

Tuesday – part 1

# RNA-Seq data analysis

## Quality control

**Michał Szcześniak, PhD**

**Faculty of Biology , Adam Mickiewicz University, Poznań**  
**ideas4biology Ltd.**

# Need help on RNA-Seq?

<http://www.rna-seqblog.com/>

DON'T MISSFeatured RNA-Seq Job – Senior Bioinformatician

**RNA-Seq Blog**  
Transcriptome Sequencing Research & Industry News

Introducing  
New SIRV Sets

LEARN MORE



HOME NEWS ▾ EVENTS ▾ JOBS ▾ TECHNOLOGY ▾ DATA ANALYSIS ▾ BLOG READER POSTS CONTACT ▾

**RNA-SEQ NEWS**

**BRIGHAM  
AND  
WOMEN'S  
HOSPITAL**

### Featured RNA-Seq Job – Senior Bioinformatician

🕒 22 hours ago 🗨️ Leave a comment 👁️ 177 Views

**GENERAL SUMMARY/ OVERVIEW STATEMENT:** This position will work with the NIH funded Accelerating Medicines Partnership, which brings high-level government, industry and non-profit foundation

partners together to identify and validate the ...

[Read More »](#)



```
graph TD; A[Raw sequence reads] --> B[Quality assessment (trimming, filtering, etc.)]; B --> C[Mapping to the reference genome]; C --> D[Feature counting];
```

### A comprehensive simulation study on classification of RNA-Seq data

🕒 22 hours ago 🗨️ Leave a comment 👁️ 562 Views

RNA sequencing (RNA-Seq) is a powerful technique for the gene-expression profiling of organisms that uses the capabilities of next-generation sequencing technologies. Developing gene-expression-based

classification algorithms is an emerging powerful method ...

[Read More »](#)

**RNA-protein interactions and RNA Structure**

### Summer School on RNA-protein interactions RNA Structure and Biology workshop

🕒 4 days ago 🗨️ Leave a comment 👁️ 724 Views

October 2–6 2017 CEITEC MU, University Campus

**STAY CONNECTED**



**SUBMIT NEWS**

**RECENT POSTS BY OUR READERS**

New Paper Demonstrates Novel Method for Detecting miRNA to Improve Liver Toxicity Testing

Driving the Future of Clinical Research: Bioinformaticians

Can bioinformatics be more accessible?

Evaluation and comparison of computational tools for RNA-seq isoform quantification

**SUBSCRIBE TO THE RNA-SEQ BLOG**

**Subscribe**

**RNA-SEQ PRODUCTS & SERVICES**



Increase mappable reads with RNA input as low as 1ng

[Learn more about CleanTag™ Library Prep](#)

1996-2016 YEARS

**TriLink**  
BIOTECHNOLOGIES

**Do your own RNA-seq analysis!**

Different from traditional, expert-only Bioinformatics software



# Need help on RNA-Seq?

<https://rnaseq.uoregon.edu/>

## RNA-seqlopedia



RNA-seq produces millions of sequences from complex RNA samples. With this powerful approach, you can:

1. Measure gene expression.
2. Discover and annotate complete transcripts.
3. Characterize alternative splicing and polyadenylation.

The RNA-seqlopedia provides an overview of RNA-seq and of the choices necessary to carry out a successful RNA-seq experiment.

RNA-seqlopedia is written by the [Cresko Lab](#) of the [University of Oregon](#) and was funded by grant R24 RR032670 (NIH, National Center for Research Resources).

[Credits.](#)

# Need help on RNA-Seq?

<https://www.ncbi.nlm.nih.gov/pubmed/26813401>

Genome Biol. 2016 Jan 26;17:13. doi: 10.1186/s13059-016-0881-8.

## A survey of best practices for RNA-seq data analysis.

[Conesa A](#)<sup>1,2</sup>, [Madrigal P](#)<sup>3,4</sup>, [Tarazona S](#)<sup>5,6</sup>, [Gomez-Cabrero D](#)<sup>7,8,9,10</sup>, [Cervera A](#)<sup>11</sup>, [McPherson A](#)<sup>12</sup>, [Szczęsniak MW](#)<sup>13</sup>, [Gaffney DJ](#)<sup>14</sup>, [Elo LL](#)<sup>15</sup>, [Zhang X](#)<sup>16,17</sup>, [Mortazavi A](#)<sup>18,19</sup>.

### ⊕ Author information

### Erratum in

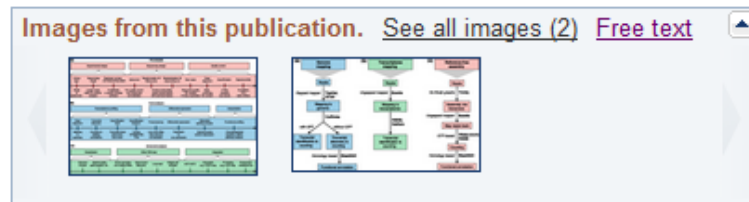
Erratum to: A survey of best practices for RNA-seq data analysis. [Genome Biol. 2016]

