Sergio Sarnataro, PhD

Bioinformatics Awareness Days
July 11th, 2022

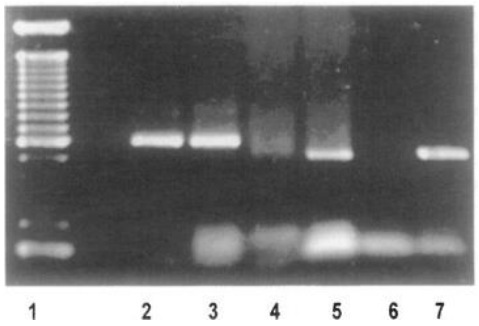
bulk RNA-Seq: different technologies and data analysis

Overview

- Historical development of bulk RNA-Seq
- Types of bulk RNA-Seq
- Focus on 3’ methods
- Common expressions: read length, coverage and depth
- Quality control
- Alignment and Mapping
- Contamination analysis
- Quantifying gene expression

Historical overview

1970s

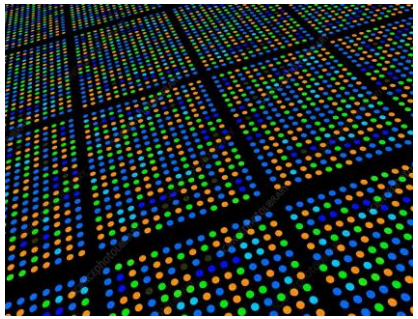


Northern Blot

1980s

qPCR

1990s



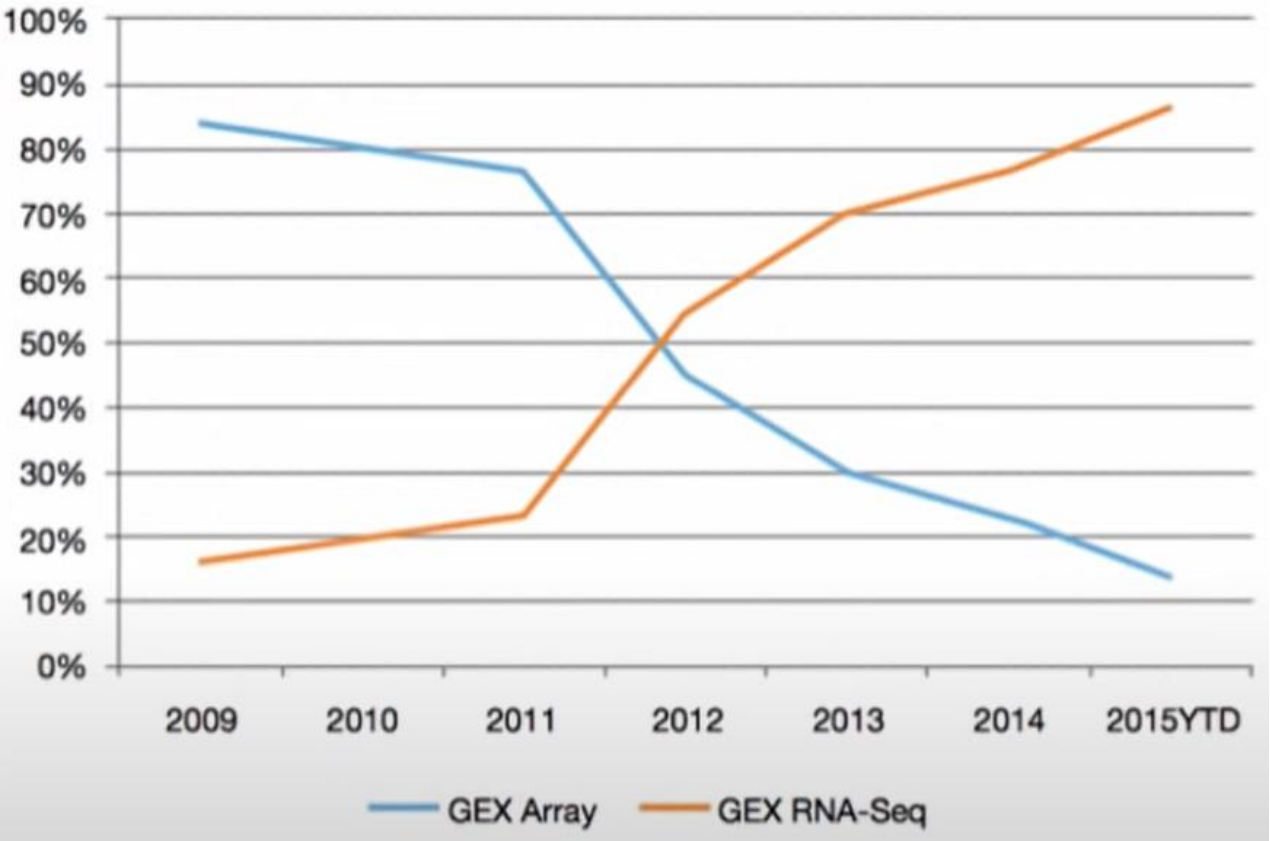
Microarray

present

RNA-Seq

RNA-Seq has replaced microarrays

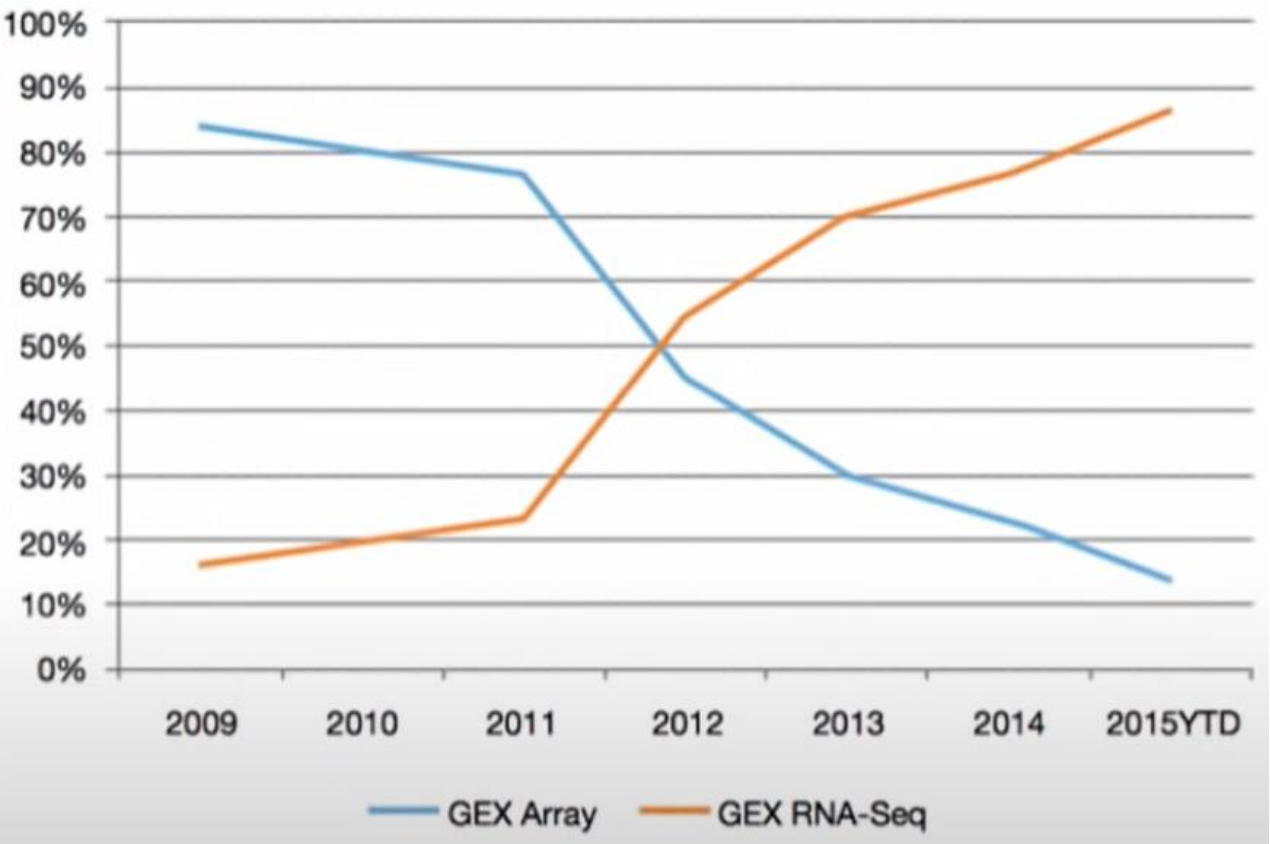
NIH Funding for Gene Expression Studies – New Grants



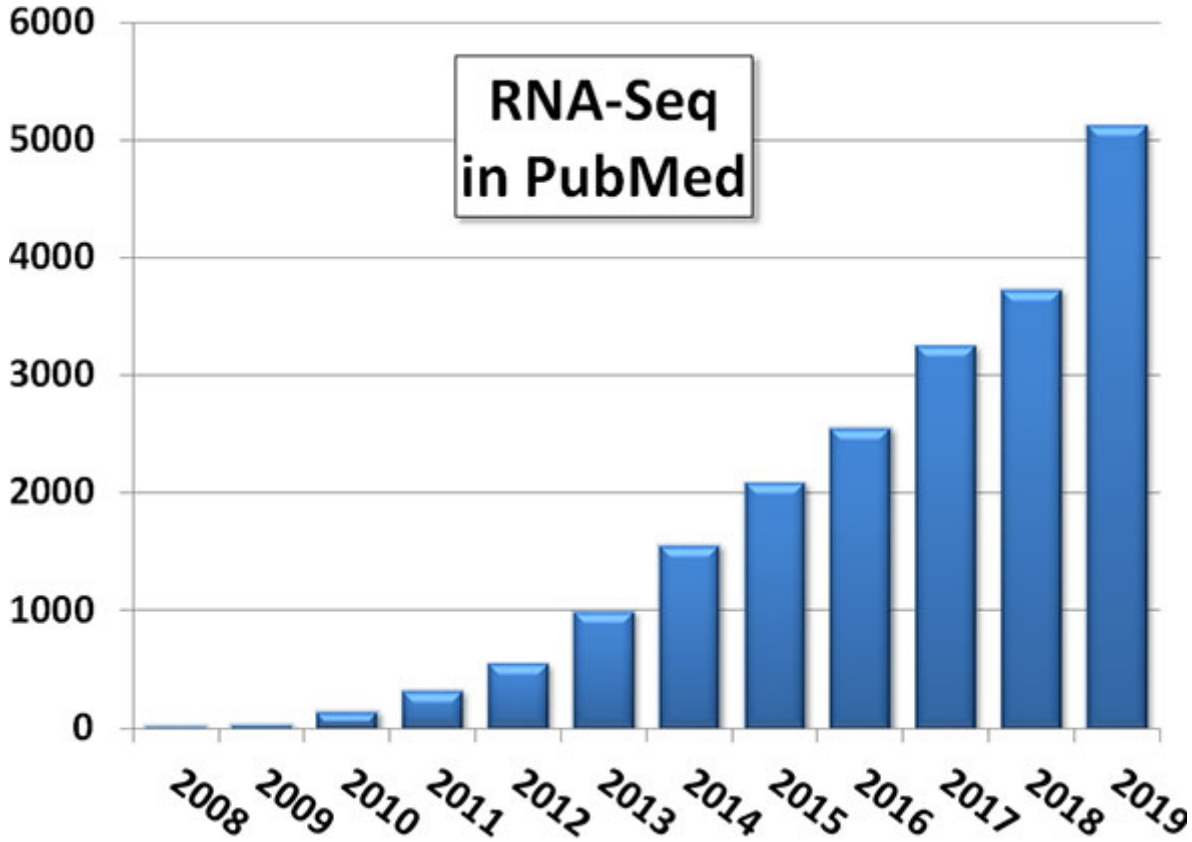
https://www.youtube.com/watch?v=7BLS_YY9HeM&ab_channel=IowaInstituteofHumanGenetics

