### SUPPLEMENTARY INFORMATION

### FastQC reports

Short-insert (200 bp)	2
Short-insert (350 bp)	32

### Mate-pair

2kb	2
5kb 5	52
12kb 6	52

### Linked-reads

R1	72
R2	
(Separate FastQC reports for read 1 and read 2; first 16 bp of read	
barcodes in linked-read sequencing.)	

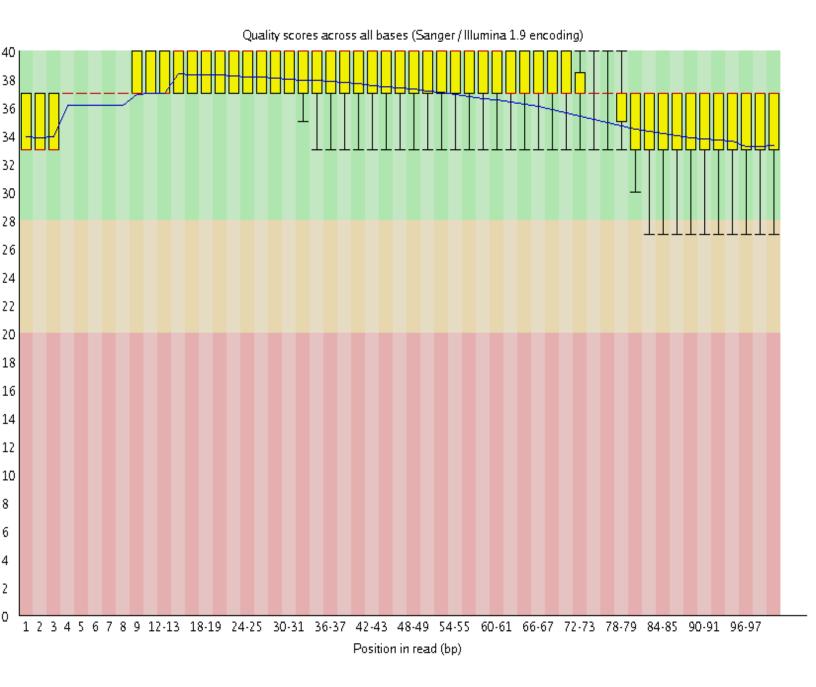
# **PastQC Report**

- 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
- Bequence Duplication Levels
- Overrepresented sequences
- Adapter Content

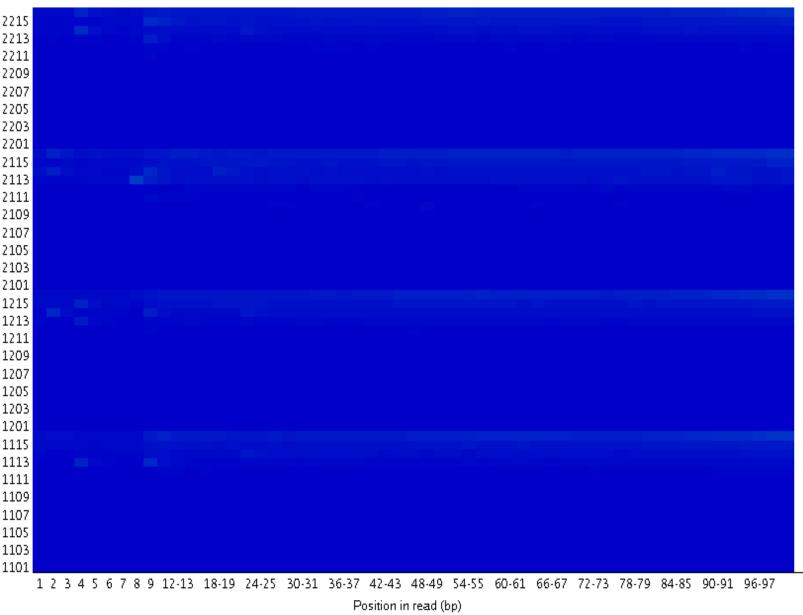
### Basic Statistics

Measure	Value
Filename	stdin
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3214623356
Sequences flagged as poor quality	0
Sequence length	100
%GC	40