### Abstract

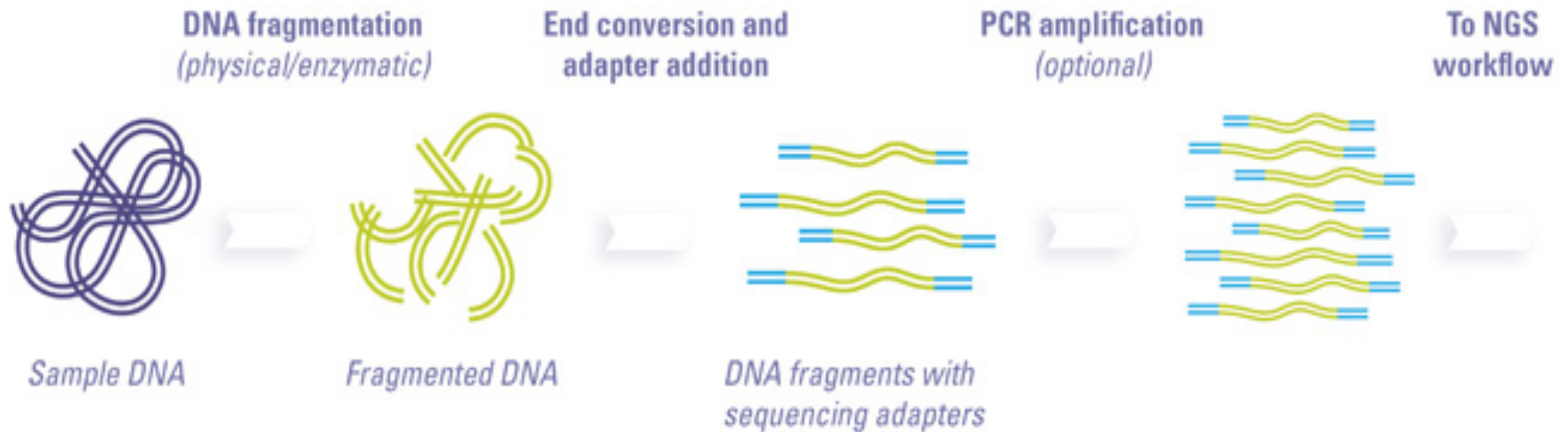
RNA-sequencing (RNA-seq) has a wide variety of applications, but no single analysis pipeline can be used in all cases. We review all of the major steps in RNA-seq data analysis, including experimental design, quality control, read alignment, quantification of gene and transcript levels, visualization, differential gene expression, alternative splicing, functional analysis, gene fusion detection and eQTL mapping. We highlight the challenges associated with each step. We discuss the analysis of small RNAs and the integration of RNA-seq with other functional genomics techniques. Finally, we discuss the outlook for novel technologies that are changing the state of the art in transcriptomics.

PMID: 26813401 PMCID: [PMC4728800](#) DOI: [10.1186/s13059-016-0881-8](#)

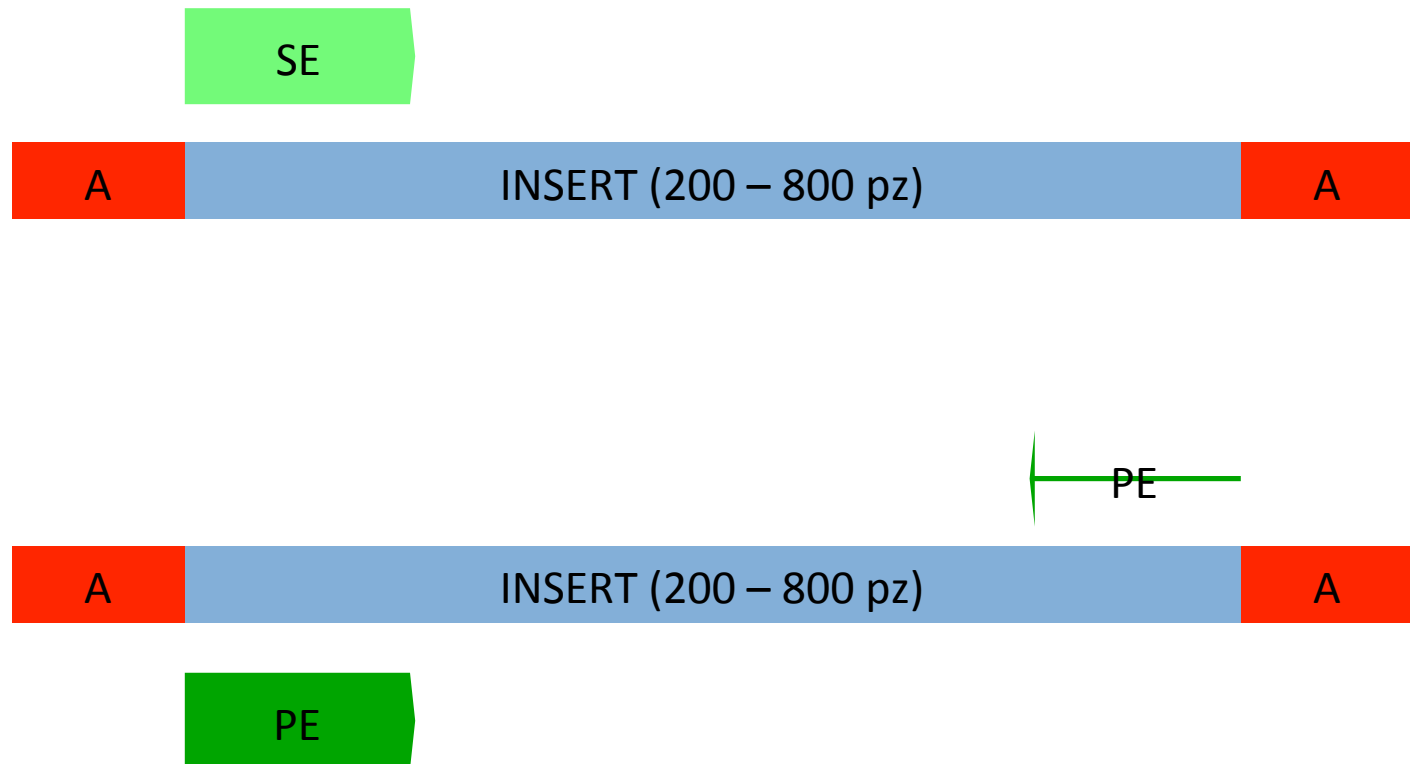
[Indexed for MEDLINE] [Free PMC Article](#)