RNA-Seq has replaced microarrays

NIH Funding for Gene Expression Studies – New Grants



RNA-Seq Publications



https://www.youtube.com/watch?v=7BLS_YY9HeM&ab_channel=IowaInstituteofHumanGenetics

Types of bulk RNA-Seq

mRNA-Seq

total RNA-Seq

small RNA-Seq

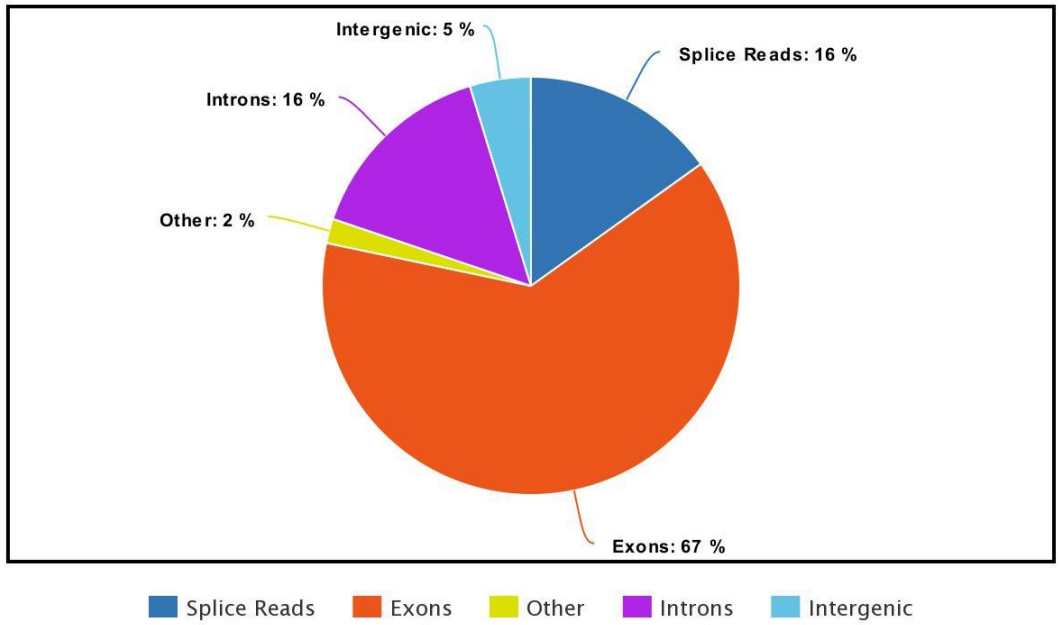
Types of bulk RNA-Seq

mRNA-Seq

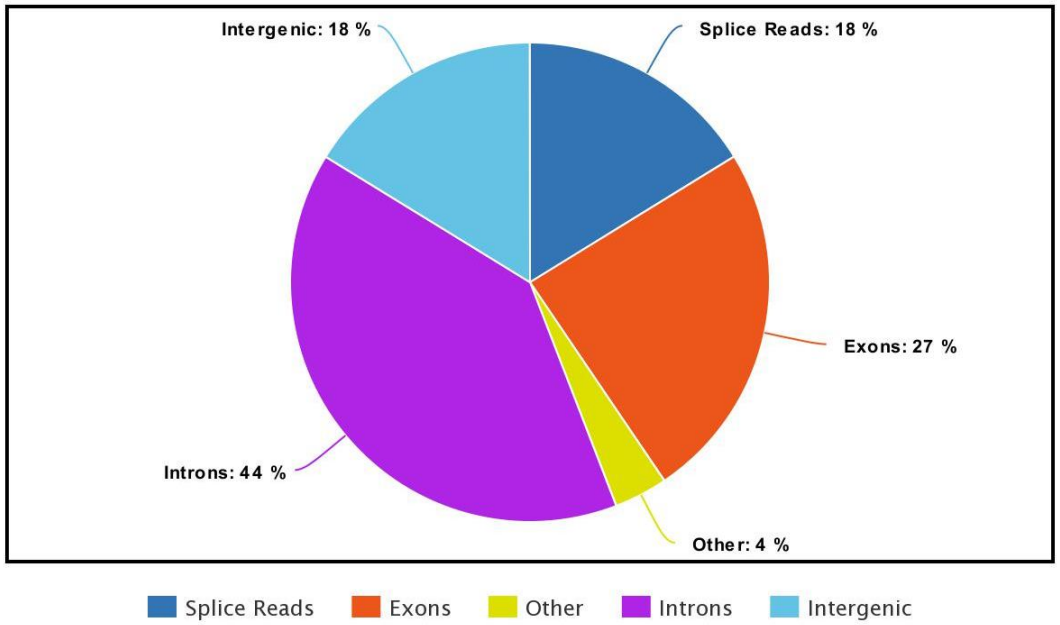
total RNA-Seq

small RNA-Seq

TruSeq Stranded mRNA

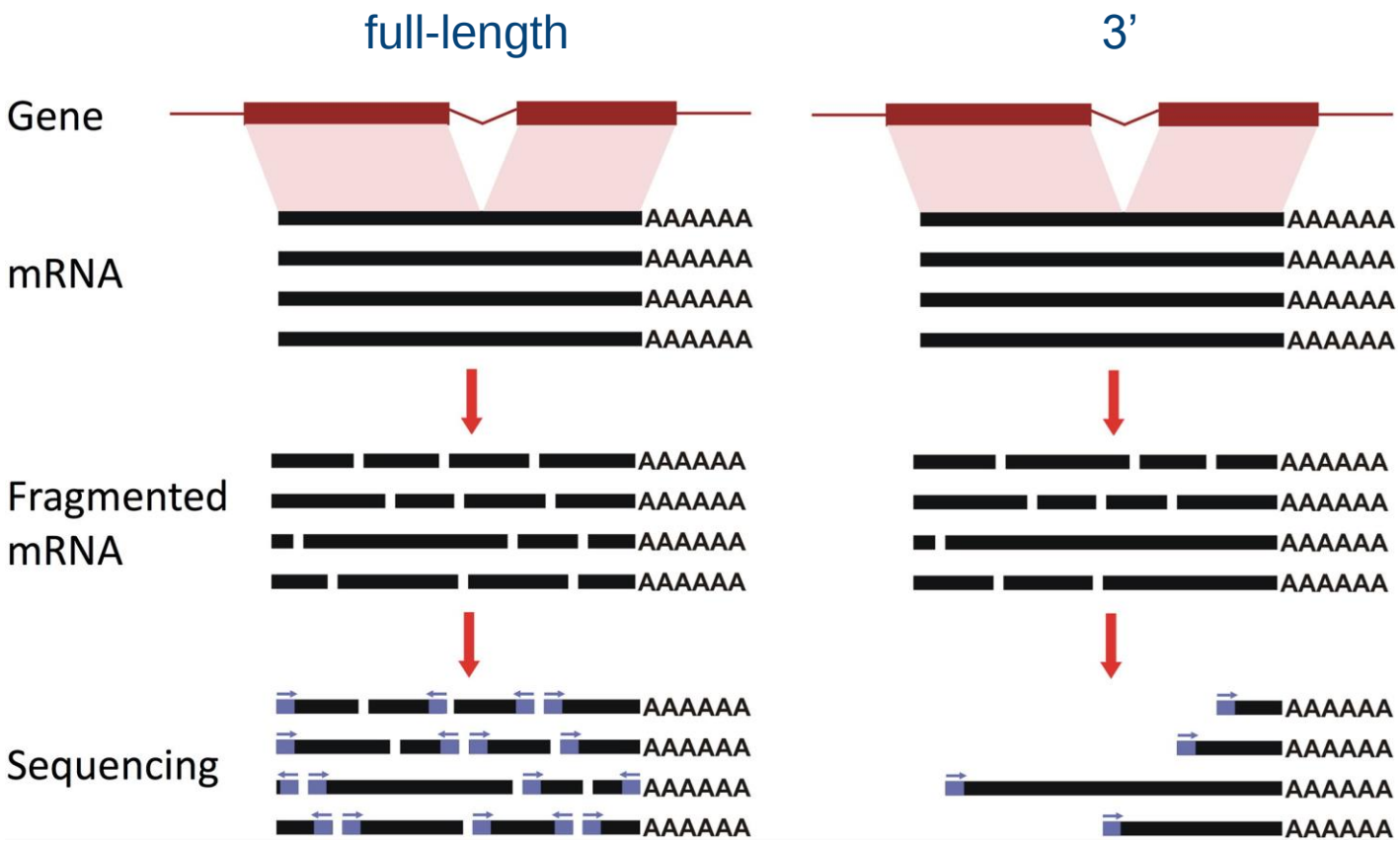


Total TNA-Seq w/RiboZero



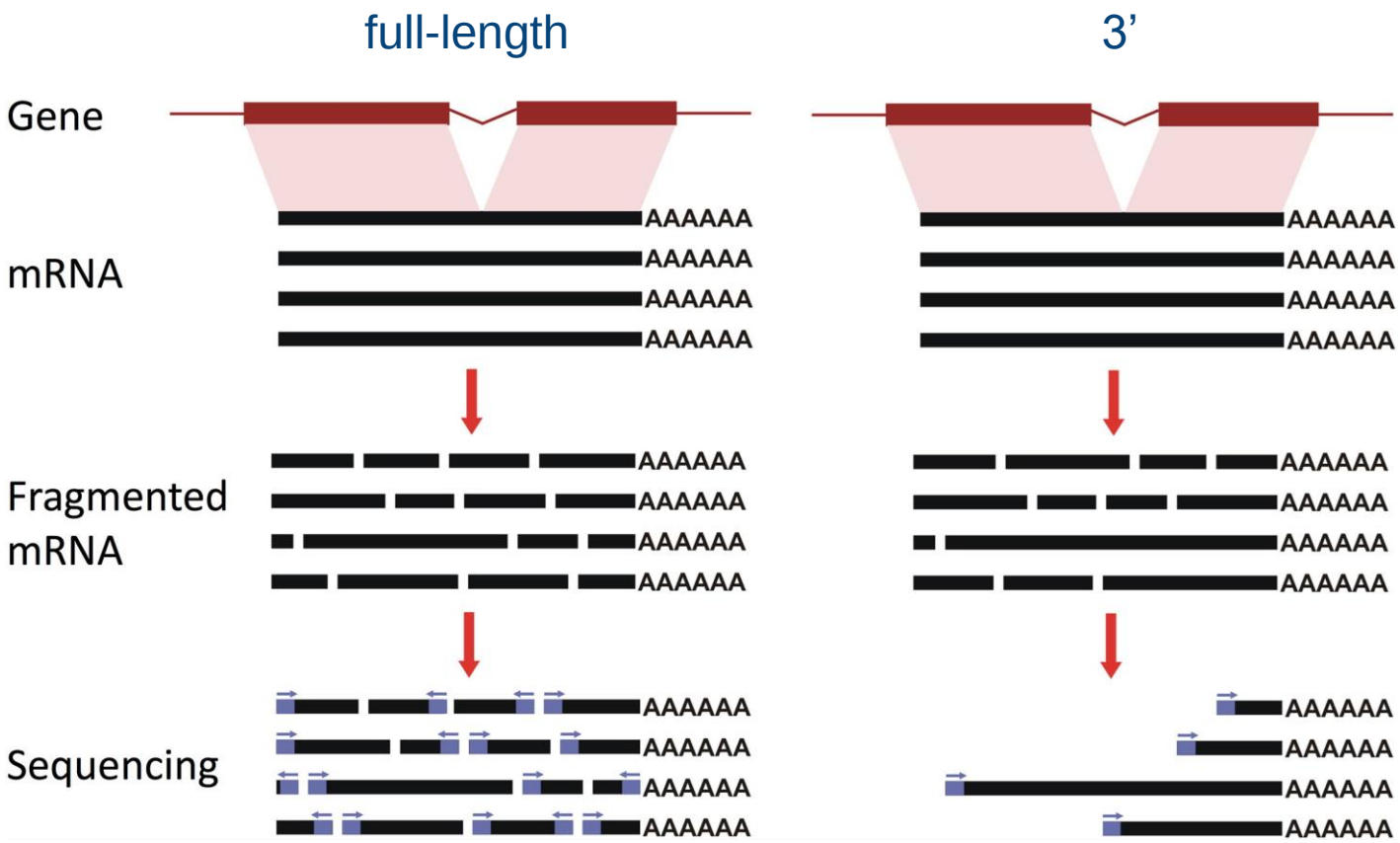
Adapted from https://www.youtube.com/watch?v=7BLS_YY9HeM&ab_channel=IowaInstituteofHumanGenetics

3’ vs full-length RNA-Seq



Tandonnet S. and Torres T. T., Genomics Data 2017

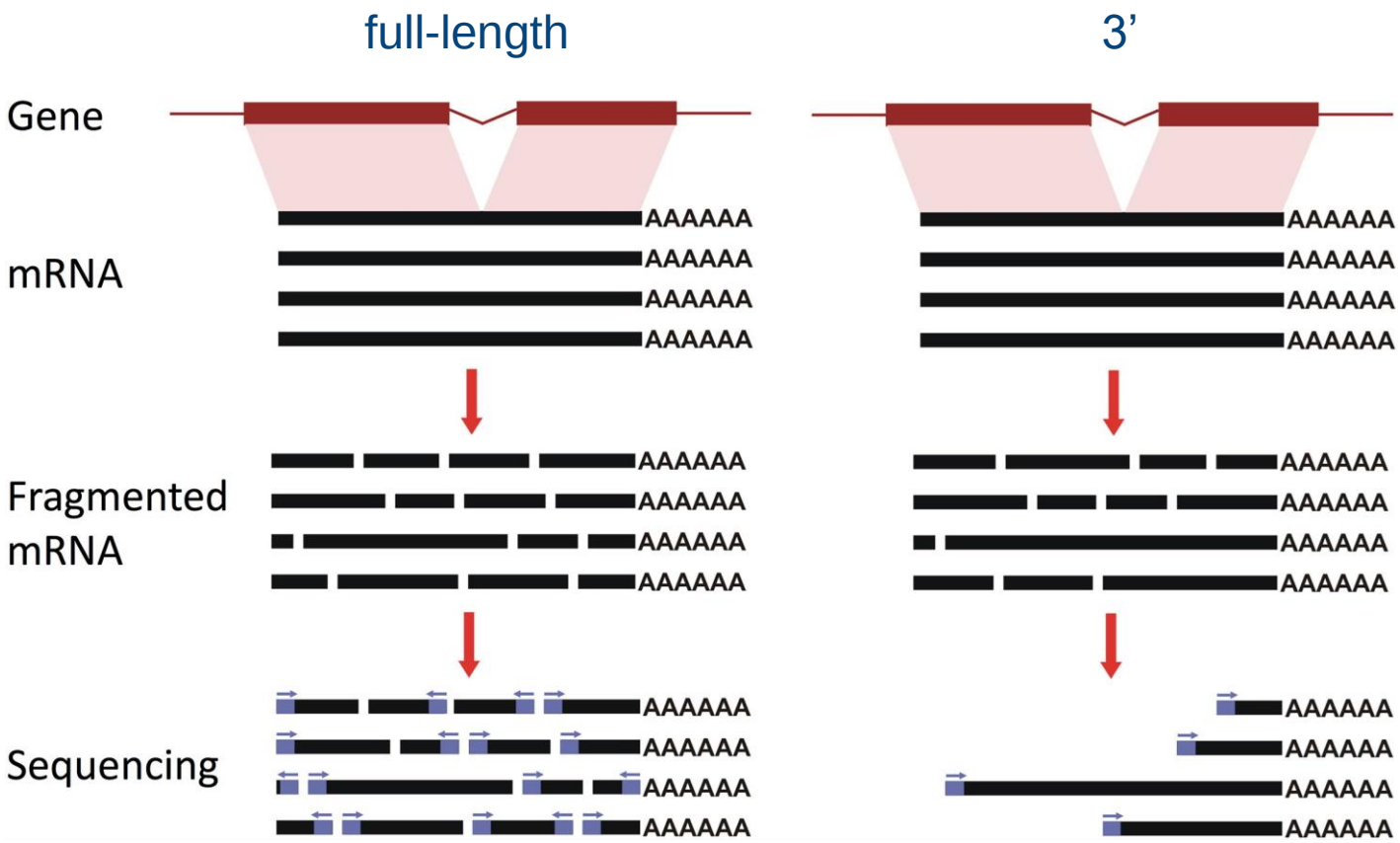
3’ vs full-length RNA-Seq



- In the classic RNA-Seq procedure, RNA is fragmented and converted in cDNA using small primers of random sequence
- In 3’ RNA-Seq library, mRNA molecules are randomly fragmented, but only 3’ portion of an mRNA is sequenced, using polyT oligos

Tandonnet S. and Torres T. T., Genomics Data 2017

3’ vs full-length RNA-Seq



Example of 3’ method: QuantSeq

Pros

- Low input and low quality samples
- Faster library preparation protocol
- Cost saving

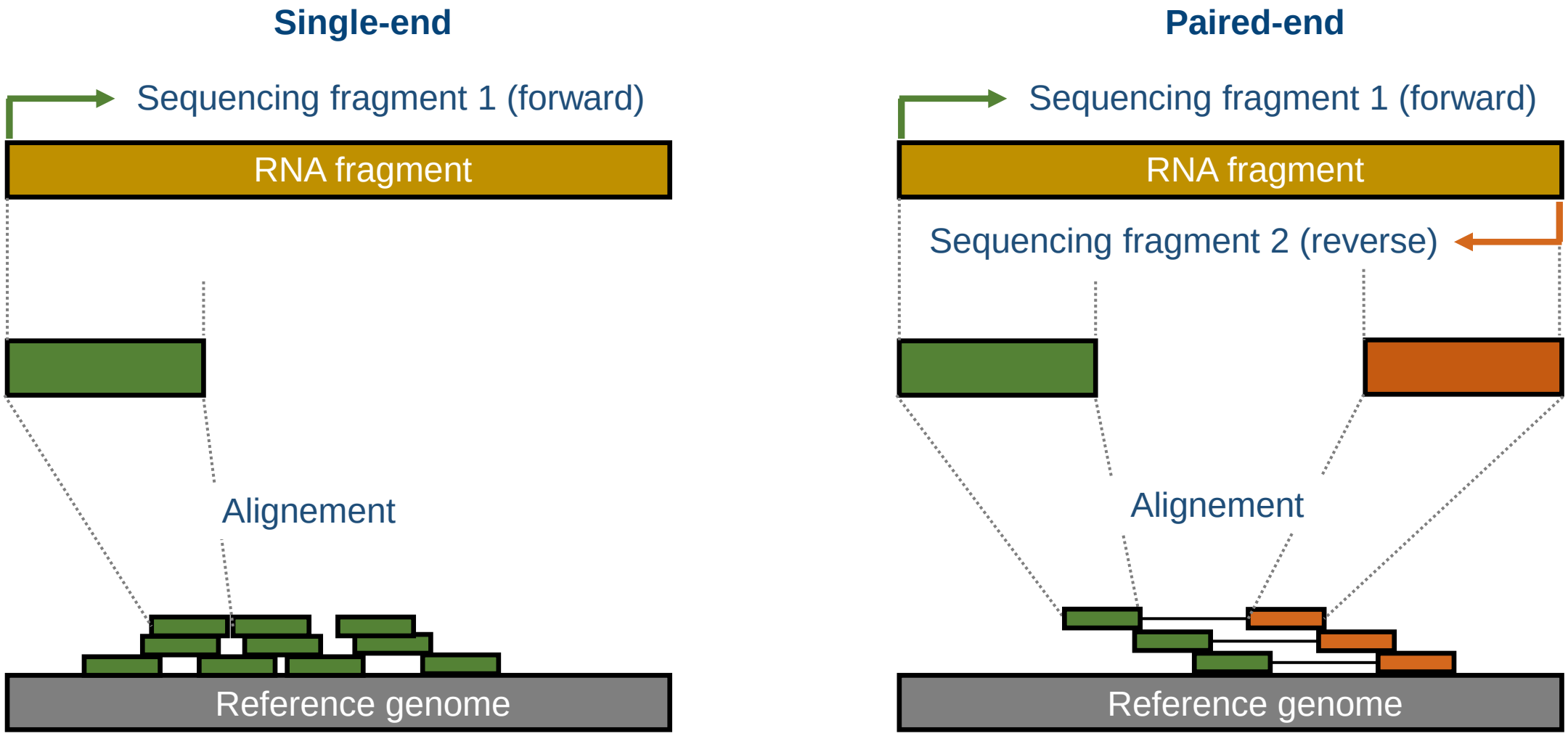
Cons

- No splicing information
- Requires reference genome
- Only eukaryotic samples (requires polyA)

- In the classic RNA-Seq procedure, RNA is fragmented and converted in cDNA using small primers of random sequence
- In 3’ RNA-Seq library, mRNA molecules are randomly fragmented, but only 3’ portion of an mRNA is sequenced, using polyT oligos

Tandonnet S. and Torres T. T., Genomics Data 2017

Single-end and Paired-end sequencing



Read length

Number of base pairs (bp) sequenced from a DNA/RNA fragment