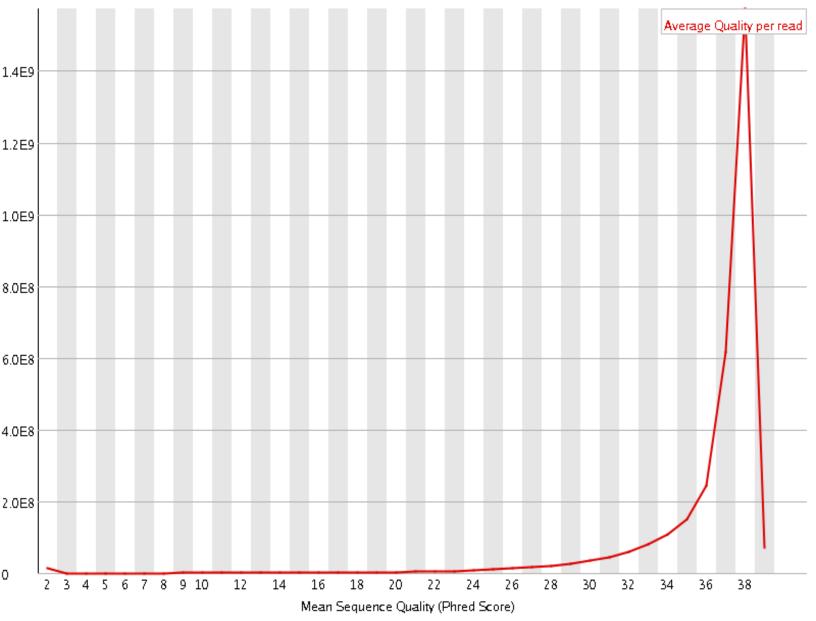






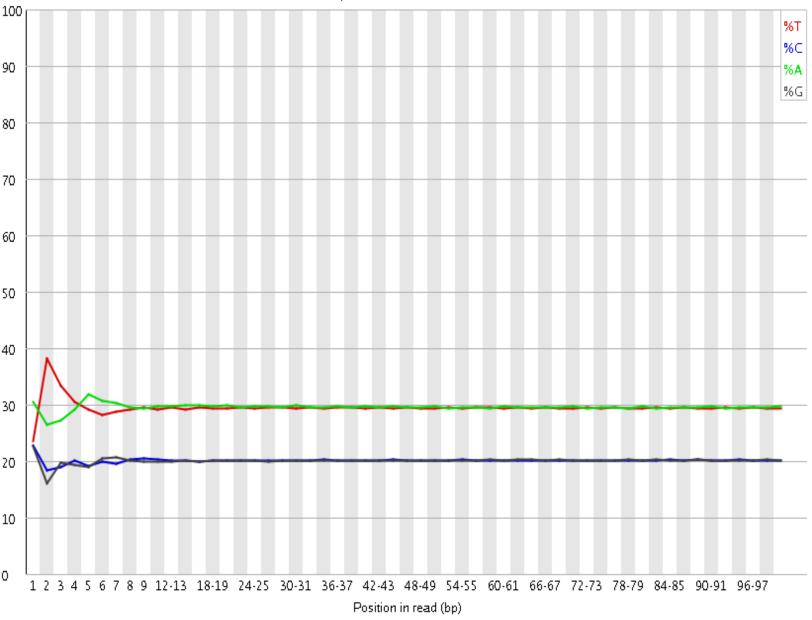


### Quality score distribution over all sequences



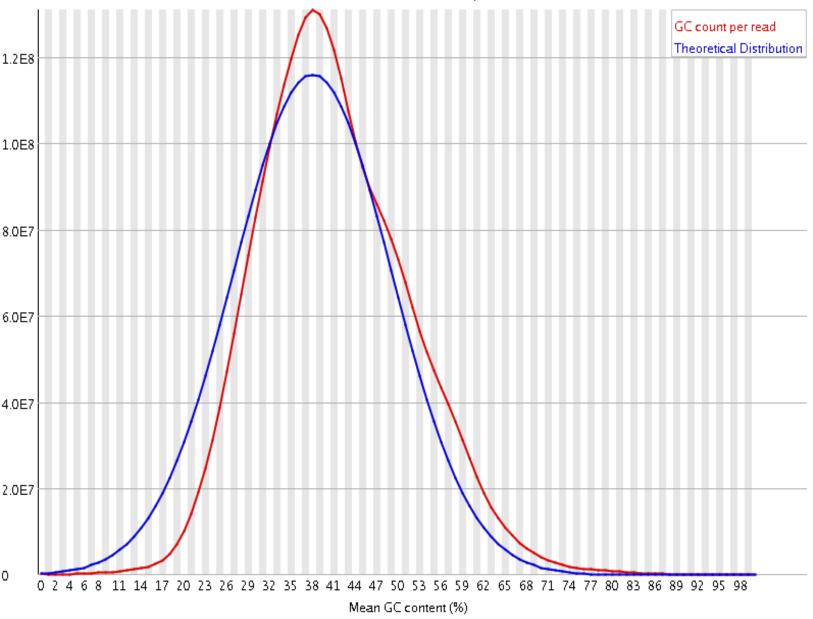


### Sequence content across all bases



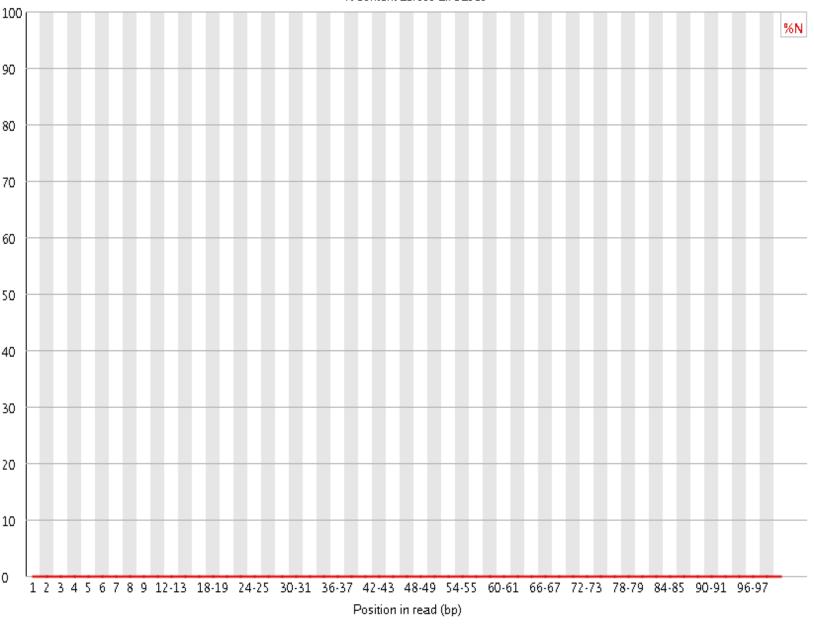


GC distribution over all sequences

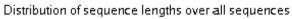


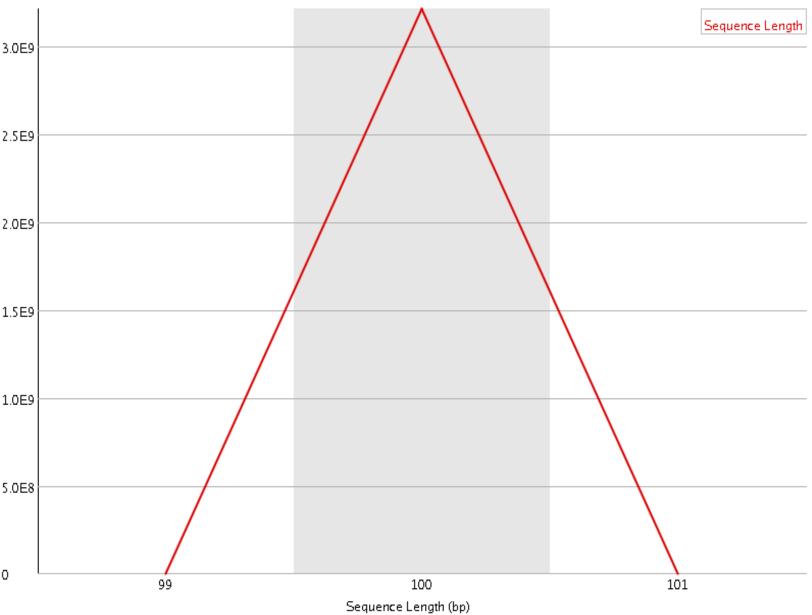


#### N content across all bases



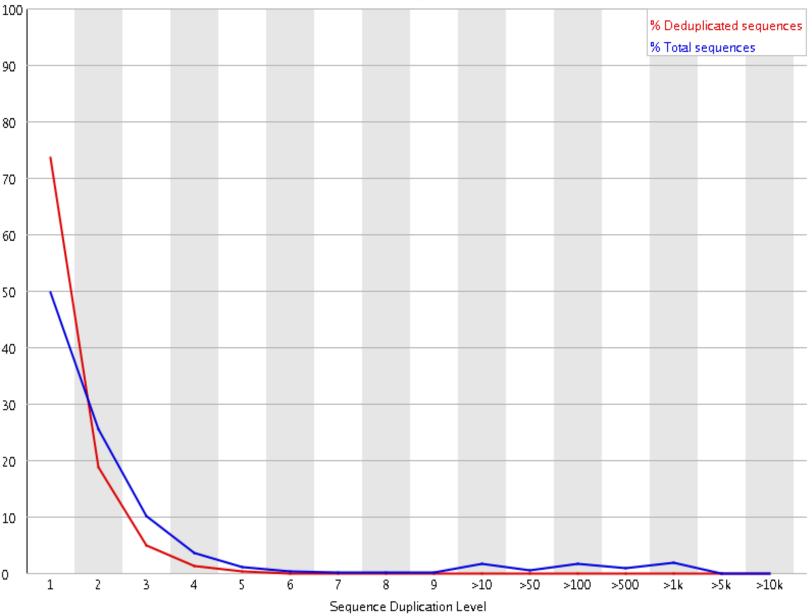








Percent of seqs remaining if deduplicated 67.74%

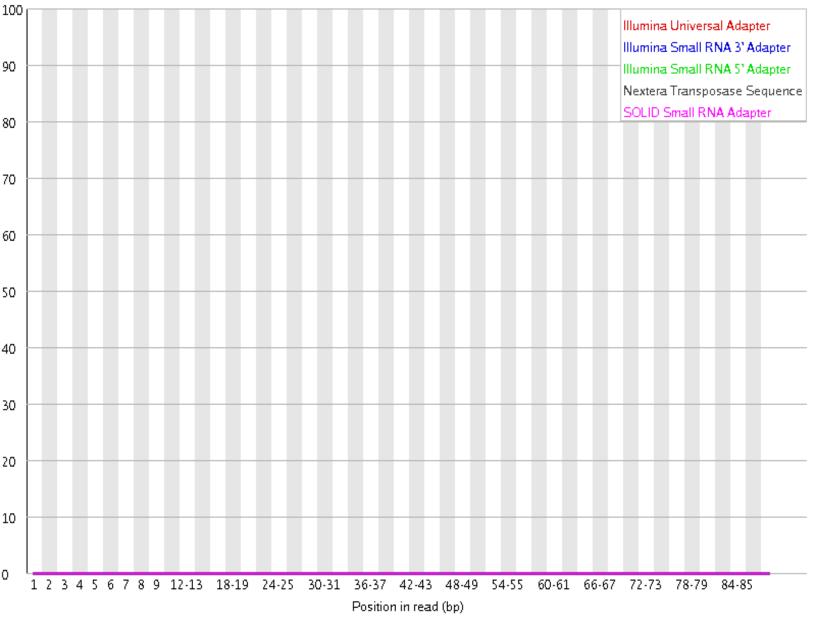


## Overrepresented sequences

No overrepresented sequences



#### % Adapter



Produced by FastQC (version 0.11.7)