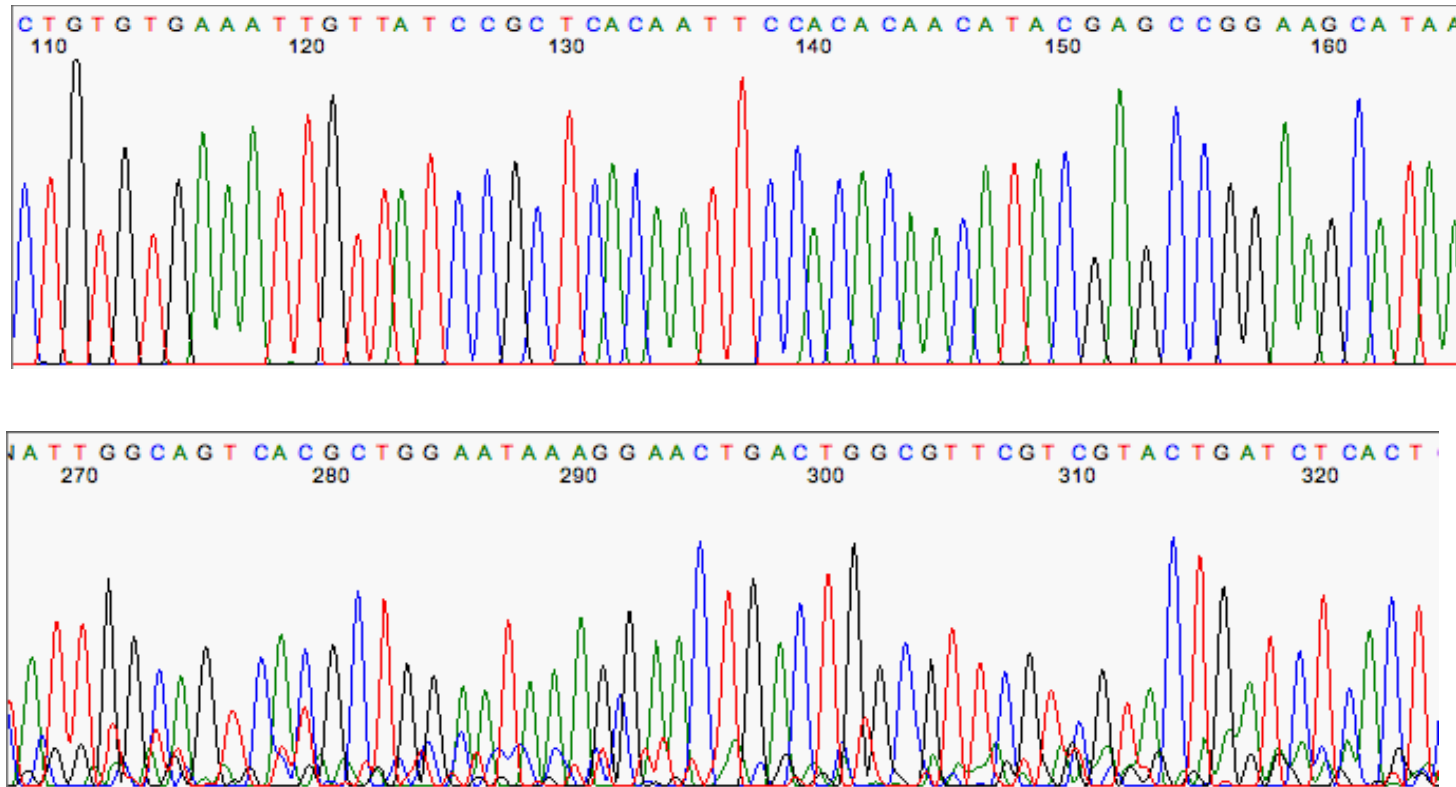
# Preparing the sequencing libraries



# Single-end vs paired-end



# Sanger sequencing



# FASTQ format

## Example record 1:

```
@SRR1586016.34 HW-ST997:228:COTA9ACXX:4:1101:1368:2342/1
CAATTTTCTGACTATAGCCTGAGGGTGGAATTCTCGGGTGCCAAGGAAAT
+
?@@DFFFDFDFF<<C<?<E2AFG>FFH<CG=@BG<F?FFH?F69FGGH@###
```

## Example record 2:

```
@SRR1747399.209 HWI-ST845:121101:C18JNACXX:5:1101:12845:2209/1
GAGCACTGCAGGCCAAAGTCAGAGTCTCCGTGGTCGGGCTGGTCTAGG
+
@@@FFFFFHGD<CFGJJGFHIIGCFECGHIJJGGGGEIGGIIGGIIEH
```



# Quality

```
>gnl|SRA|SRR191637.1221 HWUSI-EAS1600:IB_HJ1:8_5_2010:0:5:21:1153:5906  
CGTACNNNAATAGTTTAACTGTTGGTTCGTATG
```

One channel quality score

```
30 30 30 27 30  2  2  2  2 27 28 28 24 27 28 32 32 32 32 32  
32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32
```

# Phred quality score

$$Q = -10 \log_{10}(P)$$

**Phred quality scores are logarithmically linked to error probabilities**

<b>Phred Quality Score</b>	<b>Probability of incorrect base call</b>	<b>Base call accuracy</b>
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

# ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	\$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

! " # \$ % & ' ( ) \* + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J K L M N O  
P Q R S T U V W X Y Z [ \ ] ^ \_ ` a b c d e f g h i j k l m n o p q r s t u v w x y z { | } ~

# RNA-Seq Pipelines

1. FASTQ → **QC and filtering** → mapping → *ab initio* assembly
2. FASTQ → **QC and filtering** → expression estimation → differential expression analysis
3. FASTQ → **QC and filtering** → *de novo* assembly

# Software

**FastQC** – quality control

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

**FASTX-Toolkit** – reads filterings and processing (single-end only)

[http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)

**Trimmomatic** – reads processing and filtering

<http://www.usadellab.org/cms/?page=trimmomatic>

**BBDUK2**– reads processing and filtering

<http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/>

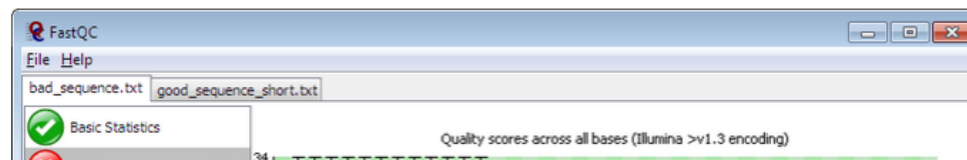
# FastQC



## FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A <a href="#">suitable Java Runtime Environment</a> The <a href="#">Picard</a> BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under <a href="#">GPL v3 or later</a> .
Initial Contact	<a href="#">Simon Andrews</a>

[Download Now](#)



# FastQC

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

## Documentation

A [copy of the FastQC](#) documentation is available for you to try before you buy (well download..).