It corresponds directly to reagents used on NGS instruments

Read length

Number of base pairs (bp) sequenced from a DNA/RNA fragment

It corresponds directly to reagents used on NGS instruments

75 bp

Short reads: gene expression profiling

100 bp

Longer reads: new transcriptome, WGS eukaryote

150 bp

Longest reads: amplicon sequencing, metagenomics

Coverage and depth

reads

$$C = \frac{\text{\# sequenced bases (= \# bases of all mapped reads)}}{\text{\# bases of reference}}$$

Coverage: number of reads that align to, or «cover» a known reference. It describes how often, in average, a reference sequence is covered by bases from the reads

Adapted from www.ecseq.com

Coverage and depth

reads

$$C = \frac{\text{\# sequenced bases (= \# bases of all mapped reads)}}{\text{\# bases of reference}}$$

Coverage: number of reads that align to, or «cover» a known reference. It describes how often, in average, a reference sequence is covered by bases from the reads

reads

Sequencing depth = total number of reads

Depth: total number of usable reads from the sequencing machine (usually used in the unit of «number of reads», in millions). Specially used for RNA-Seq

Adapted from www.ecseq.com

Coverage and depth

reads

$$C = \frac{\#\text{ sequenced bases (= \# bases of all mapped reads)}}{\#\text{ bases of reference}}$$

Coverage: number of reads that align to, or «cover» a known reference. It describes how often, in average, a reference sequence is covered by bases from the reads

reads

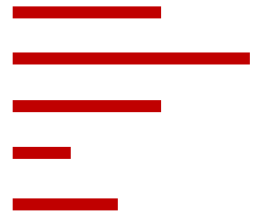
Sequencing depth = total number of reads

Depth: total number of usable reads from the sequencing machine (usually used in the unit of «number of reads», in millions). Specially used for RNA-Seq

What is a good coverage for NGS projects?