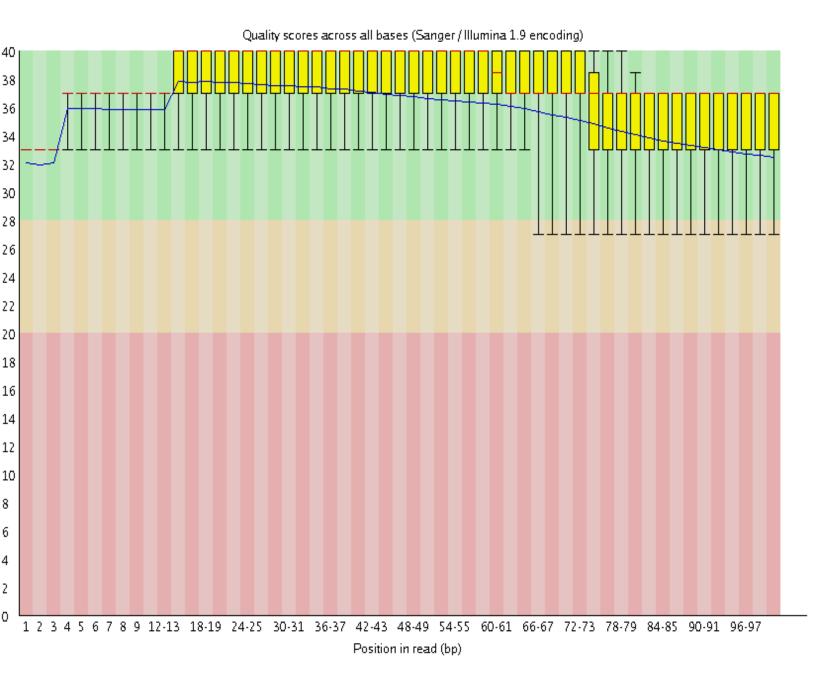
# **PastQC Report**

- Basic Statistics
  - 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

### Basic Statistics

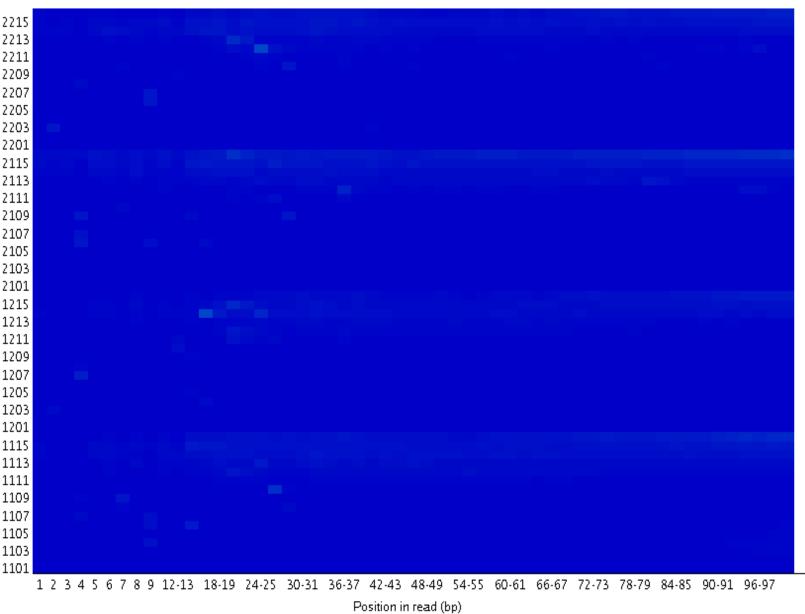
Measure	Value
Filename	stdin
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1280576580
Sequences flagged as poor quality	0
Sequence length	100
%GC	40





