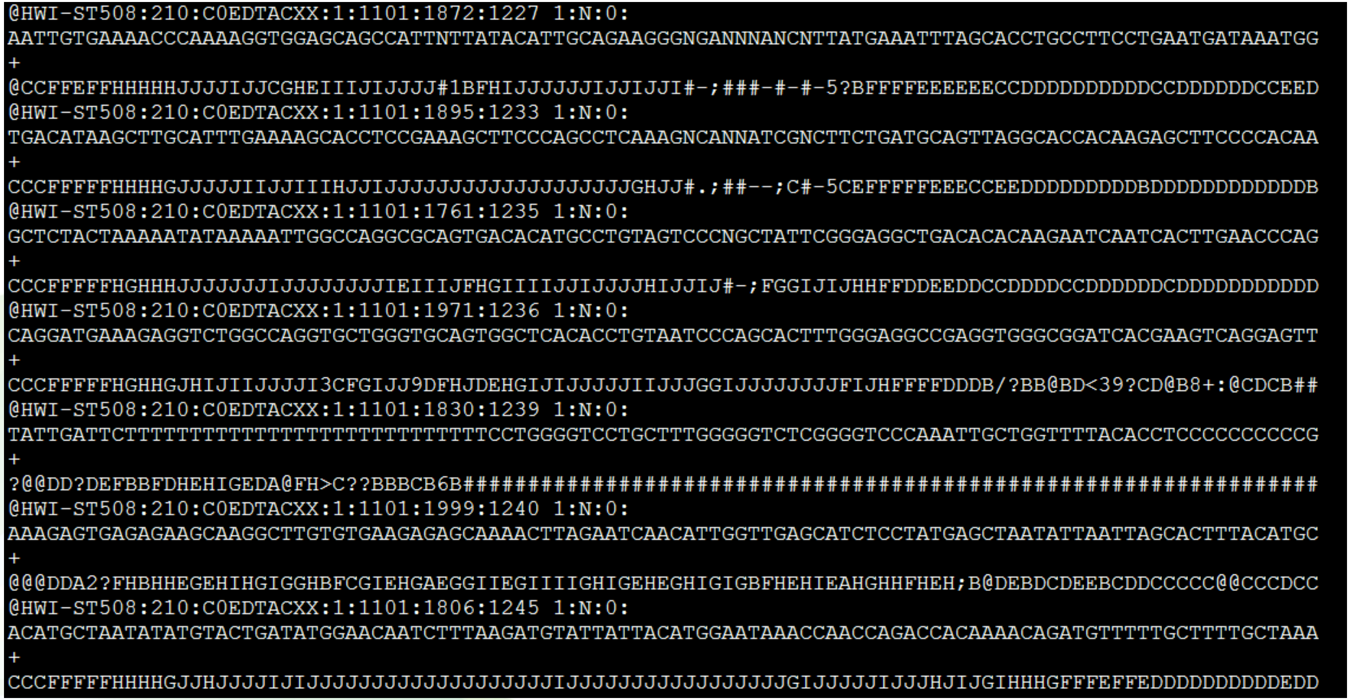
List of DE

Functional Interpretation

Biological Insights & hypothesis

FASTQ file

Line1: Sequence identifier
Line2: Raw sequence
Line3: meaningless
Line4: quality values for the sequence