Adapted from www.ecseq.com

Coverage and depth



reads

C =

Sequencing Method	Recommended Coverage
Whole genome sequencing (WGS)	30× to 50× for human WGS (depending on application and statistical model)
Whole-exome sequencing	100×
RNA sequencing	Usually calculated in terms of numbers of millions of reads to be sampled. Detecting rarely expressed genes often requires an increase in the depth of coverage.
ChIP-Seq	100×

umber of reads that
cover» a known
describes how
rage, a reference
covered by bases
ds

reads

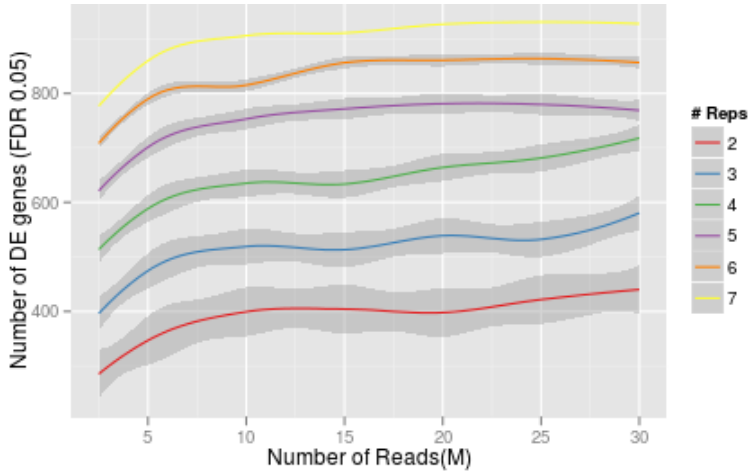
number of usable
ne sequencing
ually used in the
ber of reads», in
ecially used for

What is a good coverage for NGS projects?

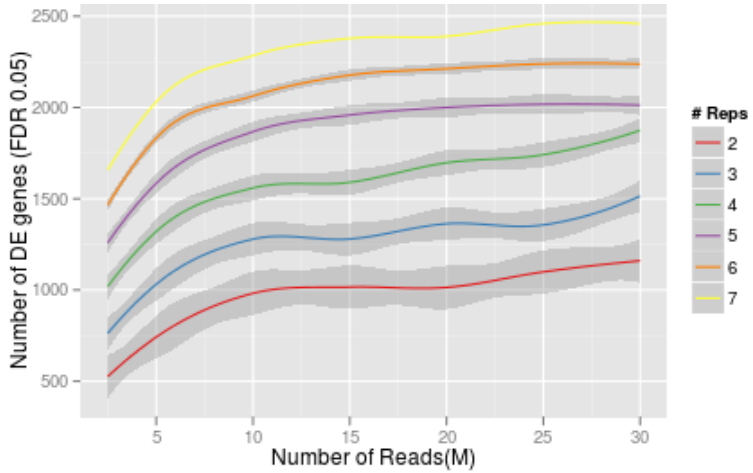
Adapted from www.ecseq.com

The importance of sequencing depth

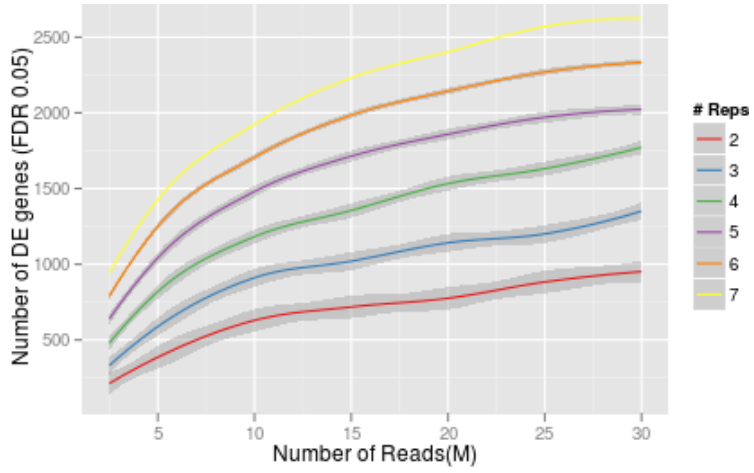
#DE genes vs. Reps vs #Reads (High Expr)



#DE genes vs. Reps vs #Reads (Medium Expr)



#DE genes vs. Reps vs #Reads (Low Expr)

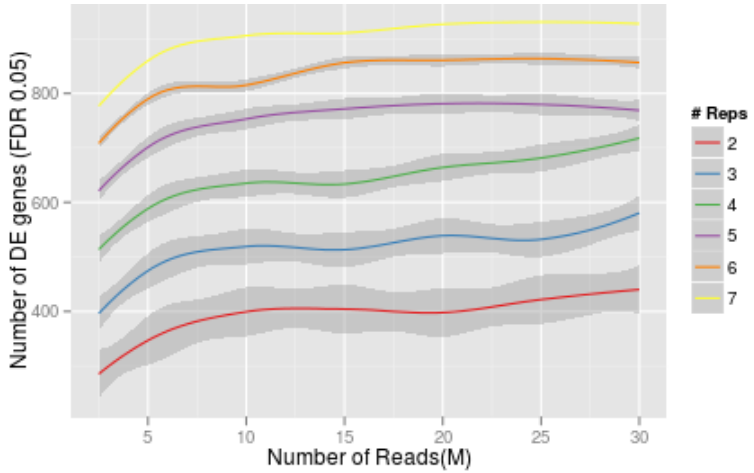


For **High Expressed genes**: Increasing sequencing depth has little effect on increasing number of DE genes detected, while biological replicates are more beneficial

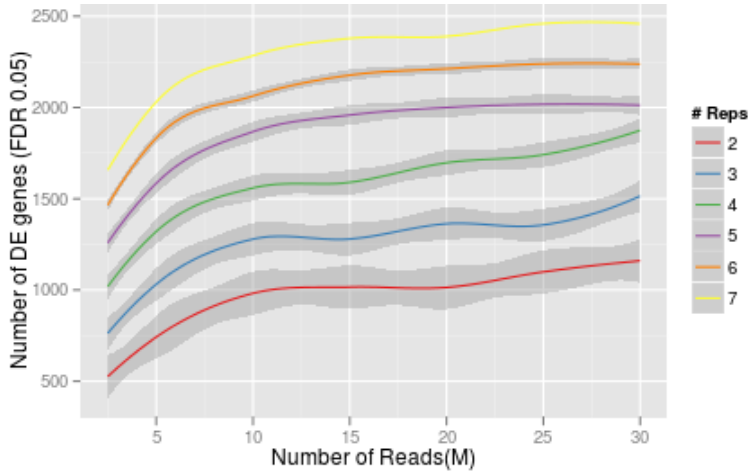
Liu et al., Bioinformatics 2014

The importance of sequencing depth

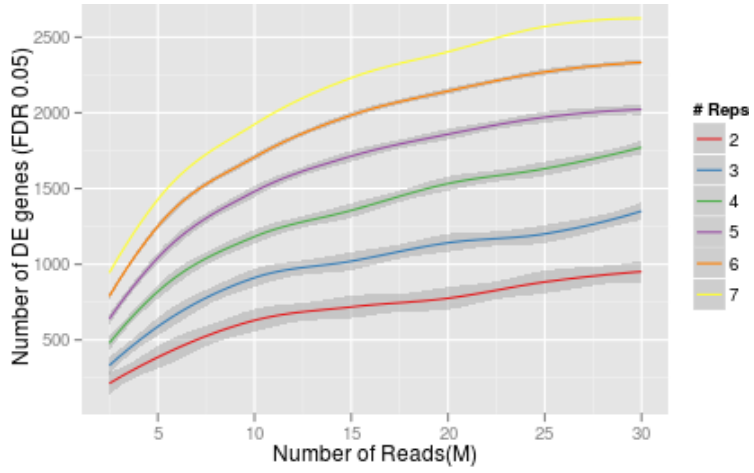
#DE genes vs. Reps vs #Reads (High Expr)



#DE genes vs. Reps vs #Reads (Medium Expr)



#DE genes vs. Reps vs #Reads (Low Expr)



For **High Expressed genes**: Increasing sequencing depth has little effect on increasing number of DE genes detected, while biological replicates are more beneficial

For **Low Expressed genes**: Both sequencing depth and biological replicates increases power to detect DE genes

Liu et al., Bioinformatics 2014

Machine output: .fastq files

```
@A00560:168:H5J7CDRXY:1:2101:11550:1016 1:N:0:NTTACC+NGGAGA
CNCTCTACCCCCAATGCCTCCAACCATAGTAATTCAGCCCTTGCCAAAGACAATAAAGCAATTCTCAAGTAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCA
+
F#:FFF,,FF:FFFF,FFFFFFFFFFFFFFF,FFFFF:FFFFF,FFFFFFF:FFFFF::F::F:FFFFFF:FFFFFFFFFFFFFFFFF,,,FFF,FFF:FF
@A00560:168:H5J7CDRXY:1:2101:8305:1031 1:N:0:NTTACC+NGGAGA
TNGTATTTGAGTGTTTTTGCTTGCATGTGTGTCTTTGCGCCATGTTCCTGCTGGCATCTGAATGAACAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCACA
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF,,FFFF:F:FF,,
@A00560:168:H5J7CDRXY:1:2101:21884:1031 1:N:0:NTTACC+NGGAGA
CNGTGTTTCATTTAGGATACTTTTTGCTGCAAGATTCTCAAGCCACACACAGTAACTGGTAGGTCACTGAGTAGTCTGAGCCCCGAGCACCTCAGCGAGCT
+
F#FFFFF:FFFFFFF,F:FFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFF,F:FFFFFFFFFFFFFFFFFFFFFF:FFFFFF,F:FFFF
@A00560:168:H5J7CDRXY:1:2101:32588:1031 1:N:0:NTTACC+NGGAGA
TNGTCTAATGAACAGTAATGTATGCTCTCTCTAATTGTTCAGAGTCTTATAAGGAAAAAAAAAAAAAAAAAAAGATCGGAATAGCACACGTCTGAACTCC
+
F#FFFFFFFFFFFFFFF:FF:F:FFFFFFFFFFFFFFF:,,FF:FFFFF,FFFFF,FFFFFFFFFFFFFFFFFF:::F:F:F:,::::,,FF,:FFF,,,FF
@A00560:168:H5J7CDRXY:1:2101:4878:1047 1:N:0:NTTACC+NGGAGA
CTAAAGAGGCCGTTTATCTTTGTAAACACAAAACATTTTTGCTTTCTCCGGTTTTATGTTAATGGCGAAAGAATGGAAGCGAATAAAGTTTTACTGATTTT
+
F:FFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF:FFFF:FFFFFFFFFFFFFFFF:F,FFFF,FFFFFFFFFFFFFFFFF
@A00560:168:H5J7CDRXY:1:2101:5746:1047 1:N:0:NTTACC+NGGAGA
CNCCTCAACCCTAATTAAAGTCTCTTCCTGCCCCTTCGGGGCTGCAAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCAAACGTCTGAACTCCAGTCACGTTA
+
F#FFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFF,,FFF,:FF:FFFF,:FFFF:,F::,FFFFF,FFFFFF
```