## Example Reports

- [Good Illumina Data](#)
- [Bad Illumina Data](#)
- [Adapter dimer contaminated run](#)
- [Small RNA with read-through adapter](#)
- [Reduced Representation BS-Seq](#)
- [PacBio](#)
- [454](#)

## Changelog

- 25-3-15: Version 0.11.3 released
  - Fixed a bug when disabling the per-tile plot from limits.txt
  - Fixed a bug which caused the program to continue when processing of multiple files was actually complete
  - Fixed a bug which meant format selection in the interactive application didn't work

**ERX1071671:** Illumina HiSeq 2500 paired end sequencing; RNA-Seq of human embryonic stem cells over expressing NANOS3 and DAZL  
1 ILLUMINA (Illumina HiSeq 2500) run: 16.5M spots, 3.3G bases, 1.5Gb downloads

**Design:** RNA-Seq of human embryonic stem cells over expressing NANOS3 and DAZL

**Submitted by:** Clintec, Karolinska Institutet (CLINTEC, KAROLINSKA INSTITUTET)

**Study:** RNA-Seq of human embryonic stem cells over expressing NANOS3 and DAZL

[PRJEB10573](#) • [ERP011837](#) • [All experiments](#) • [All runs](#)  
[show Abstract](#)

**Sample:** P1177\_112

[SAMEA3515983](#) • [ERS823132](#) • [All experiments](#) • [All runs](#)  
**Organism:** [Homo sapiens](#)

**Library:**

**Name:** P1177\_112

**Instrument:** Illumina HiSeq 2500

**Strategy:** RNA-Seq

**Source:** TRANSCRIPTOMIC

**Selection:** cDNA

**Layout:** PAIRED

**Construction protocol:** The hESC line HS401 (46,XY) was cultured on hESC-qualified Matrigel (Corning) -coated plates using mTeSR1 medium (StemCell Technologies) in 5% CO2 and atmospheric oxygen at 37°C. The cells were passaged using Accutase (Life Technologies), followed by an over night incubation with 5 µM Y-27632 (Millipore). hESCs were co-transfected with 2.5 µg piggyBac transposon (IFP2) and 2.5 µg transposase vector (MOCK, NANOS3, or DAZL) in a 6-well format using 5 µl PLUS reagent and 10 µl Lipofectamine LTX (Life Technologies) according to the manufacturer's instructions. Two days after the transfection, cells were selected with 1 µg/ml puromycin (Life Technologies) for 6 days. Transfection was repeated three times for each transposase vector and cells were collected for RNA isolation at passage 1 after transfection. RNA libraries for sequencing were prepared using TruSeq Stranded mRNA Sample prep kit with 96 dual indexes (Illumina, CA, USA) according to the manufacturers instructions with the following changes. The protocols were automated using an Agilent NGS workstation (Agilent, CA, USA) using purification steps as described in: S. Lundin/ H. Stranneheim PLoS ONE 2010) and (E. Borgstrom/ S.Lundin et al. 2011 PLoS ONE).

**Spot descriptor:**



**Experiment attributes:**

*Experimental Factor:* over expression DAZL: genotype

**Runs:** 1 run, 16.5M spots, 3.3G bases, [1.5Gb](#)

Run	# of Spots	# of Bases	Size	Published
<a href="#">ERR990413</a>	16,529,103	3.3G	1.5Gb	2016-10-20

Our data

ERR990413



# Our analysis in FastQC

## # downloading RNA-Seq data

wget [ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR990/ERR990413/ERR990413\\_1.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR990/ERR990413/ERR990413_1.fastq.gz)

wget [ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR990/ERR990413/ERR990413\\_2.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR990/ERR990413/ERR990413_2.fastq.gz)

**Study:** RNA-Seq of human embryonic stem cells over expressing NANOS3 and DAZL

**Instrument:** Illumina HiSeq 2500

RNA libraries for sequencing were prepared using TruSeq Stranded mRNA Sample prep kit

## # decompressing and keeping only the 2,500,000 first records

gunzip -c ERR990413\_1.fastq.gz | head -10000000 > ERR990413\_1.fastq

gunzip -c ERR990413\_2.fastq.gz | head -10000000 > ERR990413\_2.fastq

# Our analysis in FastQC

## # quality control: before filtering

```
mkdir FASTQC_out
```

```
mkdir FASTQC_out/ERR990413_raw
```

```
fastqc ERR990413_1.fastq ERR990413_2.fastq --quiet --noextract --nogroup  
--outdir FASTQC_out/ERR990413_raw
```

- |                          |   |
|--------------------------|---|
| <code>-q --quiet</code>  | Supress all progress messages on stdout and only report errors.   |
| <code>--noextract</code> | Do not uncompress the output file after creating it. You should set this option if you do not wish to uncompress the output when running in non-interactive mode.   |
| <code>--nogroup</code>   | Disable grouping of bases for reads >50bp. All reports will show data for every base in the read. WARNING: Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! |