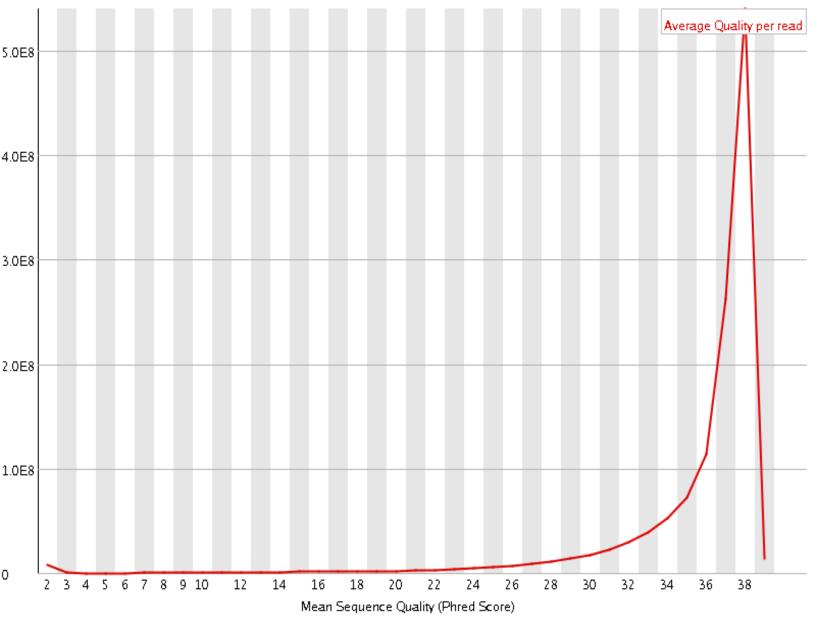


Quality per tile



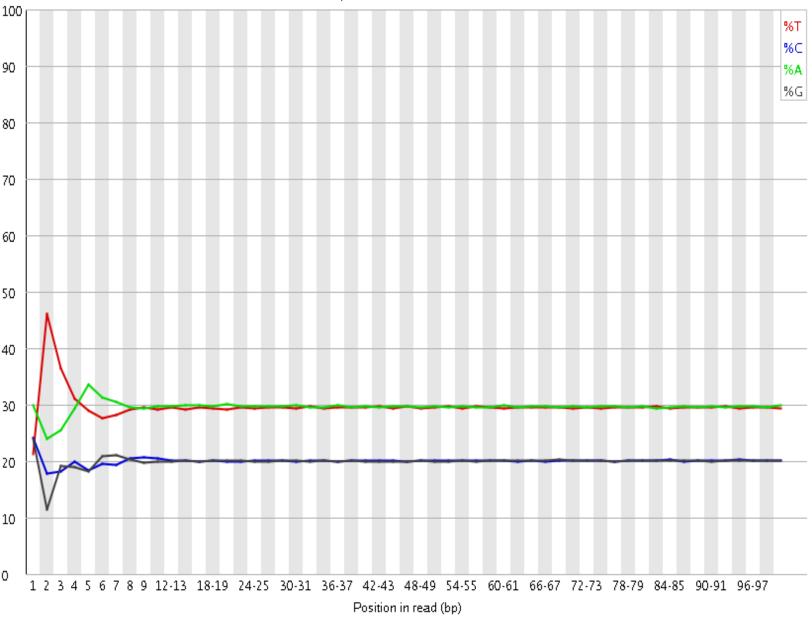


### Quality score distribution over all sequences



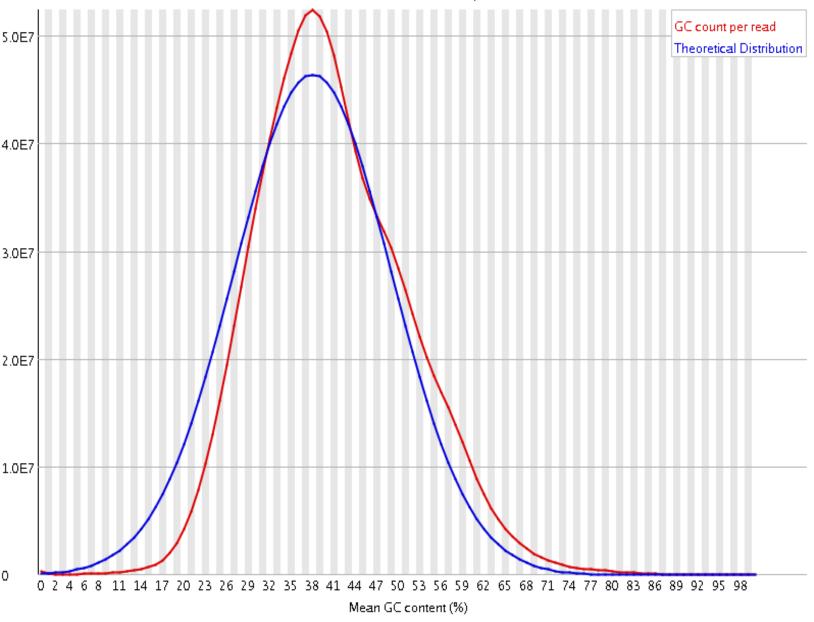


### Sequence content across all bases



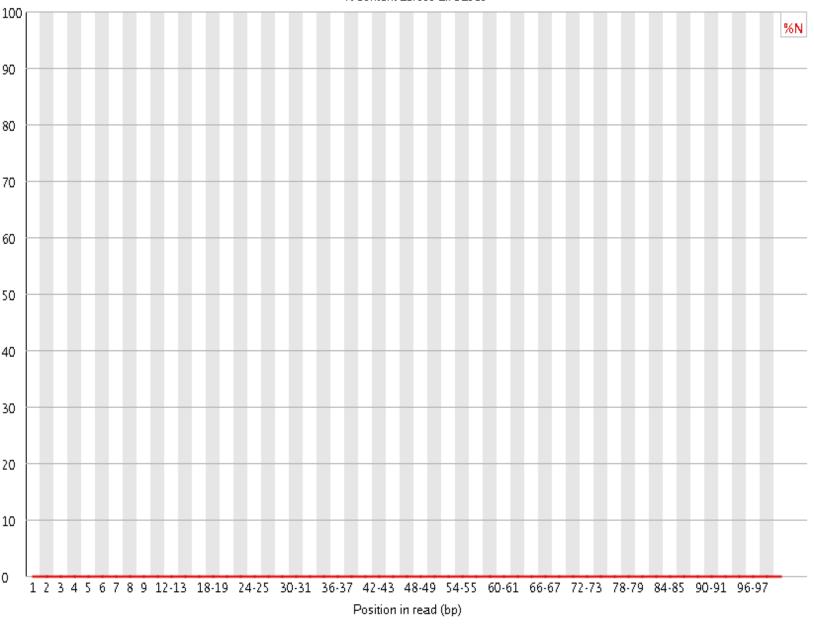


### GC distribution over all sequences

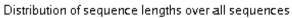


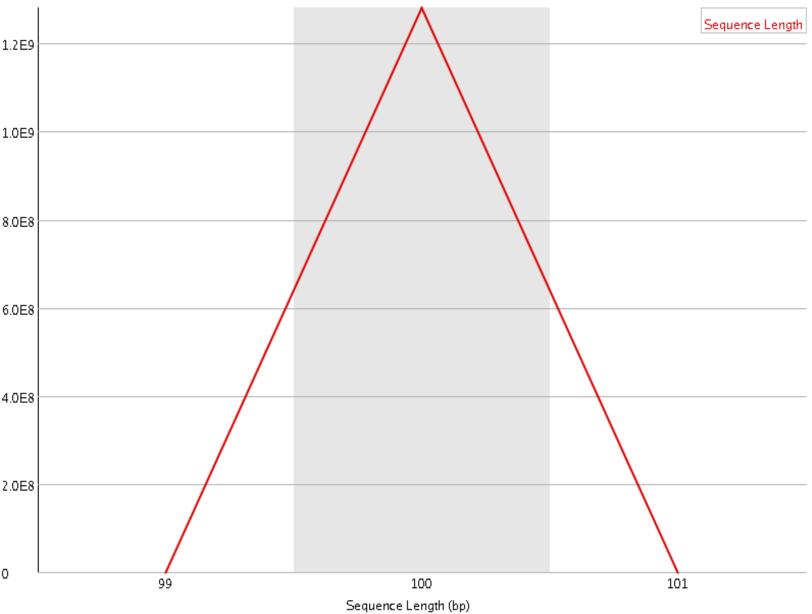


#### N content across all bases



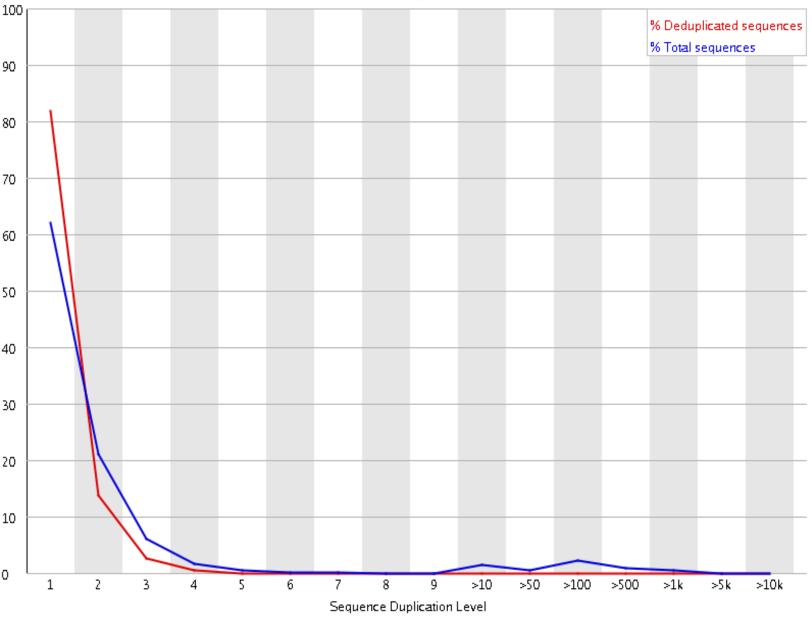








Percent of seqs remaining if deduplicated 75.85%

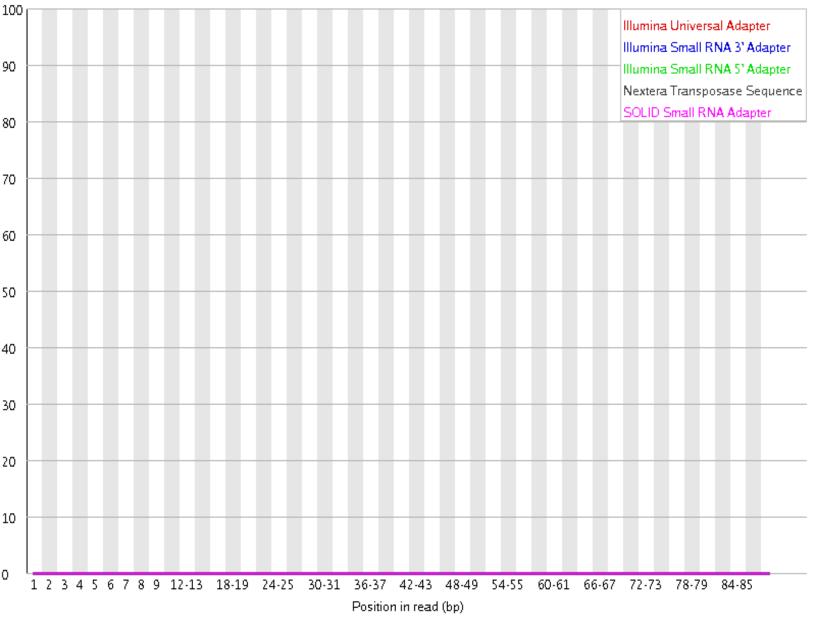


## Overrepresented sequences

No overrepresented sequences



#### % Adapter



Produced by FastQC (version 0.11.7)

# **PastQC Report**

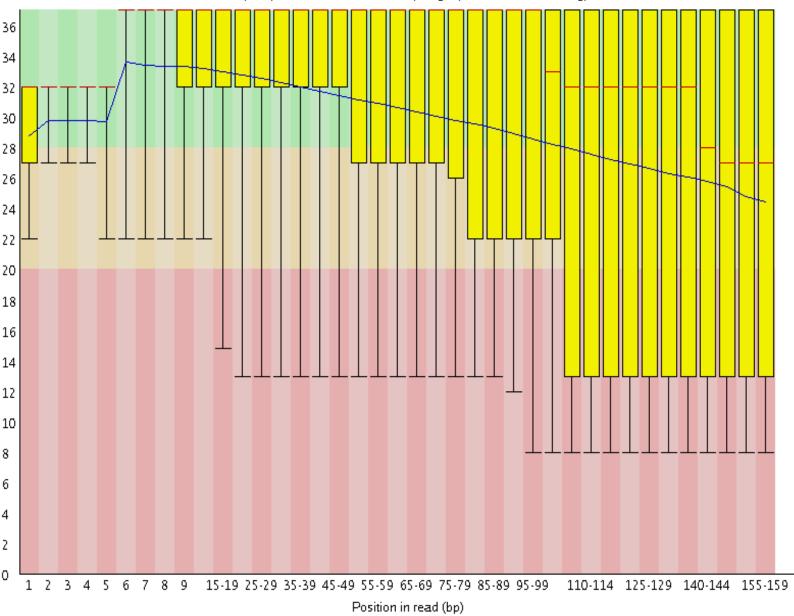
- Basic Statistics
  - Per base sequence quality
  - Per tile sequence quality
  - Per sequence quality scores
  - Per base sequence content
  - Per sequence GC content
- Per base N content
- Bequence Length Distribution
- Bequence Duplication Levels
- Overrepresented sequences
- Adapter Content

### Basic Statistics

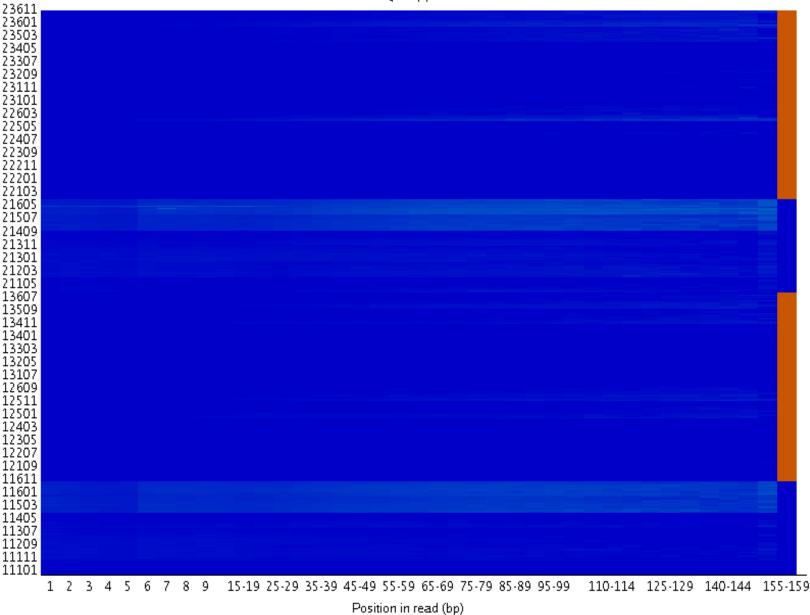
Measure	Value	
Filename	stdin	
File type	Conventional base calls	
Encoding	Sanger / Illumina 1.9	
Total Sequences	553277870	
Sequences flagged as poor quality	0	
Sequence length	35-159	
%GC	41	



Quality scores across all bases (Sanger / Illumina 1.9 encoding)