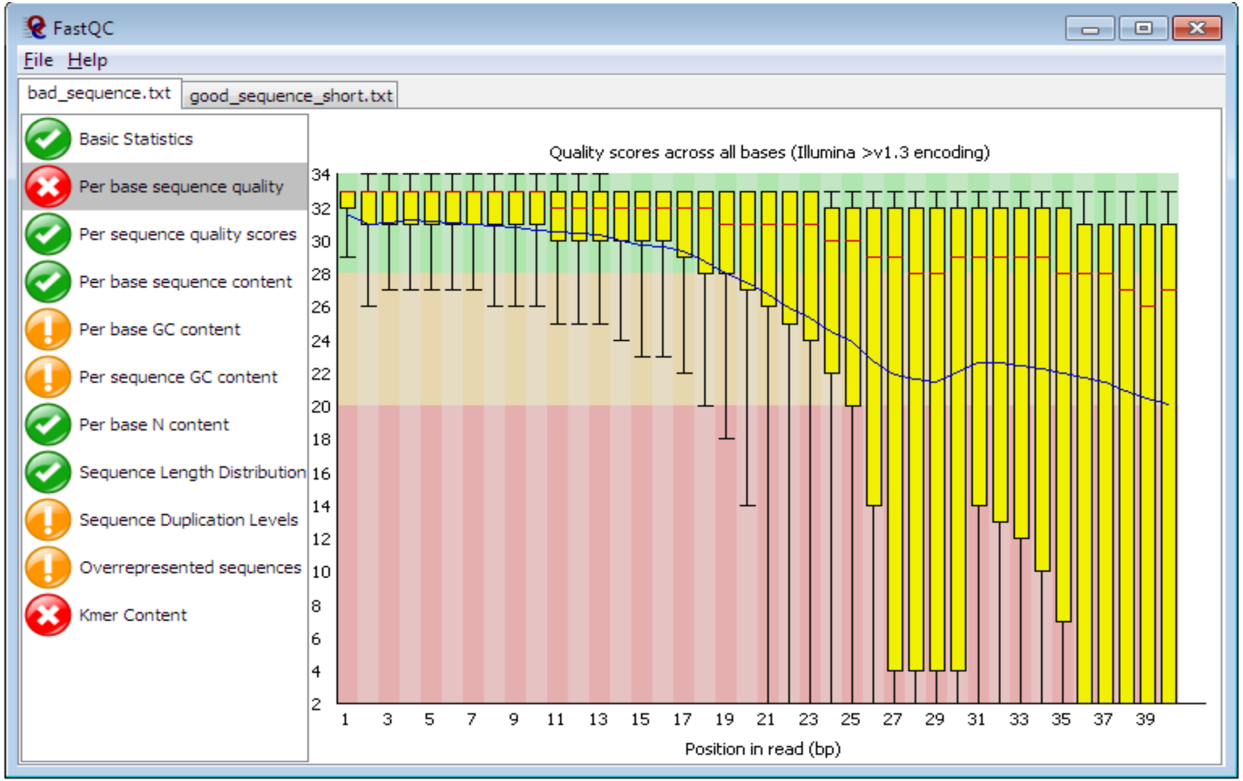
```
@HWI-ST508:210:C0EDTACXX:1:1101:1872:1227 1:N:0:
AATTGTGAAAACCCAAAAGGTGGAGCAGCCATTNTTATACATTGCAGAAGGGNGANNNANCNTTATGAAATTTAGCACCTGCCTTCCTGAATGATAAATGG
+
@CCFFEFFHHHHHJJJJIJJCGHEIIIJIJJJJ#1BFHIJJJJJJIJJIJJI#-;###-#-#-5?BFFFFEEEEEECCDDDDDDDDDDCCDDDDDDCCEED
@HWI-ST508:210:C0EDTACXX:1:1101:1895:1233 1:N:0:
TGACATAAGCTTGCATTTGAAAAGCACCTCCGAAAGCTTCCCAGCCTCAAAGNCANNATCGNCTTCTGATGCAGTTAGGCACCACAAGAGCTTCCCCACAA
+
CCCFFFFFHHHHGJJJJJIIJJIIIHJJIJJJJJJJJJJJJJJJJJJGHJJ#.;##--;C#-5CEFFFFFFEEECCEEDDDDDDDDDBDDDDDDDDDDDDB
@HWI-ST508:210:C0EDTACXX:1:1101:1761:1235 1:N:0:
GCTCTACTAAAAATATAAAAATTGGCCAGGCGCAGTGACACATGCCTGTAGTCCCNGCTATTCGGGAGGCTGACACACAAGAATCAATCACTTGAACCCAG
+
CCCFFFFFHGHHHJJJJJJIJJJJJJJIEIIIJFHGIIIIJJIJJJJHIJJIJ#-;FGGIJIJHHFFDDEEDDCCDDDDCCDDDDDDCDDDDDDDDDDD
@HWI-ST508:210:C0EDTACXX:1:1101:1971:1236 1:N:0:
CAGGATGAAAGAGGTCTGGCCAGGTGCTGGGTGCAGTGGCTCACACCTGTAATCCCAGCACTTTGGGAGGCCGAGGTGGGCGGATCACGAAGTCAGGAGTT