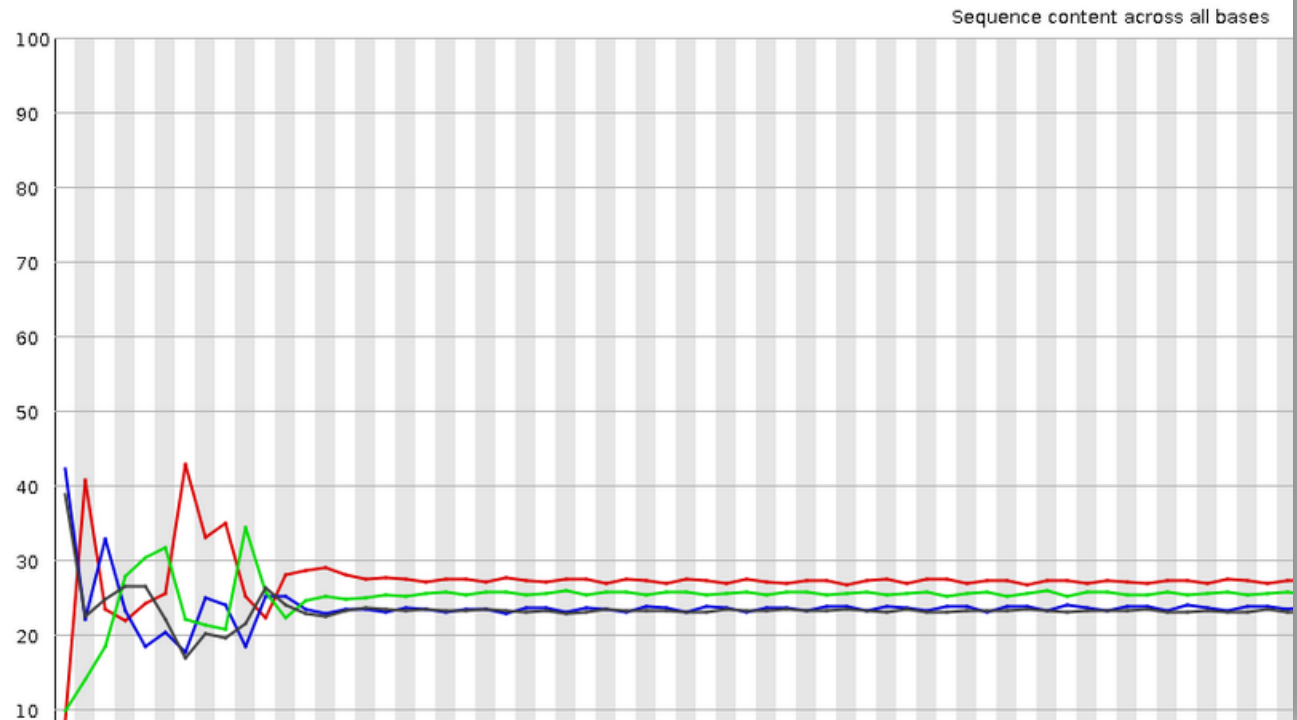
# Our analysis in FastQC

## 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](#)

### ✗ Per base sequence content



# Trimmomatic

USADELLAB.org

Home

Research

Education

Service & Software

Publications

Supporting Info

About Us

NGS, DE and other things

## Trimmomatic: A flexible read trimming tool for Illumina NGS data

### Citations

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

### Downloading Trimmomatic

Version 0.36: [binary](#), [source](#) and [manual](#)

### Quick start

#### Paired End:

```
java -jar trimmomatic-0.35.jar PE -phred33 input_forward.fq.gz input_reverse.fq.gz  
output_forward paired.fq.gz output_forward unpaired.fq.gz output_reverse paired.fq.gz
```

# Trimmomatic - functionalities

**Adapter clipping** – removing adapter sequences from the ends of the reads (standard Illumina adapter sequences are distributed with Trimmomatic)

**Quality trimming** – scanning read ends with a sliding window and cutting off regions with poor average quality

**Quality filtering** – keeping or discarding reads based on their length after trimming. Works with both single and paired-end data.

# Adapter sequences

**TruSeq3-SE, TruSeq3-PE** – adapter sequences used by Illumina HiSeq and MiSeq machines (single–end and paired–end, respectively)

**TruSeq2-SE, TruSeq2-PE** – adapter sequences used by Illumina GAII machines

**NexteraPE-PE** – adapter sequences used by Nextera sample preparation protocol

# QC for our data: paired-end reads

```
mkdir TRIMMED
```

```
mkdir STATUS
```

```
java -Xms4g -Xmx4g -jar trimmomatic/trimmomatic.jar PE -threads 1 -  
phred33 ERR990413_1.fastq ERR990413_2.fastq TRIMMED/  
ERR990413_trimmomatic_R1_trimmed.fastq /dev/null TRIMMED/  
ERR990413_trimmomatic_R2_trimmed.fastq /dev/null  
ILLUMINACLIP:trimmomatic/adapters/TruSeq3-PE.fa:2:30:10 LEADING:20  
TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50 &> STATUS/trimmomatic.log
```

## Parameters:

Input file(s)

Output file(s)

QC settings

# QC for our data: paired-end reads

```
java -Xms4g -Xmx4g -jar trimmomatic/trimmomatic.jar PE -threads 1 -  
phred33 ERR990413_1.fastq ERR990413_2.fastq TRIMMED/  
ERR990413_trimmomatic_R1_trimmed.fastq /dev/null TRIMMED/  
ERR990413_trimmomatic_R2_trimmed.fastq /dev/null  
ILLUMINACLIP:trimmomatic/adapters/TruSeq3-PE.fa:2:30:10 LEADING:20  
TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50 &> STATUS/trimmomatic.log
```

PE: paired-end

Removing adapters:TruSeq3-PE.

Two input files:

- ERR990413\_1.fastq
- ERR990413\_2.fastq



# QC for our data: paired-end reads

```
java -Xms4g -Xmx4g -jar trimmomatic/trimmomatic.jar PE -threads 1 -  
phred33 ERR990413_1.fastq ERR990413_2.fastq TRIMMED/  
ERR990413_trimmomatic_R1_trimmed.fastq /dev/null TRIMMED/  
ERR990413_trimmomatic_R2_trimmed.fastq /dev/null  
ILLUMINACLIP:trimmomatic/adapters/TruSeq3-PE.fa:2:30:10 LEADING:20  
TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50 &> STATUS/trimmomatic.log
```

Four output files (data redirected to /dev/null are discarded):

TRIMMED/ERR990413\_trimmomatic\_R1\_trimmed.fastq (paired reads, part 1)

/dev/null (unpaired reads, part 1)

TRIMMED/ERR990413\_trimmomatic\_R2\_trimmed.fastq (paired reads, part 2)

/dev/null (unpaired reads, part 2)