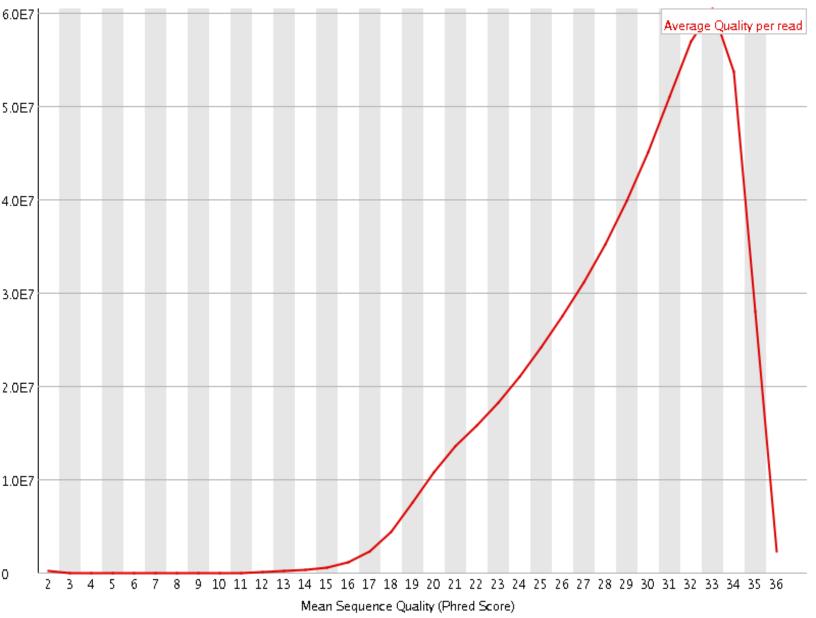






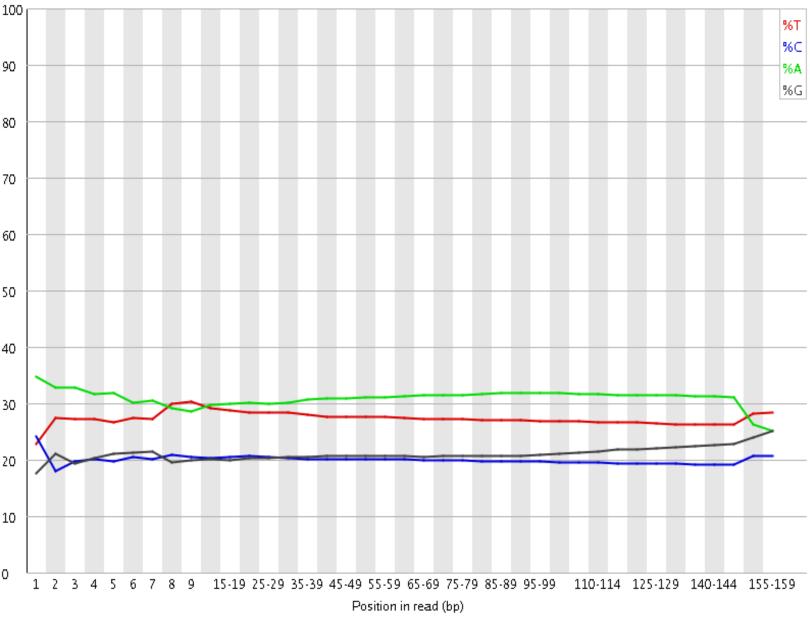


### Quality score distribution over all sequences



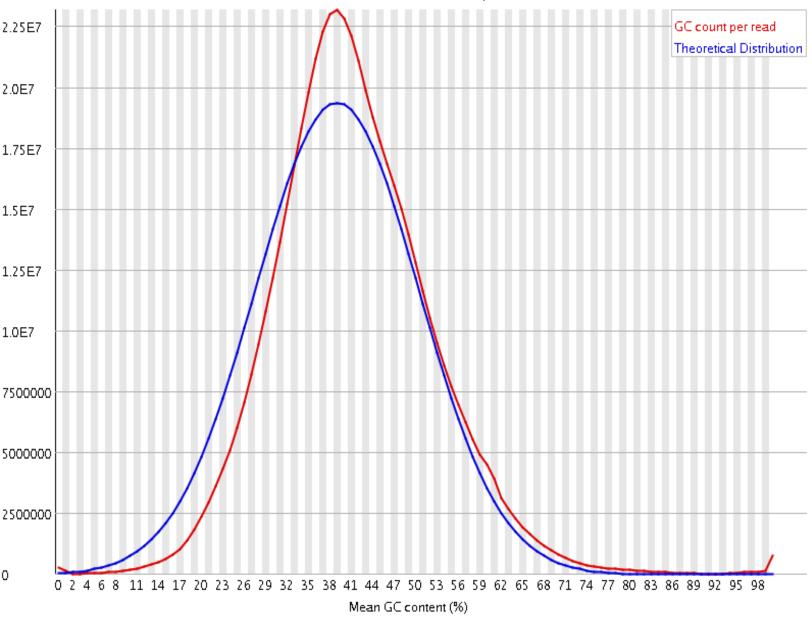


### Sequence content across all bases



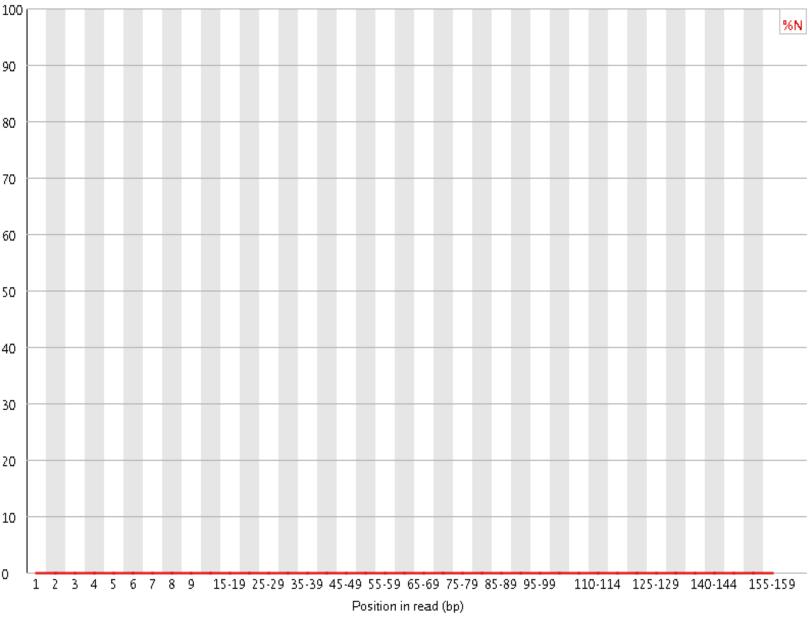


GC distribution over all sequences



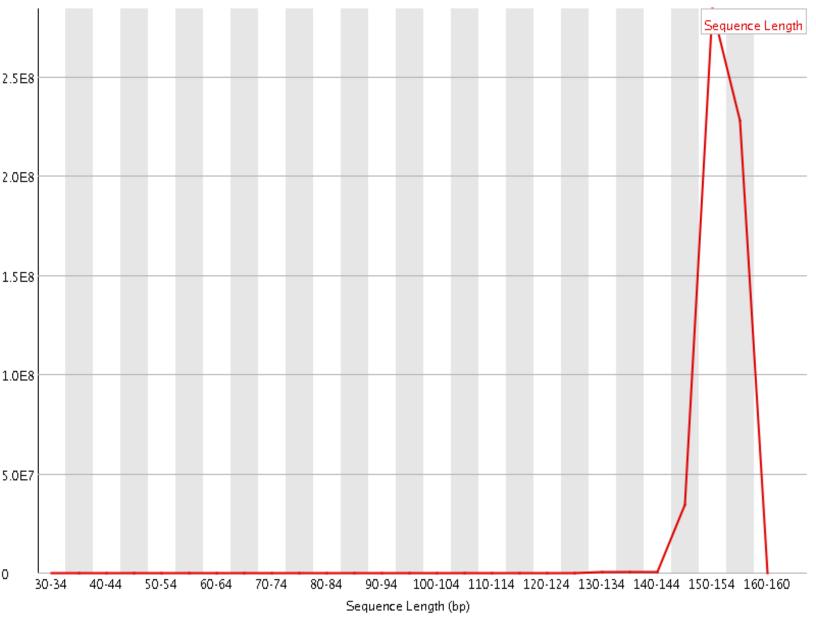


#### N content across all bases



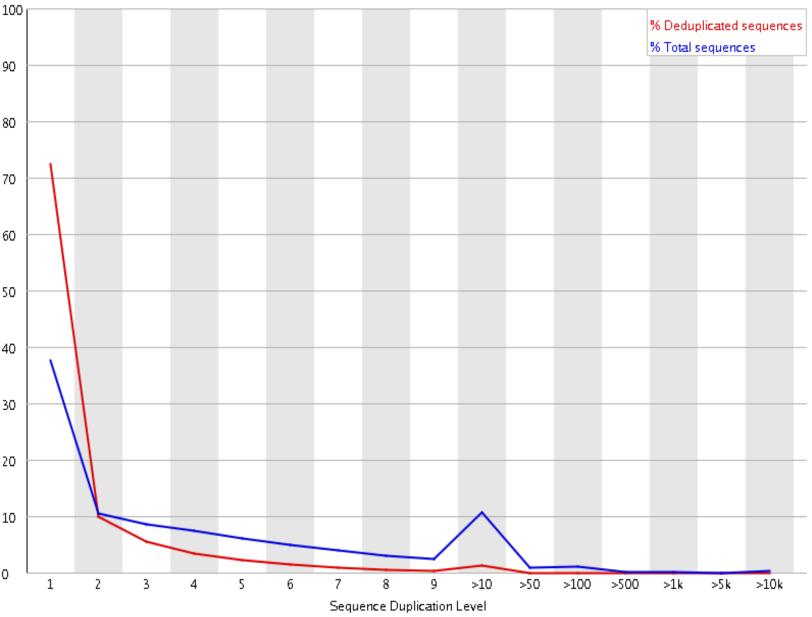
## Sequence Length Distribution

### Distribution of sequence lengths over all sequences



## Sequence Duplication Levels

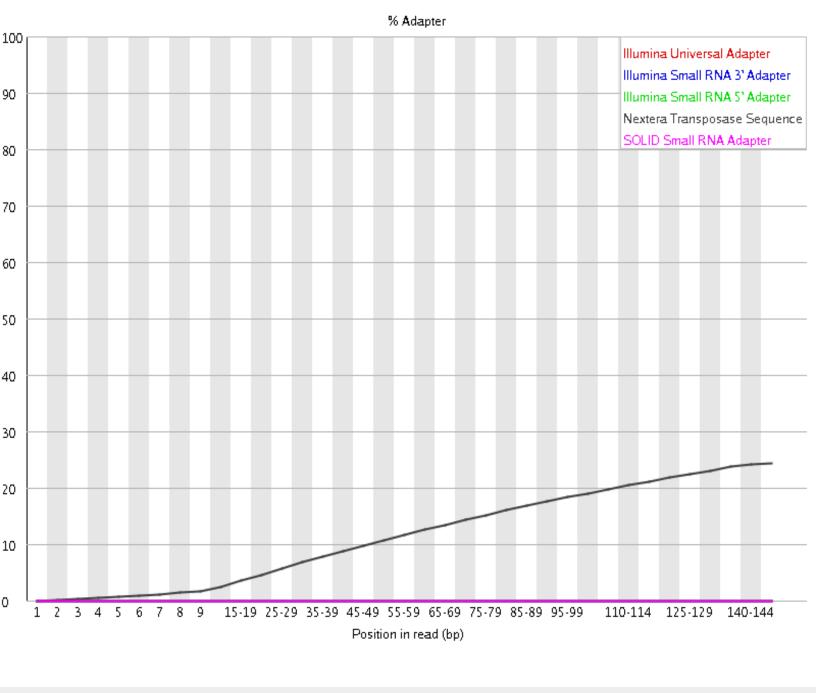