+
CCCFFFFFHGHHGJHIJIIJJJJI3CFGIJJ9DFHJDEHGIJIJJJJJIIJJJGGIJJJJJJJFIJHFFFFDDDB/?BB@BD<39?CD@B8+:@CDCB##
@HWI-ST508:210:C0EDTACXX:1:1101:1830:1239 1:N:0:
TATTGATTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTCCTGGGTCCTGCTTTGGGGGTCTCGGGGTCCCAAATTGCTGGTTTTACACCTCCCCCCCCCG
+
?@@DD?DEFBBFDHEHIGEDA@FH>C??BBBCB6B##################################################################
@HWI-ST508:210:C0EDTACXX:1:1101:1999:1240 1:N:0:
AAAGAGTGAGAGAAGCAAGGCTTGTGTGAAGAGAGCAAAACTTAGAATCAACATTGGTTGAGCATCTCCTATGAGCTAATATTAATTAGCACTTTACATGC
+
@@@DDA2?FHBHHEGEHIHGIGGHBFCGIEHGAEGGIIEGIIIIGHIGEHEGHIGIGBFHEHIEAHGHHFHEH;B@DEBDCDEEBCDDCCCCC@@CCCDCC
@HWI-ST508:210:C0EDTACXX:1:1101:1806:1245 1:N:0:
ACATGCTAATATATGTACTGATATGGAACAATCTTTAAGATGTATTATTACATGGAATAAACCAACCAGACCACAAAACAGATGTTTTTGCTTTTGCTAAA
+
CCCFFFFFHHHHGJJHJJJJIJIJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJJJJJJJGIJJJJJIJJJHJIJGIHHHGFFFEFFEDDDDDDDDDDEDD
```