# Trimmomatic PE - report

```
TrimmomaticPE: Started with arguments: -threads 2 -phred33
ERR990413_1.fastq ERR990413_2.fastq TRIMMED/
ERR990413_1_trimmed.fastq /dev/null TRIMMED/
ERR990413_2_trimmed.fastq /dev/null ILLUMINACLIP:
Trimmomatic/adapters/TruSeq3-PE.fa:2:30:10 LEADING:20
TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50Using PrefixPair:
'TACACTCTTTCCCTACACGACGCTCTTCCGATCT' and
'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'ILLUMINACLIP: Using 1
prefix pairs, 0 forward/reverse sequences, 0 forward only
sequences, 0 reverse only sequencesInput
Read Pairs: 2500000 Both Surviving: 2183039 (87.32%)
Forward Only Surviving: 236120 (9.44%) Reverse Only
Surviving: 36198 (1.45%) Dropped: 44643 (1.79%)
TrimmomaticPE: Completed successfully
```

# QC for single-end reads

```
java -Xms4g -Xmx4g -jar trimmomatic/trimmomatic.jar SE -threads 1 -phred33  
reads.fastq reads.trimmed.fastq ILLUMINACLIP:trimmomatic/adapters/  
TruSeq3-SE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20  
MINLEN:50
```

## Parameters:

Input file(s)

Output file(s)

QC settings

SE: single-end

There is one input file (**reads.fastq**) and one output file (**reads.trimmed.fastq**).

# Trimmomatic: summary

+

All-In-One tool

Supports both single-end and paired-end data

Advanced quality trimming

Built-in adapter sequences

-

Complicated usage

# Quality filtering : BBduk2

## (An alternative for Trimmomatic)

```
bbmap/bbduk2.sh -Xmx2g threads=2 in=ERR990413_1.fastq in2=ERR990413_2.fastq  
out=TRIMMED/ERR990413_bbduk2_R1.fastq out2=TRIMMED/  
ERR990413_bbduk2_R2.fastq qtrim=w trimq=20 maq=10 rref=bbmap/resources/  
adapters.fa k=23 mink=11 hdist=1 tbo tpe minlength=50 removeifeitherbad=t  
overwrite=t stats=STATUS/ERR990413.bbduk2_stats.txt 2> STATUS/  
ERR990413.bbduk2_trimming.txt
```

**in=** Main input. in=stdin.fq will pipe from stdin.

**in2=** Input for 2nd read of pairs in a different file.

**out=** (outnonmatch) Write reads here that do not contain kmers matching the database.

**out2=** (outnonmatch2) Use this to write 2nd read of pairs to a different file.

**qtrim=f** Trim read ends to remove bases with quality below trimq. Performed AFTER looking for kmers. Wybrana opcja: w (sliding window)

**trimq=6** Regions with average quality BELOW this will be trimmed.

**rref=** Comma-delimited list of fasta reference files for right-trimming.

# Quality filtering : BBduk2

## (An alternative for Trimmomatic)

```
bbmap/bbduk2.sh -Xmx2g threads=2 in=ERR990413_1.fastq in2=ERR990413_2.fastq  
out=TRIMMED/ERR990413_bbduk2_R1.fastq out2=TRIMMED/  
ERR990413_bbduk2_R2.fastq qtrim=w trimq=20 maq=10 rref=bbmap/resources/  
adapters.fa k=23 mink=11 hdist=1 tbo tpe minlength=50 removeifeitherbad=t  
overwrite=t stats=STATUS/ERR990413.bbduk2_stats.txt 2> STATUS/  
ERR990413.bbduk2_trimming.txt
```

- k=27** Kmer length used for finding contaminants. Contaminants shorter than k will not be found. k must be at least 1.
- mink=0** Look for shorter kmers at read tips down to this length, when k-trimming or masking.
- hdist** Maximum Hamming distance for ref kmers (subs only).
- tbo=f** (trimbyoverlap) Trim adapters based on where paired reads overlap.  
**tpe=f** (trimpairsevenly) When kmer right-trimming, trim both reads to the minimum length of either.

# Quality filtering : BBduk2

## (An alternative for Trimmomatic)

```
bbmap/bbduk2.sh -Xmx2g threads=2 in=ERR990413_1.fastq in2=ERR990413_2.fastq  
out=TRIMMED/ERR990413_bbduk2_R1.fastq out2=TRIMMED/  
ERR990413_bbduk2_R2.fastq qtrim=w trimq=20 maq=10 rref=bbmap/resources/  
adapters.fa k=23 mink=11 hdist=1 tbo tpe minlength=50 removeifeitherbad=t  
overwrite=t stats=STATUS/ERR990413.bbduk2_stats.txt 2> STATUS/  
ERR990413.bbduk2_trimming.txt
```

**minlength=10** (ml) Reads shorter than this after trimming will be discarded. Pairs will be discarded if both are shorter.

**removeifeitherbad=t** (rieb) Paired reads get sent to 'outmatch' if either is match (or either is trimmed shorter than minlen).

**overwrite=t** (ow) Grant permission to overwrite files.

**stats=** Write statistics about which contaminants were detected.

**threads=auto** (t) Set number of threads to use; default is number of logical processors.

# Quality filtering: BBduk2

## (An alternative for Trimmomatic)

```
bbmap/bbduk2.sh -Xmx2g threads=2 in=ERR990413_1.fastq in2=ERR990413_2.fastq  
out=TRIMMED/ERR990413_bbduk2_R1.fastq out2=TRIMMED/  
ERR990413_bbduk2_R2.fastq qtrim=w trimq=20 maq=10 rref=bbmap/resources/  
adapters.fa k=23 mink=11 hdist=1 tbo tpe minlength=50 removeifeitherbad=t  
overwrite=t stats=STATUS/ERR990413.bbduk2_stats.txt 2> STATUS/  
ERR990413.bbduk2_trimming.txt
```

**-Xmx** This will be passed to Java to set memory usage, overriding the program's automatic memory detection. -Xmx20g will specify 20 gigs of RAM, and -Xmx200m will specify 200 megs. The max is typically 85% of physical memory.