Percent of seqs remaining if deduplicated 52.18%



## Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	1464007	0.2646061010898556	No Hit





Produced by FastQC (version 0.11.7)

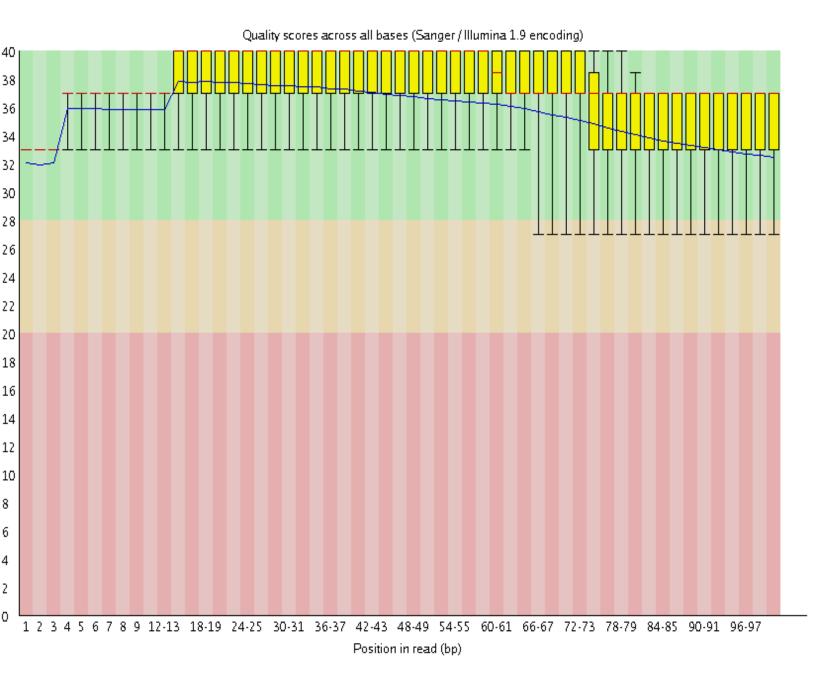
# **PastQC Report**

- Basic Statistics
  - 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

### Basic Statistics

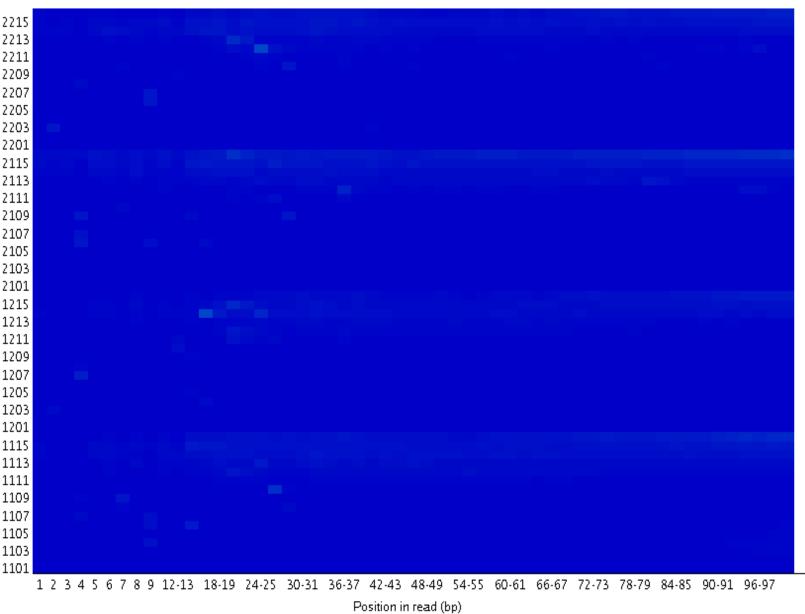
Measure	Value
Filename	stdin
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1280576580
Sequences flagged as poor quality	0
Sequence length	100
%GC	40