Sequencing QC

Information we need to check

- Basic information(total reads, sequence length, etc.)
- Per base sequence quality
- Overrepresented sequences
- GC content
- Duplication level
- Etc.

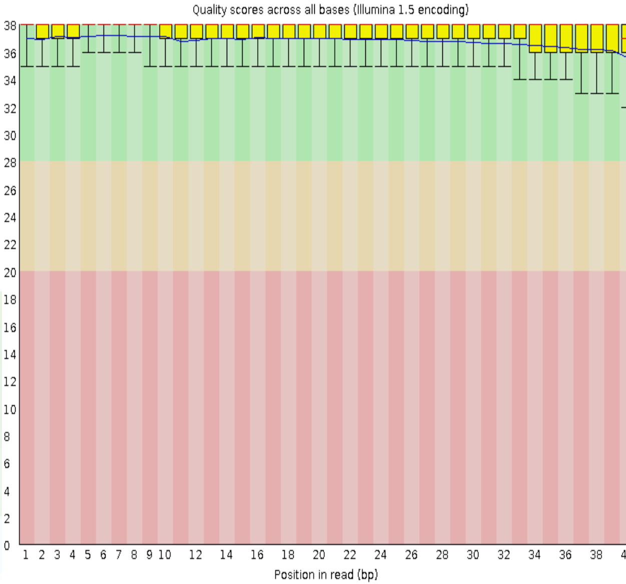
FastQC

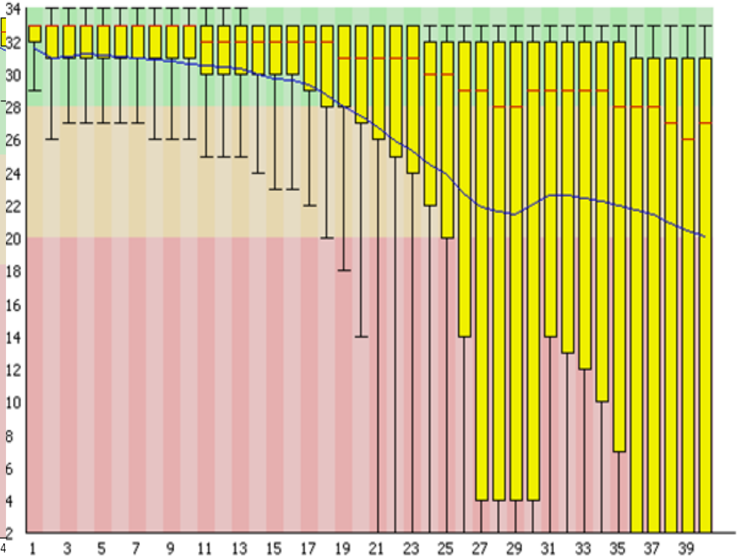


- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✗ Kmer Content

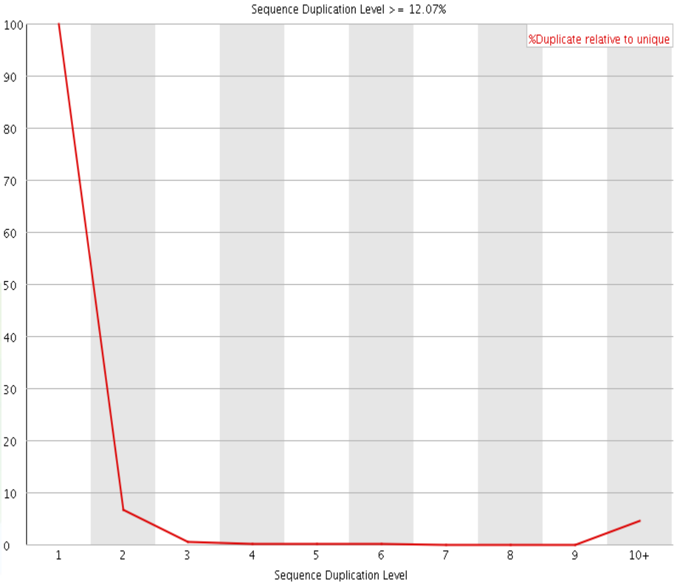
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