**minavgquality=0** (maq) Reads with average quality (after trimming) below this will be discarded.



# BBDUK2: Summary

+

All-In-One tool

Supports both single-end and paired-end data

Advanced quality trimming options

Built-in adapter sequences

-

Complicated usage

# Quality control: after filtering

## **Trimmomatic**

```
mkdir FASTQC_out/ERR990413_filtered_trimmomatic
```

```
fastqc TRIMMED/ERR990413_trimmomatic_R1_trimmed.fastq TRIMMED/  
ERR990413_trimmomatic_R2_trimmed.fastq --quiet --noextract --nogroup --outdir  
FASTQC_out/ERR990413_filtered_trimmomatic
```

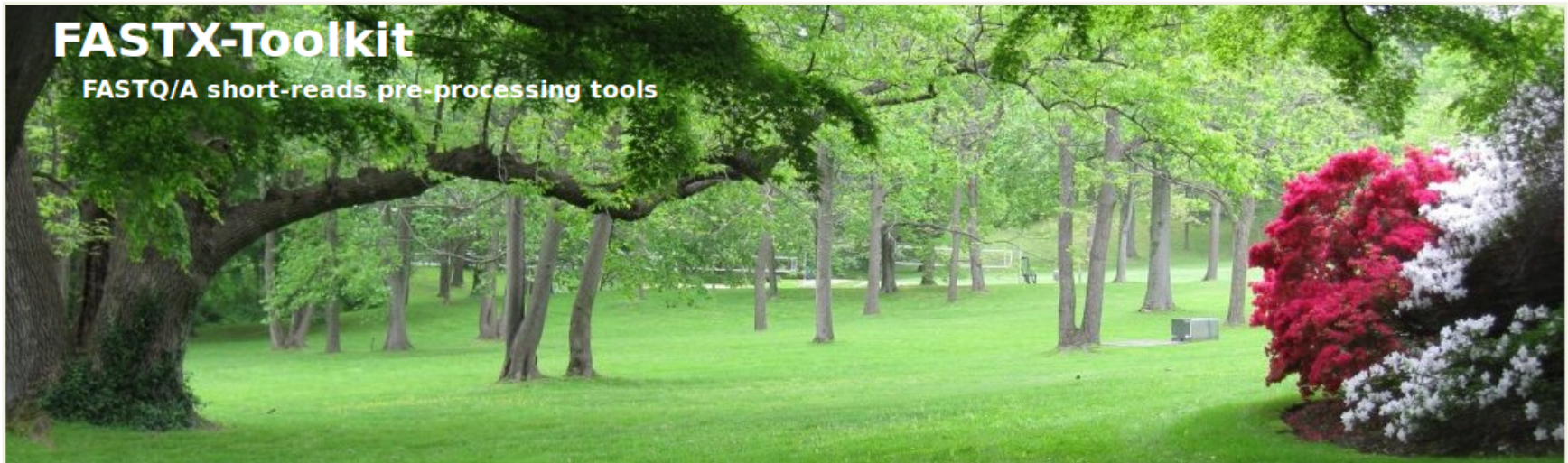
## **bbduk2**

```
mkdir FASTQC_out/ERR990413_filtered_bbduk2
```

```
fastqc TRIMMED/ERR990413_bbduk2_R1.fastq TRIMMED/  
ERR990413_bbduk2_R2.fastq --quiet --noextract --nogroup --outdir FASTQC_out/  
ERR990413_filtered_bbduk2
```

# Filtering and processing of smallRNA-Seq data

# FASTX-Toolkit



[Home](#) | [Download & Installation](#) | [Galaxy Usage](#) | [Command-line Usage](#) | [License](#) | [Useful Links](#) | [Contact](#)

## Introduction

The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

Next-Generation sequencing machines usually produce FASTA or FASTQ files, containing multiple short-reads sequences (possibly with quality information).

The main processing of such FASTA/FASTQ files is mapping (aka aligning) the sequences to reference genomes or other databases using specialized programs. Example of such mapping programs are: [Blat](#), [SHRiMP](#), [LastZ](#), [MAQ](#) and many many others.

[http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)

# Quality control: FASTX-Toolkit

**Adapter clipping** – removing adapter sequences from the ends of the reads

**Quality trimming** – removing low-quality regions from the ends of the reads

**Quality filtering** – keeping or discarding reads based on quality criteria

**Warning:** paired-end reads must be analyzed together in order to preserve the order of reads in both FASTQ files.

# Our data

## # Data download

wget <ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR158/006/SRR1586016/SRR1586016.fastq.gz>

**Study:** microRNA sequencing of HEK293 and H9c2 cells

**Sample:** HEK293\_microRNAs

**Instrument:** Illumina HiSeq 2000

Small RNA sequencing libraries were prepared using Illumina TruSeq Small RNA Sequencing kits

gunzip SRR1586016.fastq.gz

## # quality control: before filtering

mkdir FASTQC\_out/SRR1586016\_raw

fastqc **SRR1586016.fastq** --quiet --noextract --nogroup --outdir **FASTQC\_out/SRR1586016\_raw**