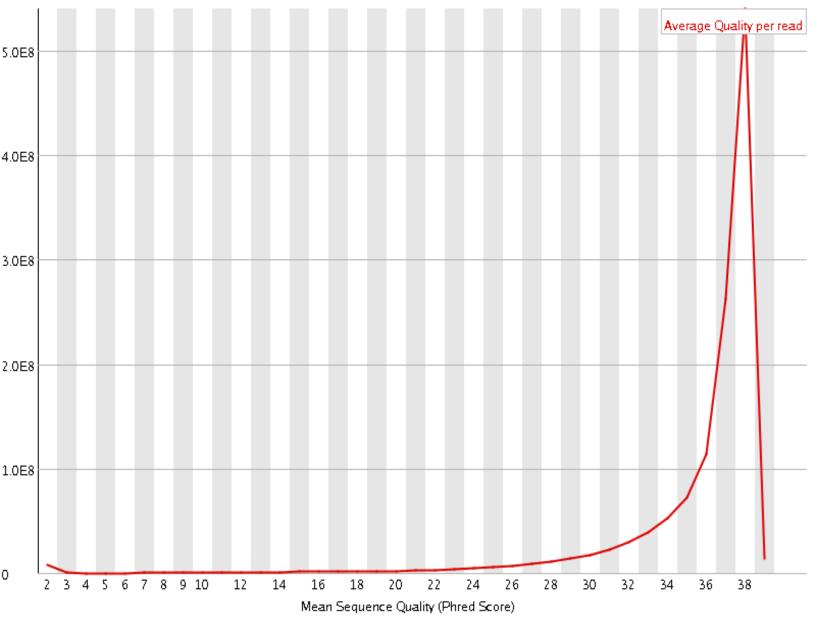


Quality per tile



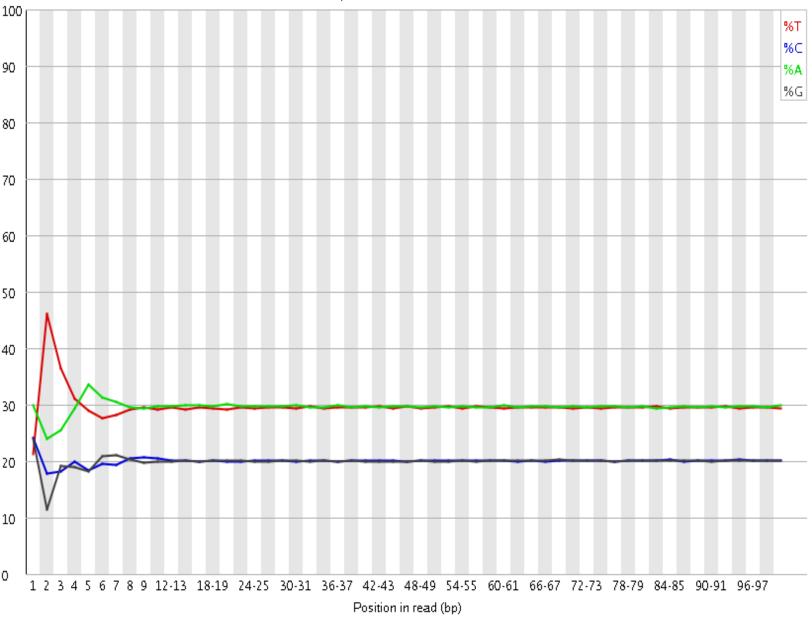


### Quality score distribution over all sequences



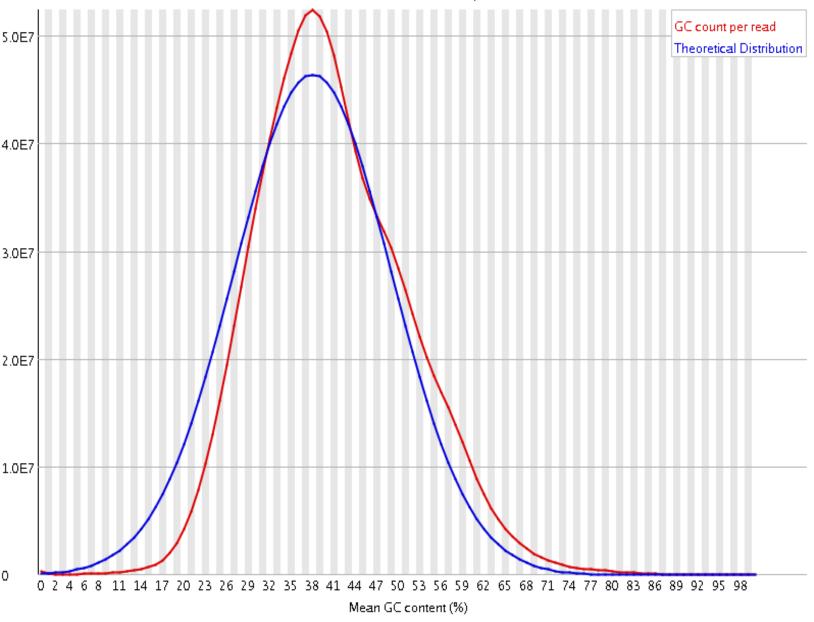


### Sequence content across all bases



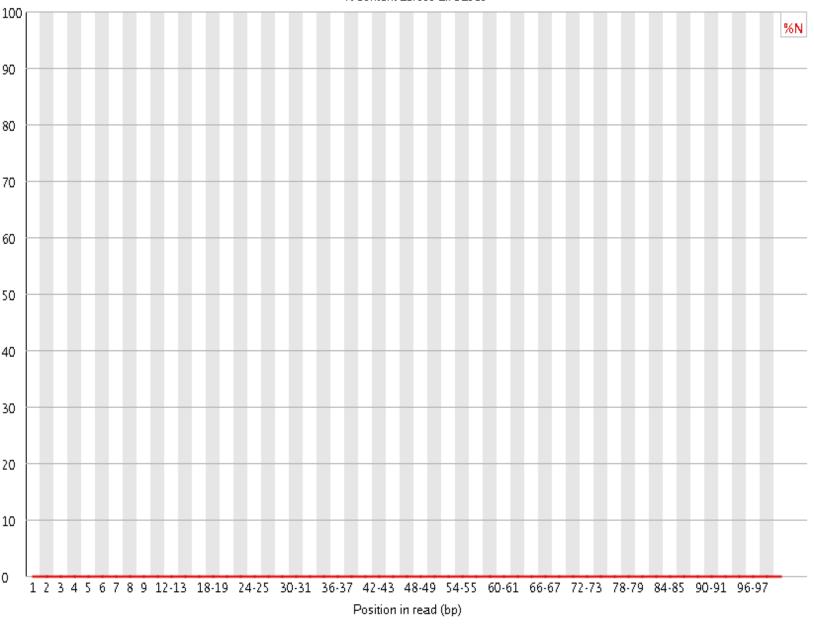


#### GC distribution over all sequences

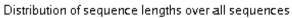


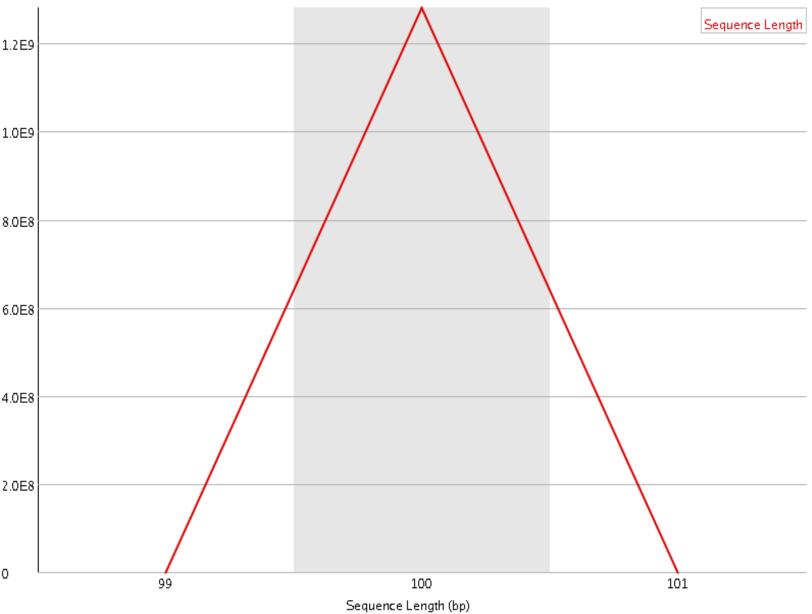


#### N content across all bases



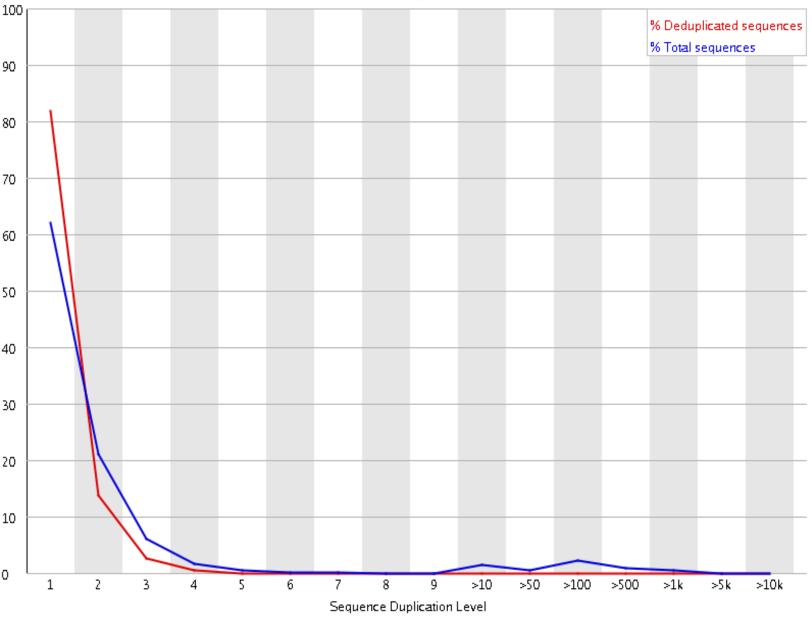








Percent of seqs remaining if deduplicated 75.85%

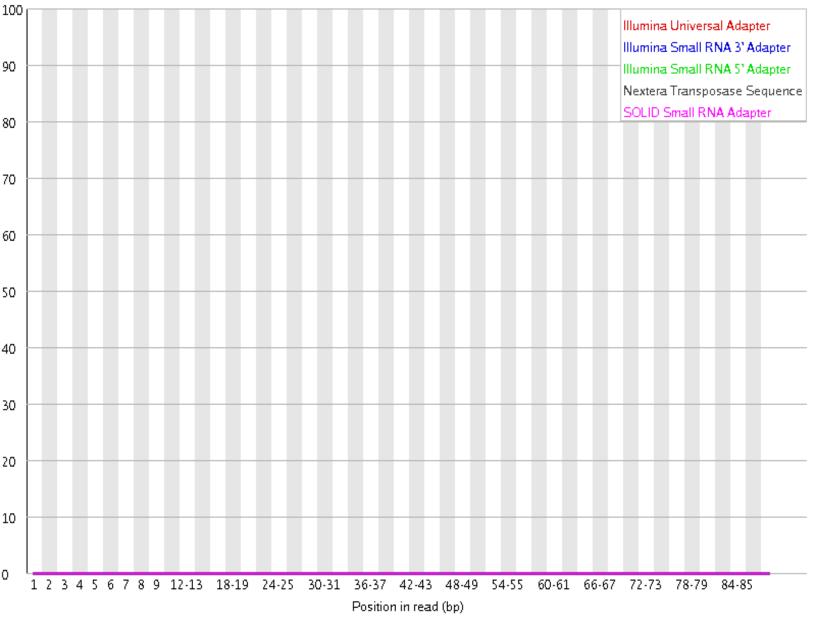


### Overrepresented sequences

No overrepresented sequences



#### % Adapter



Produced by FastQC (version 0.11.7)