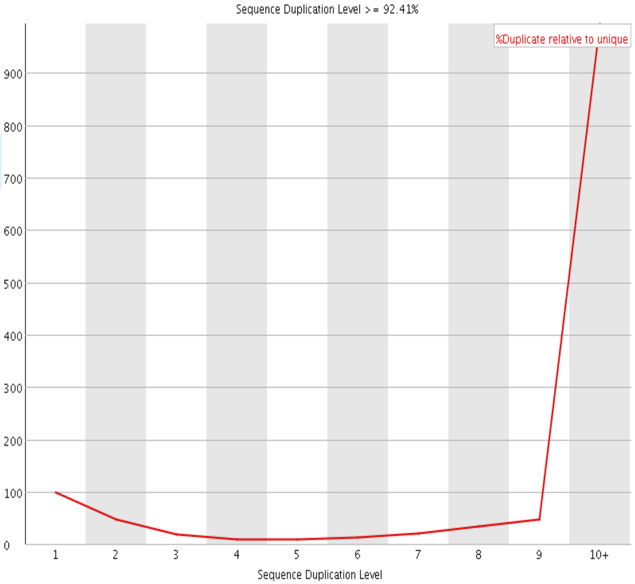
Per base sequence quality





Duplication level

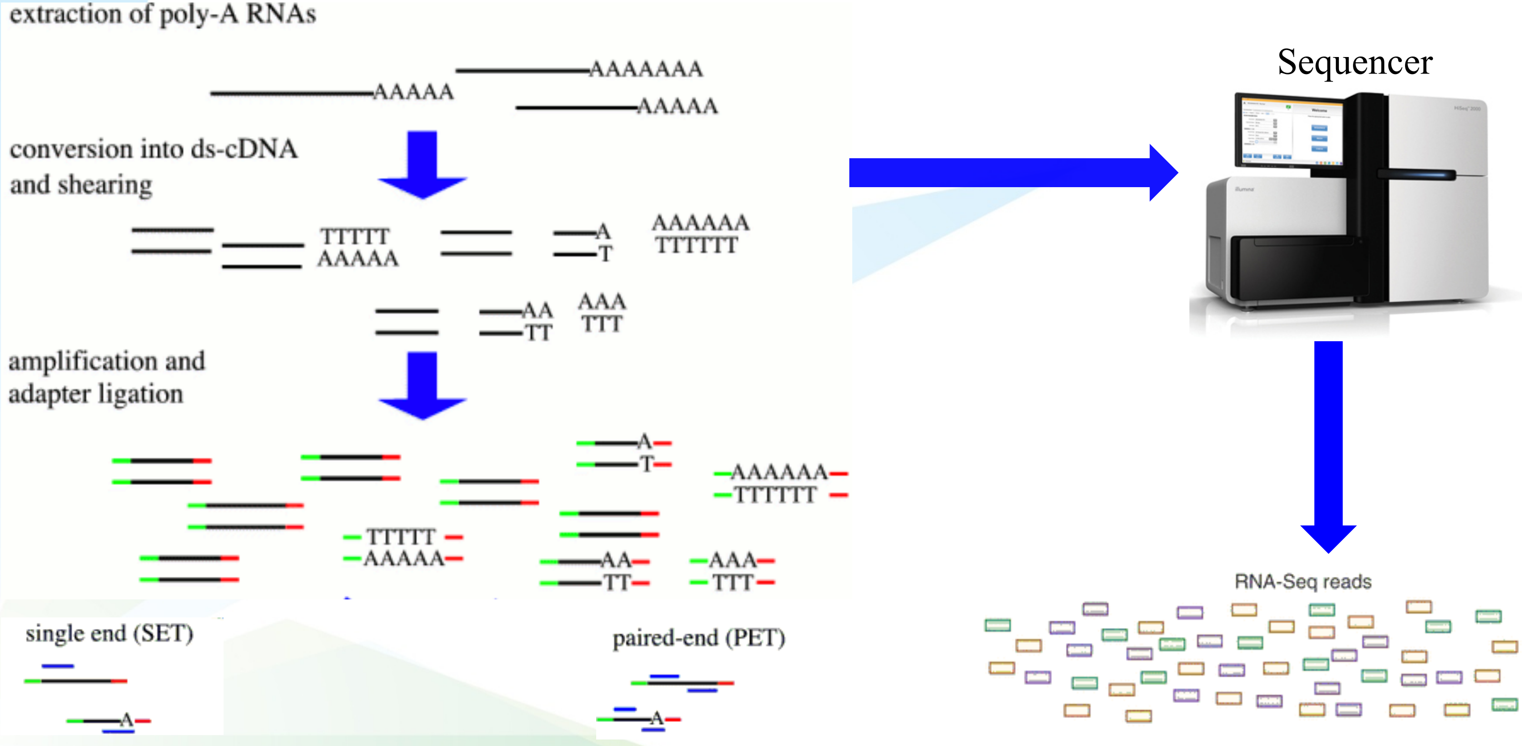




Overview of RNA-Seq

Transcriptome profiling using NGS