Machine output: .fastq files

```
@A00560:168:H5J7CDRXY:1:2101:11550:1016 1:N:0:NTTACC+NGGAGA
CNCTCTACCCCCAATGCCTCCAACCATAGTAATTCAGCCCTTGCCAAAGACAATAAAGCAATTCTCAAGTAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCA
+
F#:FFF,,FF:FFFF,FFFFFFFFFFFFFFF,FFFFF:FFFFF,FFFFFFF:FFFFF::F::F:FFFFFF:FFFFFFFFFFFFFFFFF,,,FFF,FFF:FF
@A00560:168:H5J7CDRXY:1:2101:8305:1031 1:N:0:NTTACC+NGGAGA
TNGTATTTGAGTGTTTTTGCTTGCATGTGTGTCTTTGCGCCATGTTCCTGCTGGCATCTGAATGAACAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCACA
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF,,FFFF:F:FF,,
@A00560:168:H5J7CDRXY:1:2101:21884:1031 1:N:0:NTTACC+NGGAGA
CNGTGTTTCATTTAGGATACTTTTTGCTGCAAGATTCTCAAGCCACACACAGTAACTGGTAGGTCACTGAGTAGTCTGAGCCCCGAGCACCTCAGCGAGCT
+
F#FFFFF:FFFFFFF,F:FFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFF,F:FFFFFFFFFFFFFFFFFFFFFFF:FFFFFF,F:FFFF
@A00560:168:H5J7CDRXY:1:2101:32588:1031 1:N:0:NTTACC+NGGAGA
TNGTCTAATGAACAGTAATGTATGCTCTCTCTAATTGTTCAGAGTCTTATAAGGAAAAAAAAAAAAAAAAAAAGATCGGAATAGCACACGTCTGAACTCC
+
F#FFFFFFFFFFFFFFF:FF:F:FFFFFFFFFFFFFFF:,,FF:FFFFF,FFFFF,FFFFFFFFFFFFFFFFFF:::F:F:F:,::::,,FF,:FFF,,,FF
@A00560:168:H5J7CDRXY:1:2101:4878:1047 1:N:0:NTTACC+NGGAGA
CTAAAGAGGCCGTTTATCTTTGTAAACACAAAACATTTTTGCTTTCTCCGGTTTTATGTTAATGGCGAAAGAATGGAAGCGAATAAAGTTTTACTGATTTT
+
F:FFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF:FFFF:FFFFFFFFFFFFFFFF:F,FFFF,FFFFFFFFFFFFFFFFF
@A00560:168:H5J7CDRXY:1:2101:5746:1047 1:N:0:NTTACC+NGGAGA
CNCCTCAACCCTAATTAAAGTCTCTTCCTGCCCCTTCGGGGCTGCAAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCAAACGTCTGAACTCCAGTCACGTTA
+
F#FFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFF,,FFF,:FF:FFFF,:FFFF:,F::,FFFFF,FFFFFF
```

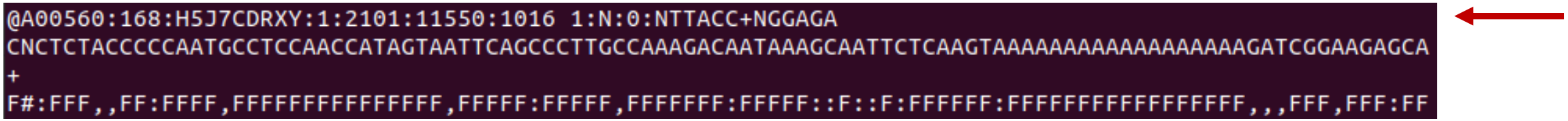
Machine output: .fastq files

```
@A00560:168:H5J7CDRXY:1:2101:11550:1016 1:N:0:NTTACC+NGGAGA
CNCTCTACCCCCAATGCCTCCAACCATAGTAATTCAGCCCTTGCCAAAGACAATAAAGCAATTCTCAAGTAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCA
+
F#:FFF..FF:FFFF.FFFFFFFFFFFFFFF.FFFFF:FFFFF.FFFFFFF:FFFFF::F::F:FFFFFF:FFFFFFFFFFFFFFFFF...FFF.FFF:FF
@A00560:168:H5J7CDRXY:1:2101:8305:1031 1:N:0:NTTACC+NGGAGA
TNGTATTTGAGTGTTTTTGCTTGCATGTGTGTCTTTGCGCCATGTTCCTGCTGGCATCTGAATGAACAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCACA
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF..FFFF:F:FF..
@A00560:168:H5J7CDRXY:1:2101:21884:1031 1:N:0:NTTACC+NGGAGA
CNGTGTTTCATTTAGGATACTTTTTGCTGCAAGATTCTCAAGCCACACACAGTAACTGGTAGGTCACTGAGTAGTCTGAGCCCCGAGCACCTCAGCGAGCT
+
F#FFFFF:FFFFFFF,F:FFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFF,F:FFFFFFFFFFFFFFFFFFFFFFF:FFFFFF,F:FFFF
@A00560:168:H5J7CDRXY:1:2101:32588:1031 1:N:0:NTTACC+NGGAGA
TNGTCTAATGAACAGTAATGTATGCTCTCTCTAATTGTTCAGAGTCTTATAAGGAAAAAAAAAAAAAAAAAAAGATCGGAATAGCACACGTCTGAACTCC
+
F#FFFFFFFFFFFFFFF:FF:F:FFFFFFFFFFFFFFF:,,FF:FFFFF,FFFFF,FFFFFFFFFFFFFFFFFF:::F:F:F:,::::,,FF,:FFF,,,FF
@A00560:168:H5J7CDRXY:1:2101:4878:1047 1:N:0:NTTACC+NGGAGA
CTAAAGAGGCCGTTTATCTTTGTAAACACAAAACATTTTTGCTTTCTCCGGTTTTATGTTAATGGCGAAAGAATGGAAGCGAATAAAGTTTTACTGATTTT
+
F:FFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFF:FFFF:FFFFFFFFFFFFFFFF:F,FFFF,FFFFFFFFFFFFFFFFF
@A00560:168:H5J7CDRXY:1:2101:5746:1047 1:N:0:NTTACC+NGGAGA
CNCCTCAACCCTAATTAAAGTCTCTTCCTGCCCCTTCGGGGCTGCAAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCAAACGTCTGAACTCCAGTCACGTTA
+
F#FFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFF,,FFF,:FF:FFFF,:FFFF:,F::,FFFFF,FFFFFF
```

Machine output: .fastq files

```
@A00560:168:H5J7CDRXY:1:2101:11550:1016 1:N:0:NTTACC+NGGAGA
CNCTCTACCCCCAATGCCTCCAACCATAGTAATTCAGCCCTTGCCAAAGACAATAAAGCAATTCTCAAGTAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCA
+
F#:FFF,,FF:FFFF,FFFFFFFFFFFFFFF,FFFFF:FFFFF,FFFFFFF:FFFFF::F::F:FFFFFF:FFFFFFFFFFFFFFFFF,,,FFF,FFF:FF
@A00560:168:H5J7CDRXY:1:2101:8305:1031 1:N:0:NTTACC+NGGAGA
TNGTATTTGAGTGTTTTTGCTTGCATGTGTGTCTTTGCGCCATGTTCCTGCTGGCATCTGAATGAACAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCACA
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF,,FFFF:F:FF,,
@A00560:168:H5J7CDRXY:1:2101:21884:1031 1:N:0:NTTACC+NGGAGA
CNGTGTTTCATTTAGGATACTTTTTGCTGCAAGATTCTCAAGCCACACACAGTAACTGGTAGGTCACTGAGTAGTCTGAGCCCCGAGCACCTCAGCGAGCT
+
F#FFFFF:FFFFFFF,F:FFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFF,F:FFFFFFFFFFFFFFFFFFFFFFF:FFFFFF,F:FFFF
@A00560:168:H5J7CDRXY:1:2101:32588:1031 1:N:0:NTTACC+NGGAGA
TNGTCTAATGAACAGTAATGTATGCTCTCTCTAATTGTTCAGAGTCTTATAAGGAAAAAAAAAAAAAAAAAAAGATCGGAATAGCACACGTCTGAACTCC
+
F#FFFFFFFFFFFFFFF:FF:F:FFFFFFFFFFFFFFF:,,FF:FFFFF,FFFFF,FFFFFFFFFFFFFFFFFF:::F:F:F:,::::,,FF,:FFF,,,FF
@A00560:168:H5J7CDRXY:1:2101:4878:1047 1:N:0:NTTACC+NGGAGA
CTAAAGAGGCCGTTTATCTTTGTAAACACAAAACATTTTTGCTTTCTCCGGTTTTATGTTAATGGCGAAAGAATGGAAGCGAATAAAGTTTTACTGATTTT
+
F:FFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF:FFFF:FFFFFFFFFFFFFFFF:F,FFFF,FFFFFFFFFFFFFFFFF
@A00560:168:H5J7CDRXY:1:2101:5746:1047 1:N:0:NTTACC+NGGAGA
CNCCTCAACCCTAATTAAAGTCTCTTCCTGCCCCTTCGGGGCTGCAAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCAAACGTCTGAACTCCAGTCACGTTA
+
F#FFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFF,,FFF,:FF:FFFF,:FFFF:,F::,FFFFF,FFFFFF
```

Machine output: .fastq files



Line 1: always begins with a @, followed by a sequence identifier and an optional description

Adapted from scilifelab.github.io

Machine output: .fastq files

```
@A00560:168:H5J7CDRXY:1:2101:11550:1016 1:N:0:NTTACC+NGGAGA
CNCTCTACCCCCAATGCCTCCAACCATAGTAATTCAGCCCTTGCCAAAGACAATAAAGCAATTCTCAAGTAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCA
+
F#:FFF,,FF:FFFF,FFFFFFFFFFFFFFF,FFFFF:FFFFF,FFFFFFF:FFFFF::F::F:FFFFFF:FFFFFFFFFFFFFFFFF,,,FFF,FFF:FF
```

Line 1: always begins with a @, followed by a sequence identifier and an optional description