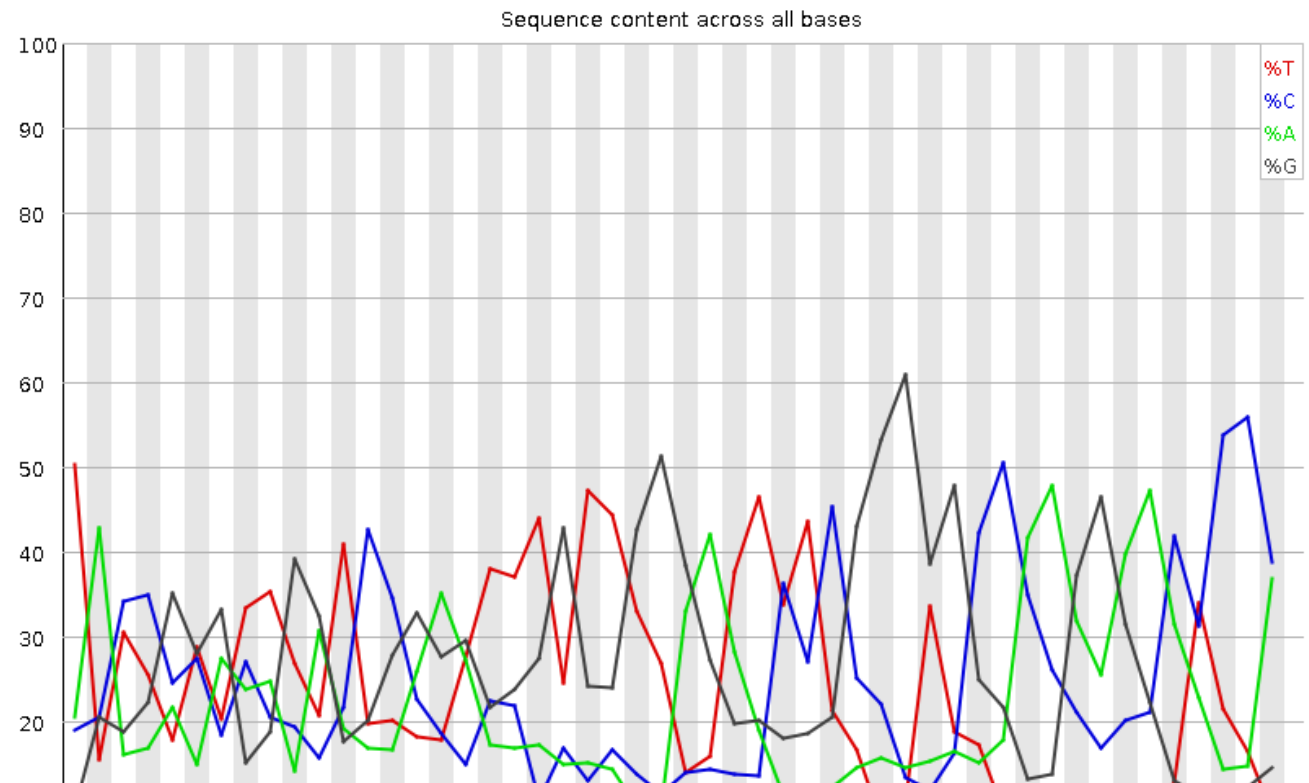
# FastQC: report

## 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](#)

### ✗ Per base sequence content



# Adapter clipping

```
fastx_clipper -Q33 -v -l 17 -a TGGGAATTCTCGGGTGCCAAGG -c  
-i SRR1586016.fastq -o SRR1586016_clipped.fastq
```

Input and output file are specified using -i and -o flags

Print clipping report (-v)

Only reads that have at least 17 bp after clipping (-l 17) are kept

Adapter sequence is specified using -a flag

Discarded are sequences without adapter(-c)



# Adapter clipping - report

Clipping Adapter: TGG AATTCTCGGGTGCCAAGG

Min. Length: 17

Non-Clipped reads - discarded.

Input: 3713258 reads.

Output: 2812780 reads.

discarded 825861 too-short reads.

discarded 47338 adapter-only reads.

discarded 19604 non-clipped reads.

discarded 7675 N reads.

# Quality filtering

```
fastq_quality_filter -Q33 -v -q 20 -p 95 -i SRR1586016_clipped.fastq  
-o SRR1586016_QC.fastq
```

**Input** and **output** file are specified using -i and -o flags

Print filtering report (-v)

Reads are only kept if at least 95% of their bases have the quality score of min. 20 (-q 20 -p 95)

# Quality filtering – report

Quality cut-off: 20

Minimum percentage: 95

Input: 2812780 reads.

Output: 2667324 reads.

discarded 145456 (5%) low-quality reads.

# QC pipeline

```
fastx_clipper -Q33 -v -l 17 -a TGGGAATTCTCGGGTGCCAAGG -c  
-i SRR1586016.fastq | fastq_quality_filter -Q33 -v -q 20 -p 95 -o  
SRR1586016_QC.fastq
```

**Input** and **output** file are specified using -i and -o flags

Different FASTX–Toolkit commands are chained using Unix pipes – the output of one command becomes the input data for the next one

# QC pipeline - report

Clipping Adapter: TGGAATTCTCGGGTGCCAAGG

Min. Length: 17

Non-Clipped reads - discarded.

Input: 3713258 reads.

Output: 2812780 reads.

**Adapter clipping**

discarded 825861 too-short reads.

discarded 47338 adapter-only reads.

discarded 19604 non-clipped reads.

discarded 7675 N reads.

Quality cut-off: 20

Minimum percentage: 95

Input: 2812780 reads.

Output: 2667324 reads.

**Quality filtering**

discarded 145456 (5%) low-quality reads.

# Additional steps

**FASTA conversion** – converting data from FASTQ to FASTA format

**Sequence collapsing** – removing redundant sequences from a FASTQ or FASTA file

# FASTA conversion

```
fastq_to_fasta -i SRR1586016.fastq -o SRR1586016.fasta -v
```

**Input** and **output** file are specified using -i and -o flags

Print conversion report (-v)

Reads that contain **N** characters are discarded

# Sequence collapsing

```
fastx_collapser -i SRR1586016.fastq -o  
SRR1586016_collapsed.fasta -v
```

**Input** and **output** file are specified using -i and -o flags

Print sequence collapsing report (-v)

Redundant sequences are discarded but the information about their abundance is preserved

Operation can be reversed by the 'fastx\_uncollapser' command



# fastx\_collapser: results

Input: 2675102 sequences (representing 2675102 reads)

Output: 219894 sequences (representing 2675102 reads)

>1-199647

TATTGCACTTGTCCCGGCCTGT

# [hsa-miR-92a-3p](#)

>2-161038

TACCCTGTAGATCCGAATTTGT

# [hsa-miR-10a-5p](#)

>3-121255

TACCCTGTAGAACCGAATTTGT

# [hsa-miR-10b-5p](#)

>4-88112

TACCCTGTAGATCCGAATTTGTG

# [hsa-miR-10a-5p](#)

>5-58922

ACCCTGTAGATCCGAATTTGTG

# [hsa-miR-10a-5p](#)

>6-45673

TATTGCACTTGTCCCGGCCTGTAA

# [hsa-miR-92a-3p](#)

>7-44625

AGCTACATCTGGCTACTGGGTCTC

# [hsa-miR-222-3p](#)

# FASTX–Toolkit: summary

+

Detailed control over QC steps and their order

Best suited for relatively short, single–end reads e.g. short RNAs, degradome sequencing or CLIP–Seq data

–

Cannot process paired–end data

Inflexible quality trimming

**Thank you for  
your attention**