مدينة الملك عبدالعزيز
KACST للعلوم والتقنية



extraction of poly-A RNAs

conversion into ds-cDNA and shearing

amplification and adapter ligation

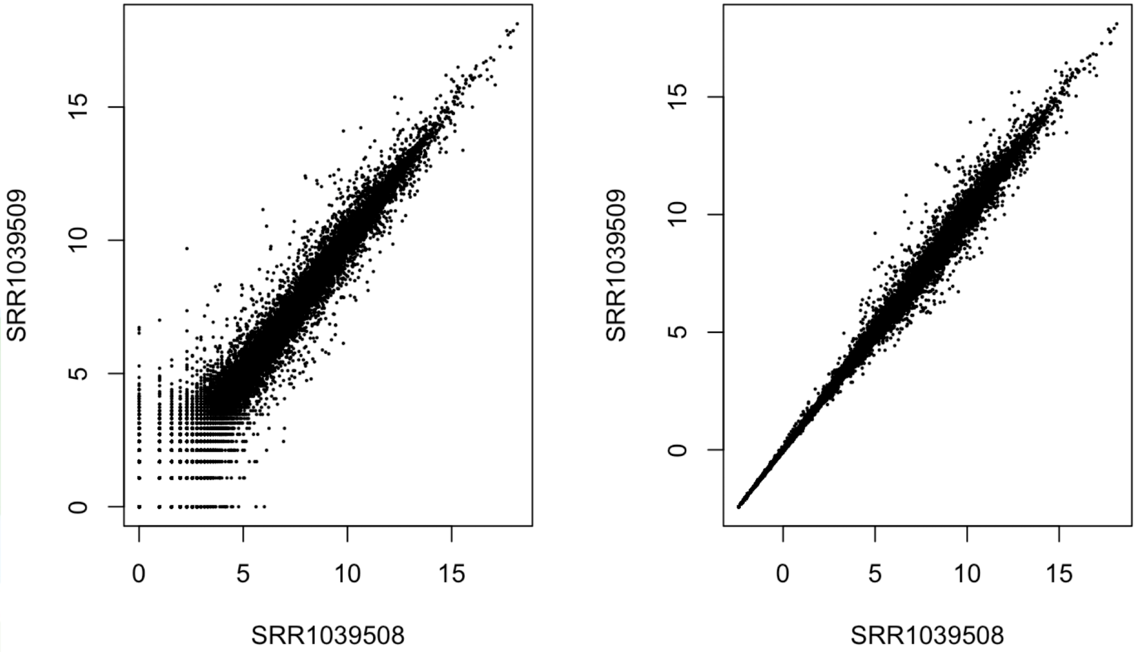
single end (SET)