Line 2: raw sequence letters

Adapted from scilifelab.github.io

Machine output: .fastq files

```
@A00560:168:H5J7CDRXY:1:2101:11550:1016 1:N:0:NTTACC+NGGAGA
CNCTCTACCCCCAATGCCTCCAACCATAGTAATTCAGCCCTTGCCAAAGACAATAAAGCAATTCTCAAGTAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCA
+
F#:FFF,,FF:FFFF,FFFFFFFFFFFFFFF,FFFFF:FFFFF,FFFFFFF:FFFFF::F::F:FFFFFF:FFFFFFFFFFFFFFFFF,,,FFF,FFF:FF
```

Line 1: always begins with a @, followed by a sequence identifier and an optional description

Line 2: raw sequence letters

Line 3: begins with a + and it optionally followed by the same sequence identifier

Adapted from scilifelab.github.io

Machine output: .fastq files

```
@A00560:168:H5J7CDRXY:1:2101:11550:1016 1:N:0:NTTACC+NGGAGA
CNCTCTACCCCCAATGCCTCCAACCATAGTAATTCAGCCCTTGCCAAAGACAATAAAGCAATTCTCAAGTAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCA
+
F#:FFF,,FF:FFFF,FFFFFFFFFFFFFFF,FFFFF:FFFFF,FFFFFFF:FFFFF::F::F:FFFFFF:FFFFFFFFFFFFFFFFF,,,FFF,FFF:FF
```

Line 1: always begins with a @, followed by a sequence identifier and an optional description

Line 2: raw sequence letters

Line 3: begins with a + and it optionally followed by the same sequence identifier

Line 4: quality scores for the sequence in line 2. It must have the same number of symbols than line 2

Adapted from scilifelab.github.io

Machine output: .fastq files

```
@A00560:168:H5J7CDRXY:1:2101:11550:1016 1:N:0:NTTACC+NGGAGA
CNCTCTACCCCCAATGCCTCCAACCATAGTAATTCAGCCCTTGCCAAAGACAATAAAGCAATTCTCAAGTAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCA
+
F#:FFF,,FF:FFFF,FFFFFFFFFFFFFFF,FFFFF:FFFFF,FFFFFFF:FFFFF::F::F:FFFFFF:FFFFFFFFFFFFFFFFF,,,FFF,FFF:FF
```

Line 1: always begins with a @, followed by a sequence identifier and an optional description

Line 2: raw sequence letters

Line 3: begins with a + and it optionally followed by the same sequence identifier

Line 4: quality scores for the sequence in line 2. It must have the same number of symbols than line 2

The quality score is calculated by using the **Phred Quality Score**:
Q = -10log(P), where P is the probability of base calling being incorrect

Adapted from scilifelab.github.io

Machine output: .fastq files

```
@A00560:168:H5J7CDRXY:1:2101:11550:1016 1:N:0:NTTACC+NGGAGA
CNCTCTACCCCCAATGCCTCCAACCATAGTAATTCAGCCCTTGCCAAAGACAATAAAGCAATTCTCAAGTAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCA
+
F#:FFF,,FF:FFFF,FFFFFFFFFFFFFFF,FFFFF:FFFFF,FFFFFFF:FFFFF::F::F:FFFFFF:FFFFFFFFFFFFFFFFF,,,FFF,FFF:FF
```

Line 1: always begins with a @, followed by a sequence identifier and an optional description

Line 2: raw sequence letters

Line 3: begins with a + and it optionally followed by the same sequence identifier

Line 4: quality scores for the sequence in line 2. It must have the same number of symbols than line 2

The quality score is calculated by using the **Phred Quality Score**:
Q = -10log(P), where P is the probability of base calling being incorrect

Base-Calling of Automated Sequencer Traces Using*Phred*. I. Accuracy Assessment

Brent Ewing[1], LaDeana Hillier[2], Michael C. Wendl[2] and Phil Green[1,3]

Adapted from scilifelab.github.io

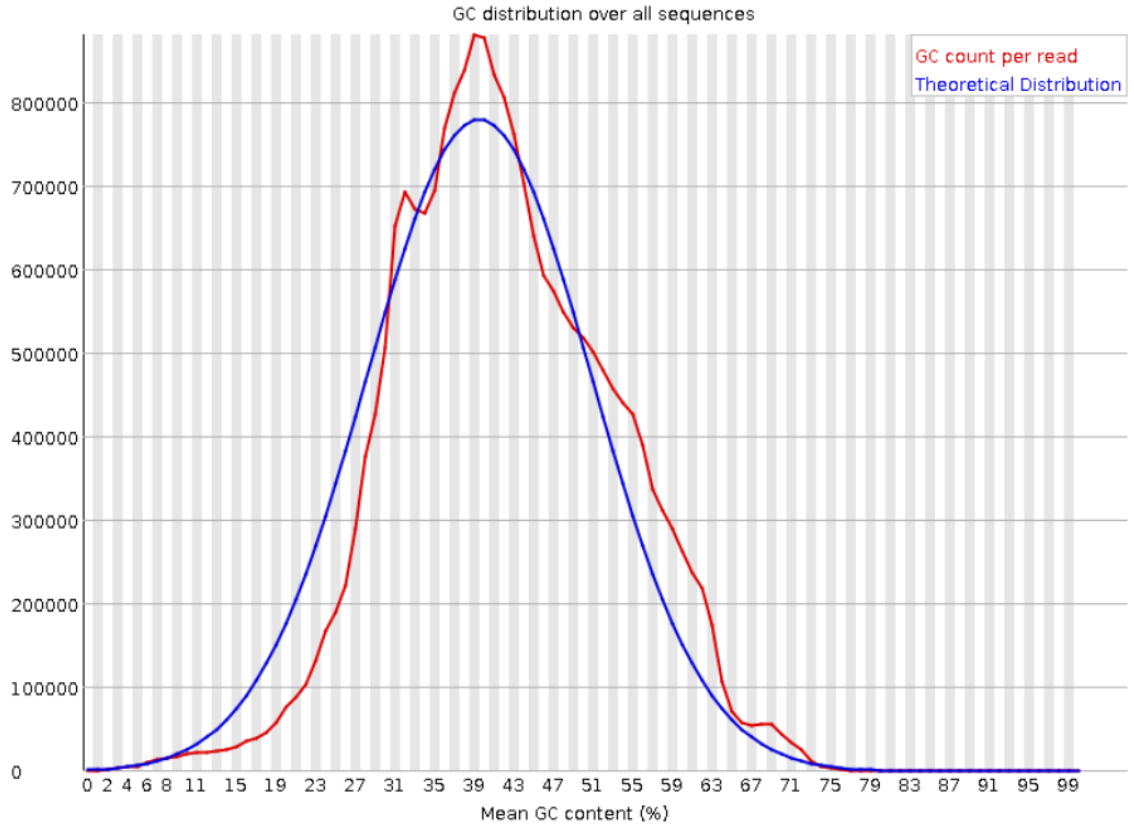
Quality control

Basic Statistics

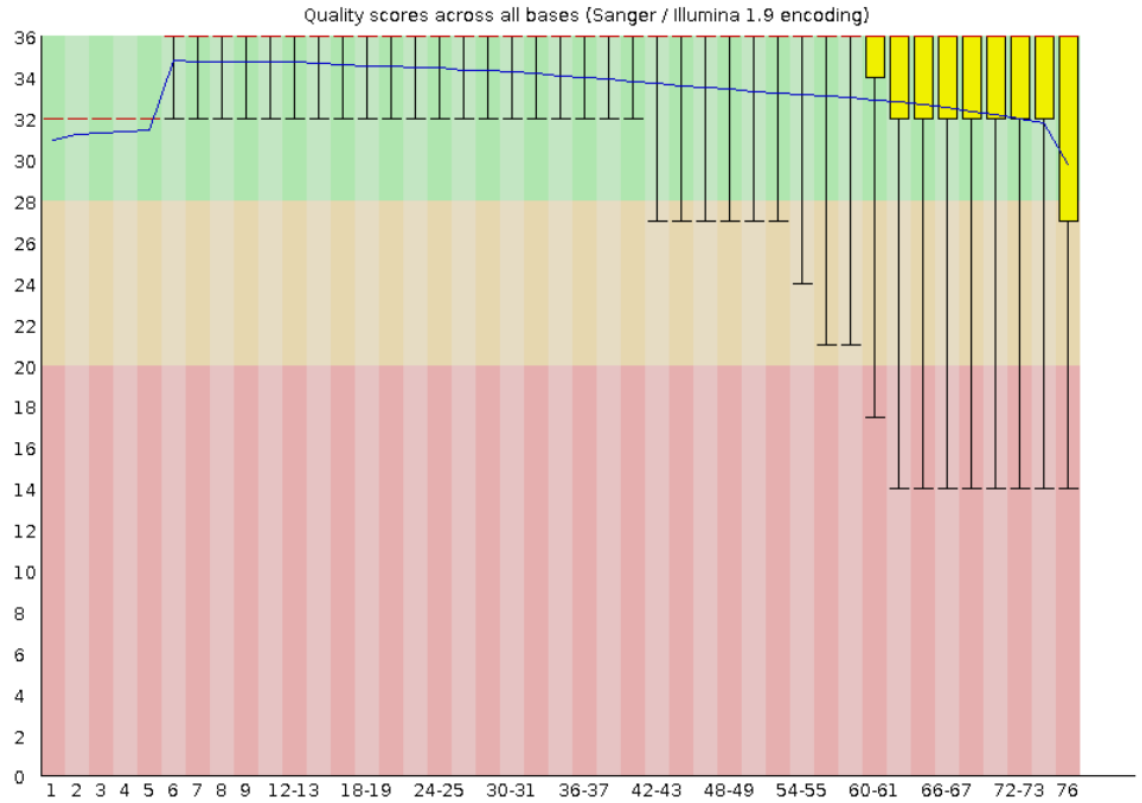
Measure	Value
Filename	R35_S54_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	16883937
Sequences flagged as poor quality	0
Sequence length	76
%GC	41

Quality control

Per sequence GC content



Per base sequence quality



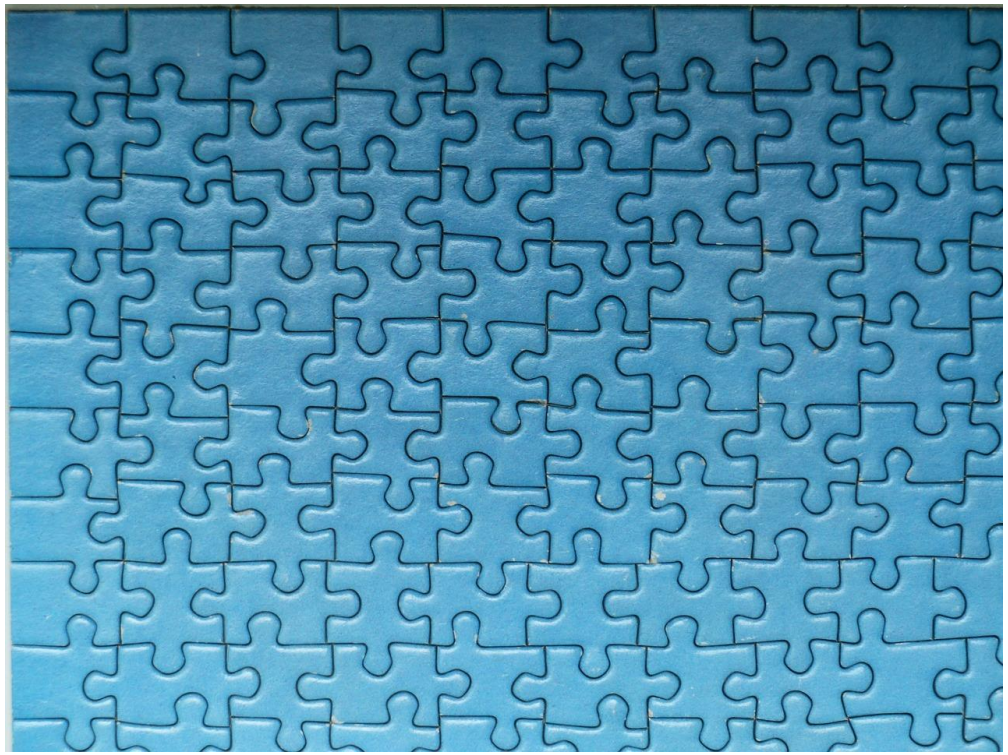
Alignment/mapping

What to do with all the collected reads?