paired-end (PET)

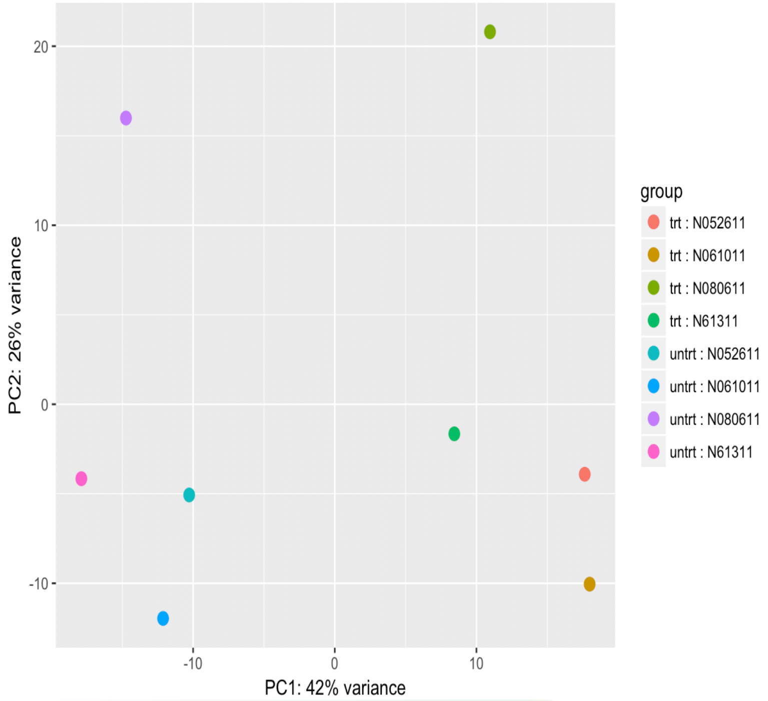
Sequencer

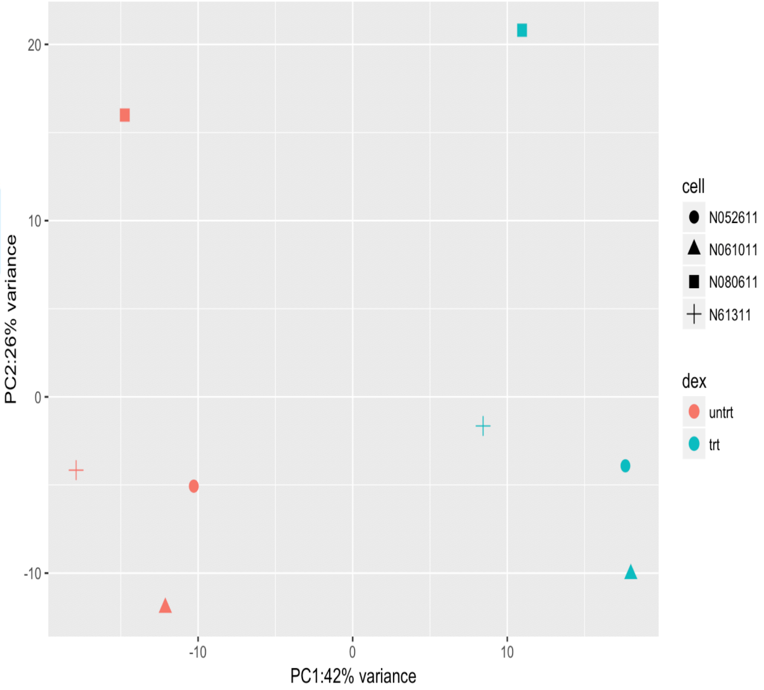
RNA-Seq reads

Before and After Normalization

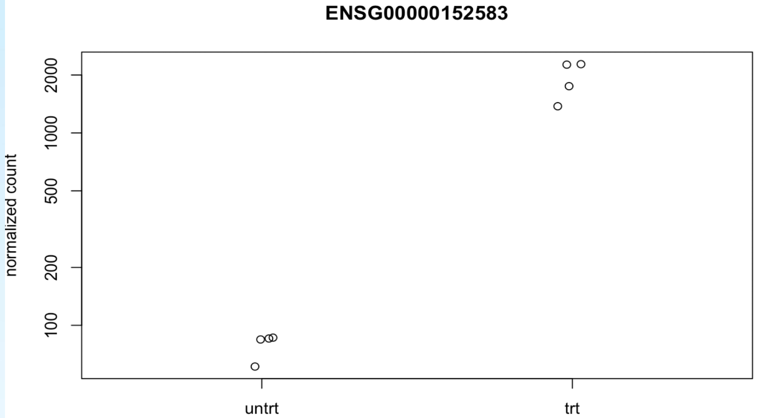


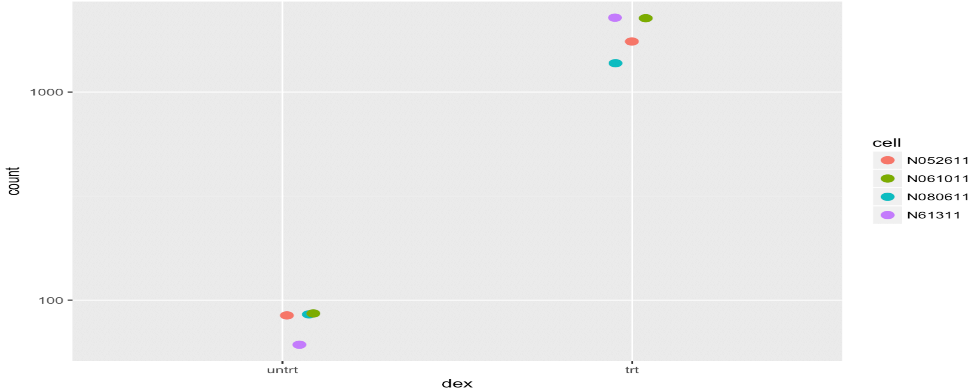
Principal component analysis (PCA) Plots

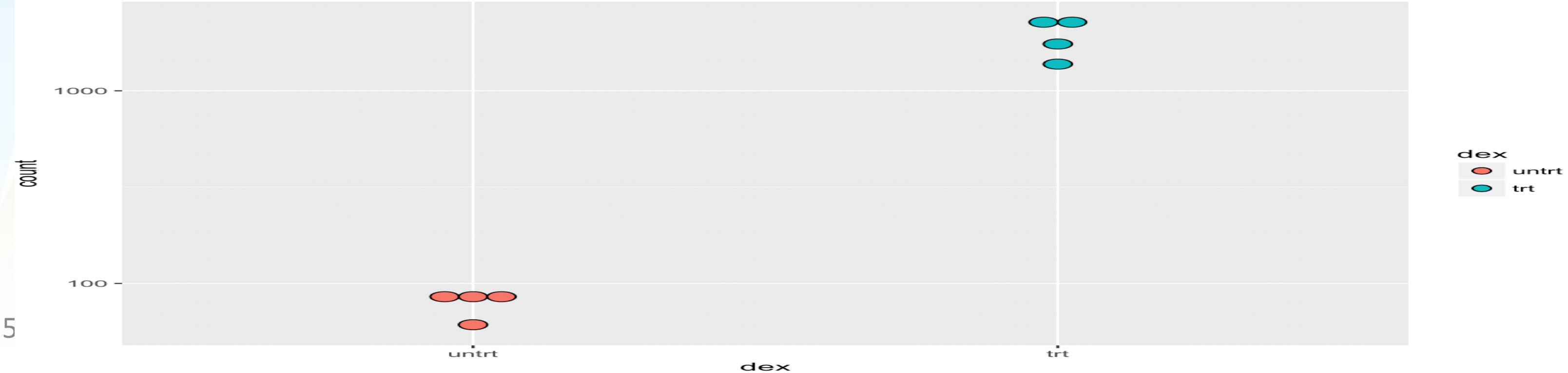




Principal component analysis (PCA) Plots

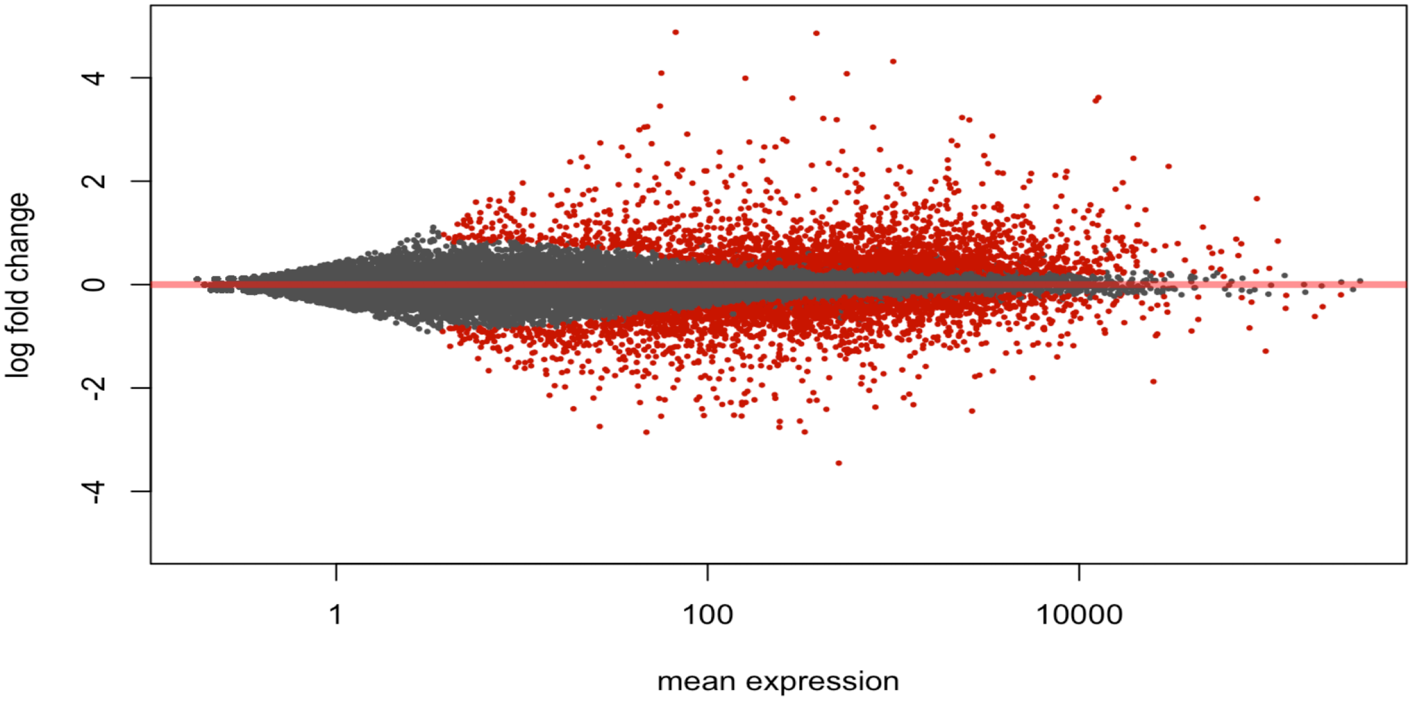






MP-Plot

مدينة الملك عبدالعزيز
KACST للعلوم والتقنية


log fold change
mean expression