Alignment/mapping

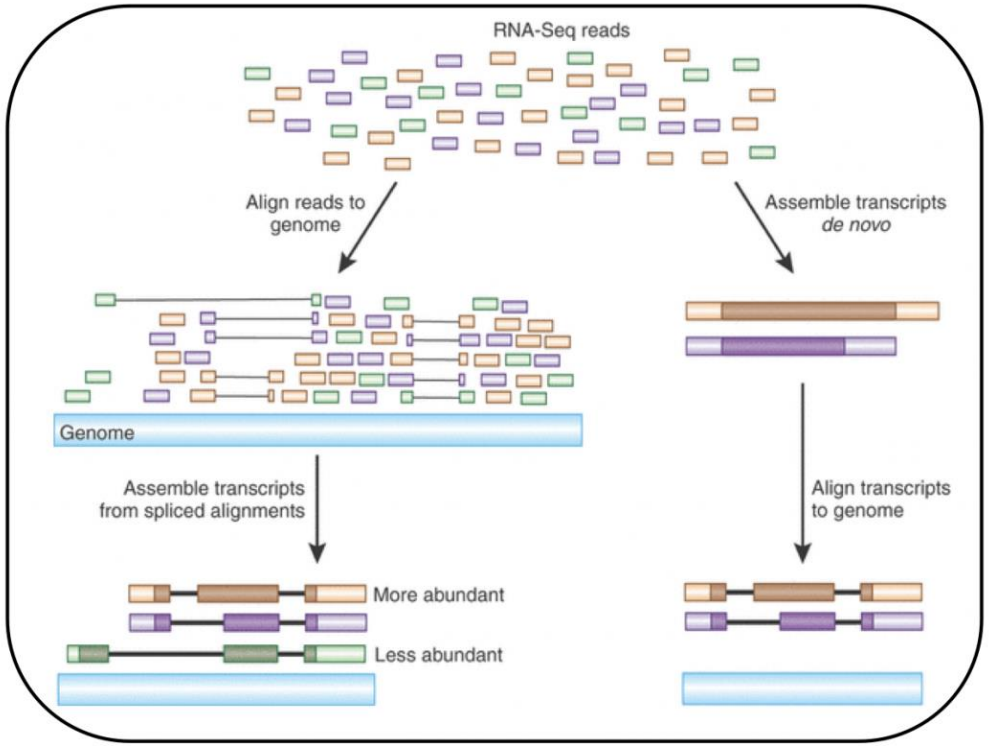
What to do with all the collected reads?

We need to understand which gene they come from, their position on the genome



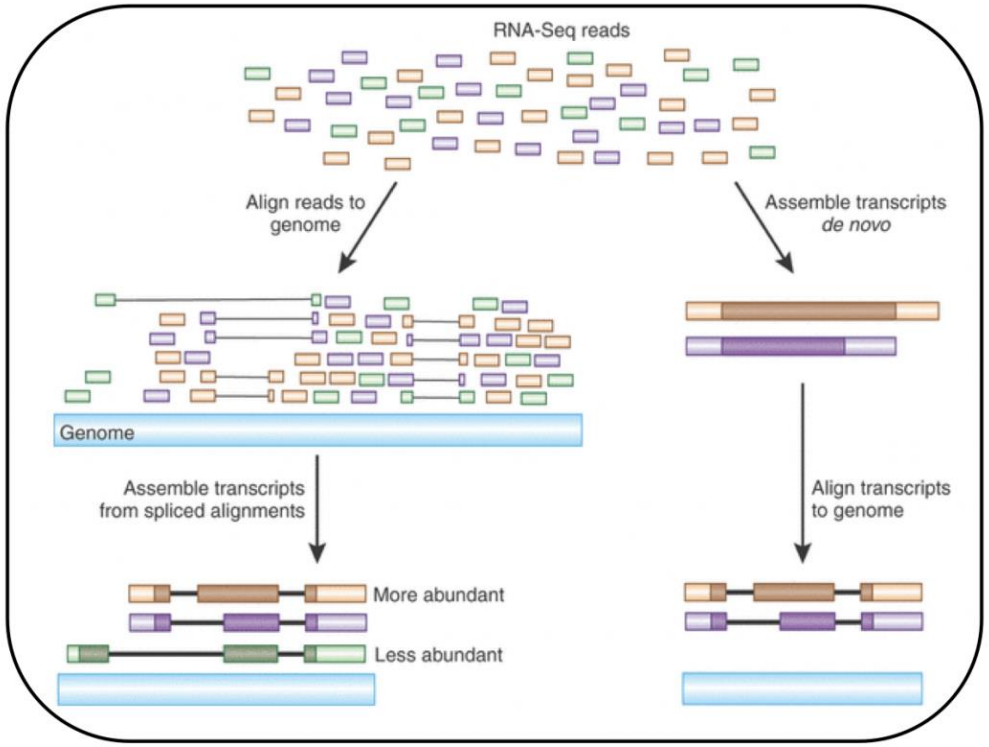
Alignment/mapping

Align to a reference genome

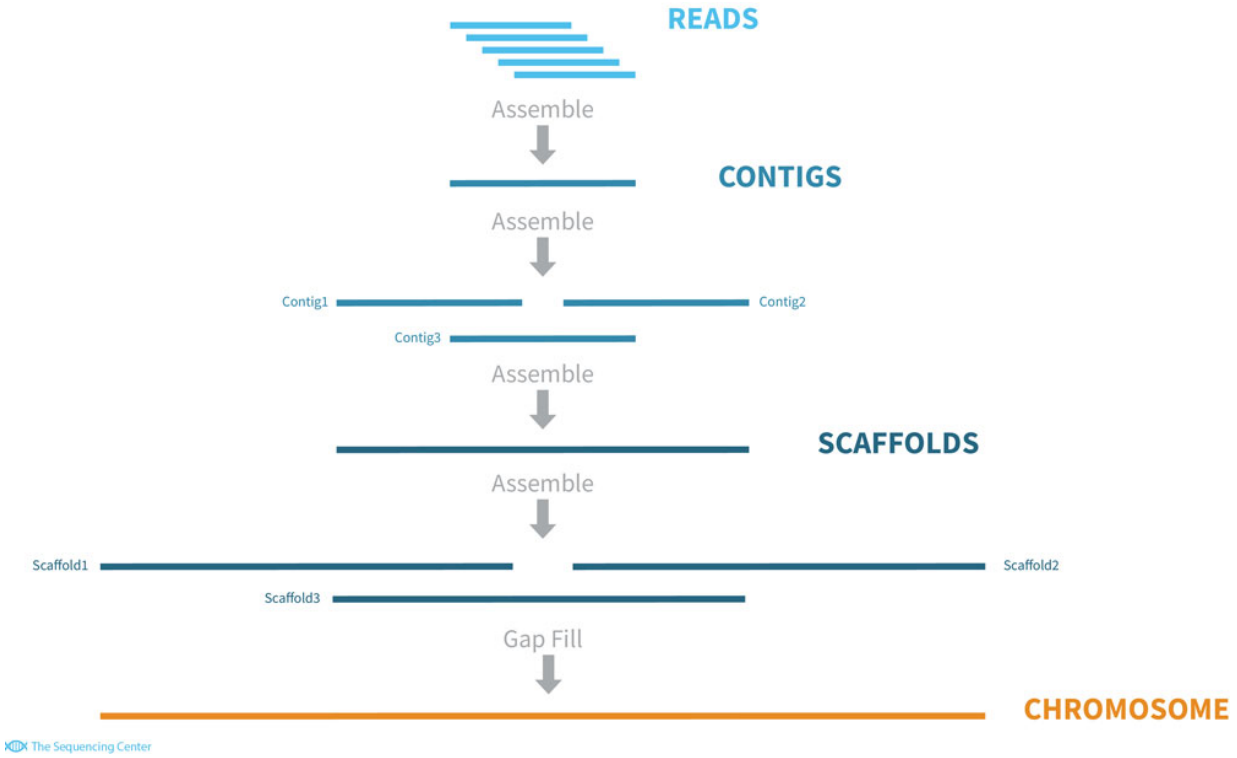


Alignment/mapping

Align to a reference genome

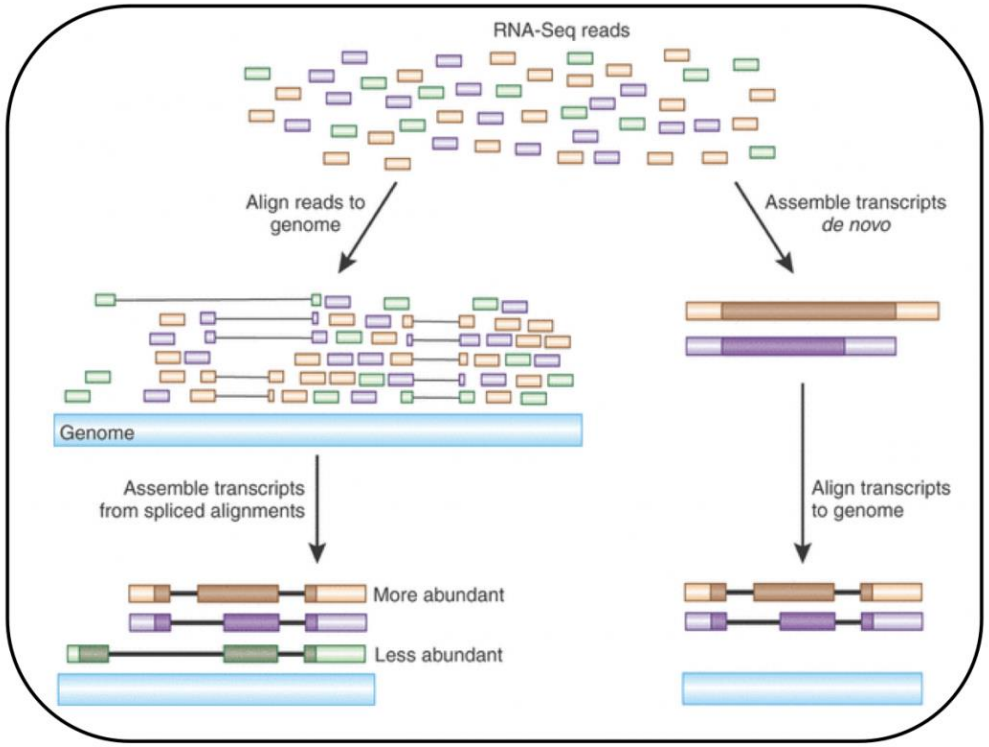


De novo alignment

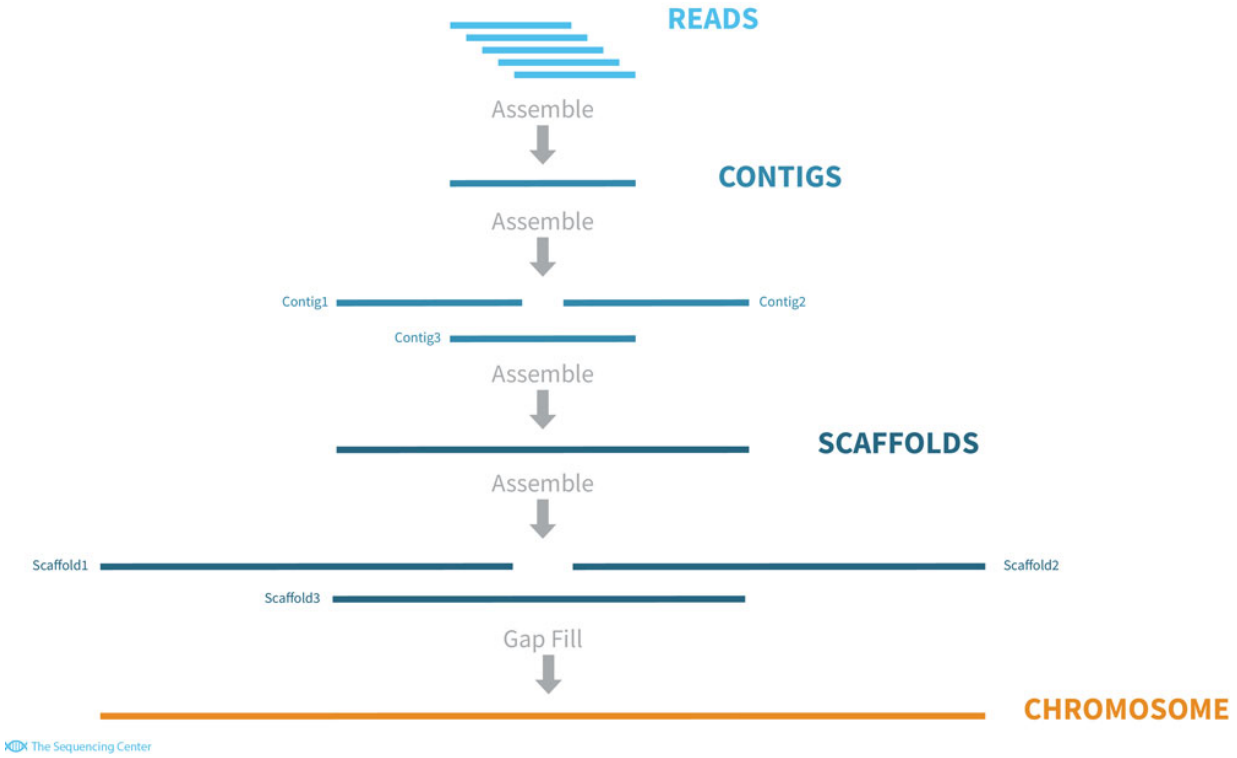


Alignment/mapping

Align to a reference genome



De novo alignment



In both cases, the output is a SAM/BAM file
(**S**equence **A**lignment/**M**ap format)

Alignment/mapping – computational example

Bash script to align some fastq files coming from a QuantSeq experiment

```
####STAR generate indices
STAR --runThreadN 12 --runMode genomeGenerate --genomeDir /home/tigem/s.sarnataro/ngs_backup/genome_reference/mm10/STAR_2.8_indices --genomeFastaFiles /home/tigem/s.sar\
nataro/ngs_backup/genome_reference/mm10/all_chr.fa --sjdbGTFfile /home/tigem/s.sarnataro/ngs_backup/genome_annotation/mm10/gencode.vM25.annotation.gtf

####alignement with STAR
for clean_sample in ${reads_folder}/*_clean.fastq.gz;
do
    STAR --runThreadN 8 \
         --genomeDir ${STAR_indices} \
         --readFilesIn ${clean_sample} \
         --outFilterType BySJout \
         --outFilterMultimapNmax 20 \
         --alignSJoverhangMin 8 \
         --alignSJDBoverhangMin 1 \
         --outFilterMismatchNmax 999 \
         --outFilterMismatchNoverLmax 0.1 \
         --alignIntronMin 20 \
         --alignIntronMax 1000000 \
         --alignMatesGapMax 1000000 \
         --outSAMattributes NH HI NM MD \
         --readFilesCommand gunzip -c \
         --outTmpDir ${tmp_files_folder}\${clean_sample} \
         --outReadsUnmapped Fastx\
         --outSAMtype BAM SortedByCoordinate \
         --outFileNamePrefix ${clean_sample%%\_clean*}_;\

done
```

Alignment/mapping – computational example

Bash script to align some fastq files coming from a QuantSeq experiment

```
####STAR generate indices
STAR --runThreadN 12 --runMode genomeGenerate --genomeDir /home/tigem/s.sarnataro/ngs_backup/genome_reference/mm10/STAR_2.8_indices --genomeFastaFiles /home/tigem/s.sar\
nataro/ngs_backup/genome_reference/mm10/all_chr.fa --sjdbGTFfile /home/tigem/s.sarnataro/ngs_backup/genome_annotation/mm10/gencode.vM25.annotation.gtf

####alignement with STAR
for clean_sample in ${reads_folder}/*_clean.fastq.gz;
do
    STAR --runThreadN 8 \
         --genomeDir ${STAR_indices} \
         --readFilesIn ${clean_sample} \
         --outFilterType BySJout \
         --outFilterMultimapNmax 20 \
         --alignSJoverhangMin 8 \
         --alignSJDBoverhangMin 1 \
         --outFilterMismatchNmax 999 \
         --outFilterMismatchNoverLmax 0.1 \
         --alignIntronMin 20 \
         --alignIntronMax 1000000 \
         --alignMatesGapMax 1000000 \
         --outSAMattributes NH HI NM MD \
         --readFilesCommand gunzip -c \
         --outTmpDir ${tmp_files_folder}\${clean_sample} \
         --outReadsUnmapped Fastx\
         --outSAMtype BAM SortedByCoordinate \
         --outFileNamePrefix ${clean_sample%%\_clean*}_;\

done
```

The tools and the parameters that have been used **strictly depends on the experiment**.

Contamination analysis

Given some typical microorganisms usually responsible for samples contamination, we try to understand if unmapped reads come from some of these microorganisms.

At the end, we obtain contamination reports.

Contamination analysis

Given some typical microorganisms usually responsible for samples contamination, we try to understand if unmapped reads come from some of these microorganisms.

At the end, we obtain contamination reports.

```
Mapping information...

                                 UNIQUE READS:
                Uniquely mapped reads number |       2697632
                     Uniquely mapped reads % |       62.37%
                       Average mapped length |       86.02
                    Number of splices: Total |       446786
         Number of splices: Annotated (sjdb) |       440865
                    Number of splices: GT/AG |       442271
                    Number of splices: GC/AG |       3619
                    Number of splices: AT/AC |       212
           Number of splices: Non-canonical |       684
                  Mismatch rate per base, % |       0.67%
                     Deletion rate per base |       0.01%
                    Deletion average length |       2.21
                    Insertion rate per base |       0.01%
                   Insertion average length |       1.77
                        MULTI-MAPPING READS:
      Number of reads mapped to multiple loci |       549686
           % of reads mapped to multiple loci |       12.71%
      Number of reads mapped to too many loci |       29393
           % of reads mapped to too many loci |       0.68%
                             UNMAPPED READS:
     % of reads unmapped: too many mismatches |       0.00%
               % of reads unmapped: too short |       24.06%
                   % of reads unmapped: other |       0.18%
                             CHIMERIC READS:
                   Number of chimeric reads |       0
                        % of chimeric reads |       0.00%
```

Contamination analysis

Given some typical microorganisms usually responsible for samples contamination, we try to understand if unmapped reads come from some of these microorganisms.

At the end, we obtain contamination reports.

```
Mapping information...

                                UNIQUE READS:
                Uniquely mapped reads number |       2697632
                     Uniquely mapped reads % |       62.37%
                       Average mapped length |       86.02
                    Number of splices: Total |       446786
         Number of splices: Annotated (sjdb) |       440865
                    Number of splices: GT/AG |       442271
                    Number of splices: GC/AG |       3619
                    Number of splices: AT/AC |       212
            Number of splices: Non-canonical |       684
                   Mismatch rate per base, % |       0.67%
                      Deletion rate per base |       0.01%
                     Deletion average length |       2.21
                     Insertion rate per base |       0.01%
                    Insertion average length |       1.77
                          MULTI-MAPPING READS:
       Number of reads mapped to multiple loci |       549686
            % of reads mapped to multiple loci |       12.71%
       Number of reads mapped to too many loci |       29393
            % of reads mapped to too many loci |       0.68%
                              UNMAPPED READS:
      % of reads unmapped: too many mismatches |       0.00%
                % of reads unmapped: too short |       24.06%
                    % of reads unmapped: other |       0.18%
                              CHIMERIC READS:
                    Number of chimeric reads |       0
                         % of chimeric reads |       0.00%
```

```
BACTERIA

MLEN    31
GAPN    5
MISMN   5
TOTAL_ALIGNMENTS        1940607
LOW_QUALITY_ALIGNMENTS  1318471
HI_QUALITY_ALIGNMENTS   622136
AMBIG_ALIGNMENTS        0
VALID_ALIGNMENTS        622136
TOTAL_READS     917156
HI_QUALITY_READS        553051
VALID_READS     553051
TOTAL_BLASTED_ORGANISMS 27
HI_QUALITY_BLASTED_ORGANISMS    7
VALID_BLASTED_ORGANISMS 5
```

```
VALID READS BACTERIA
        S258_20210830112938_10_S125_L002_R
Acholeplasma laidlawii  0       0       0
Mycoplasma arginini     1.75026082352306
Mycoplasma fermentans   4.37782921676307
Mycoplasma hominis      1.16904673574353
Mycoplasma hyorhinis    92.1192107192807
Mycoplasma orale        0.583652504689598
```

Quantifying gene/other features expression

"The genomic coordinates of where the read is mapped (BAM) are cross-referenced with the genomic coordinates of whichever feature you are interested in counting expression of (GTF), it can be exons, genes or transcripts"

GTF

```
chr1 unknown  exon      113217048  113217252   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  exon      113217048  113217351   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown  exon      113217470  113217671   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  CDS       113217535  113217671   .   +   0   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  start_codon  113217535  113217537   .   +   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

feature type

feature

Quantifying gene/other features expression

Quantification output: A count matrix, where rows are genes (or transcripts or other features) and columns are samples

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

Quantifying gene/other features expression

Quantification output: A count matrix, where rows are genes (or transcripts or other features) and columns are samples

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

How to report the results? CPM, FPKM, RPKM, etc…