# **PastQC Report**

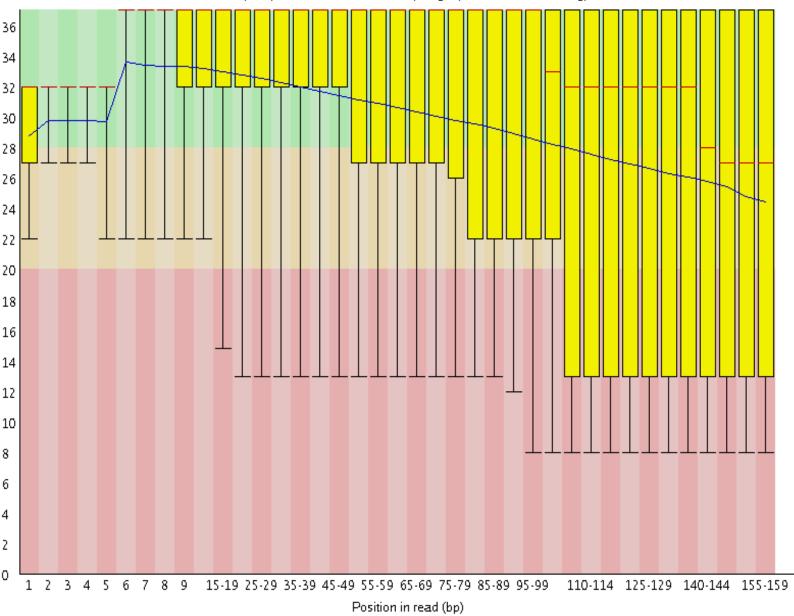
- Basic Statistics
  - Per base sequence quality
  - Per tile sequence quality
  - Per sequence quality scores
  - Per base sequence content
  - Per sequence GC content
- Per base N content
- Bequence Length Distribution
- Bequence Duplication Levels
- Overrepresented sequences
- Adapter Content

### Basic Statistics

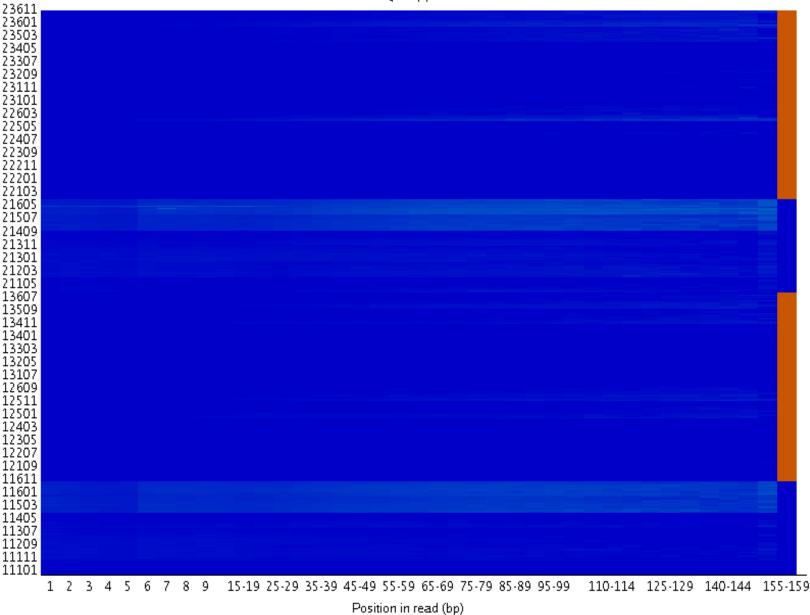
Measure	Value
Filename	stdin
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	553277870
Sequences flagged as poor quality	0
Sequence length	35-159
%GC	41



Quality scores across all bases (Sanger / Illumina 1.9 encoding)

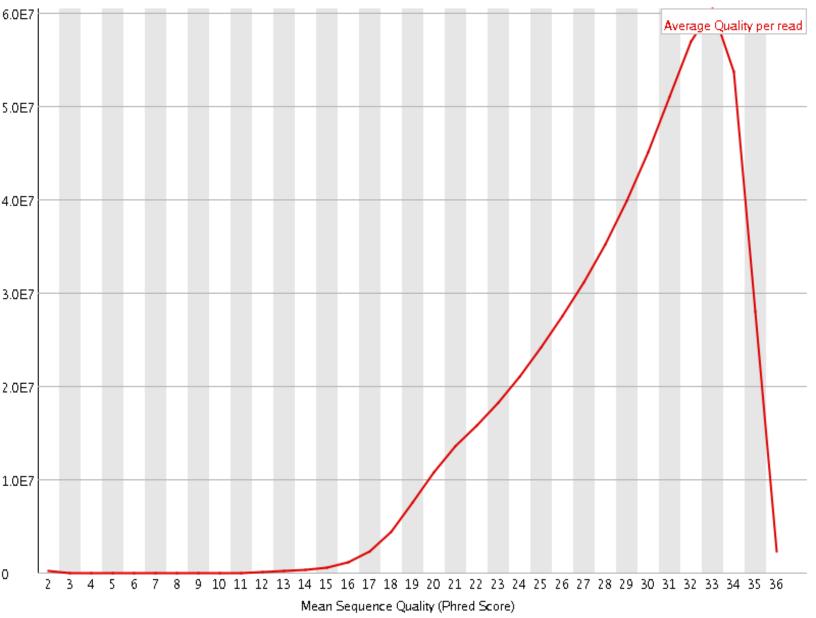






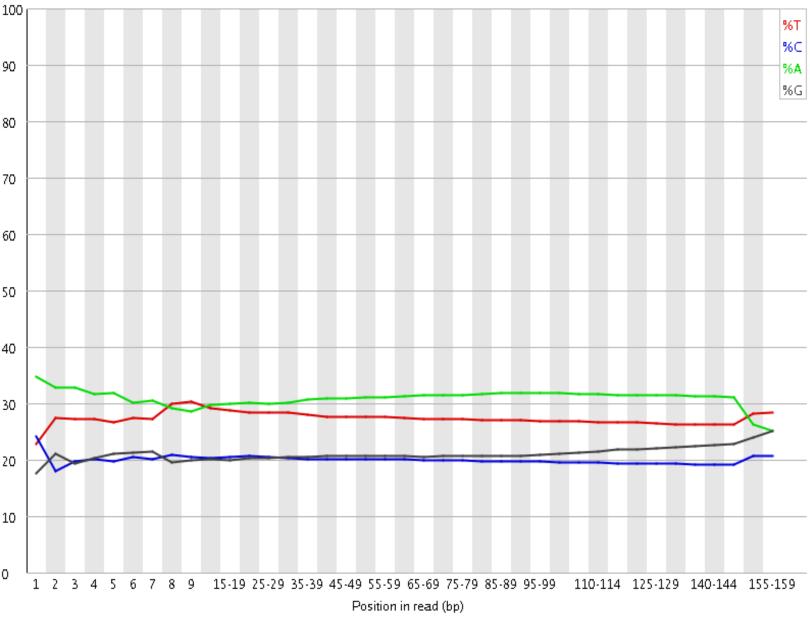


#### Quality score distribution over all sequences



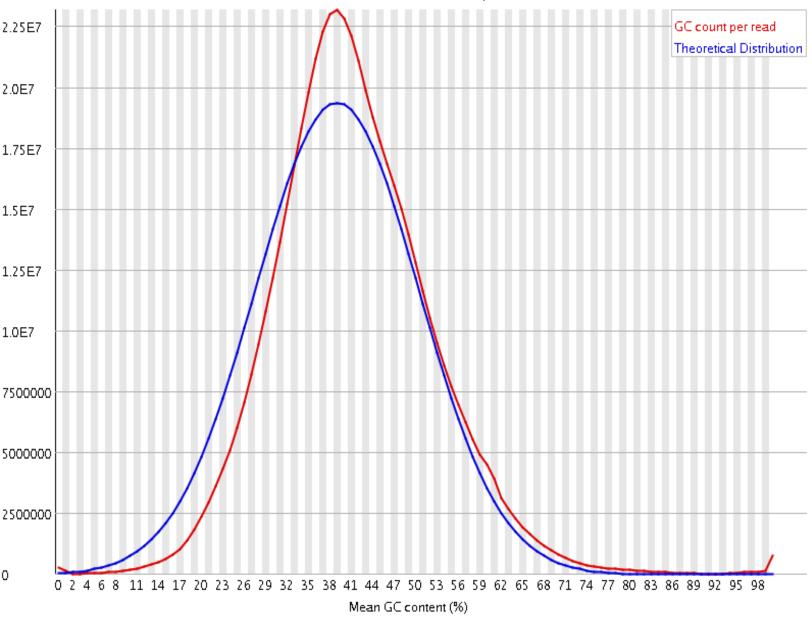


#### Sequence content across all bases



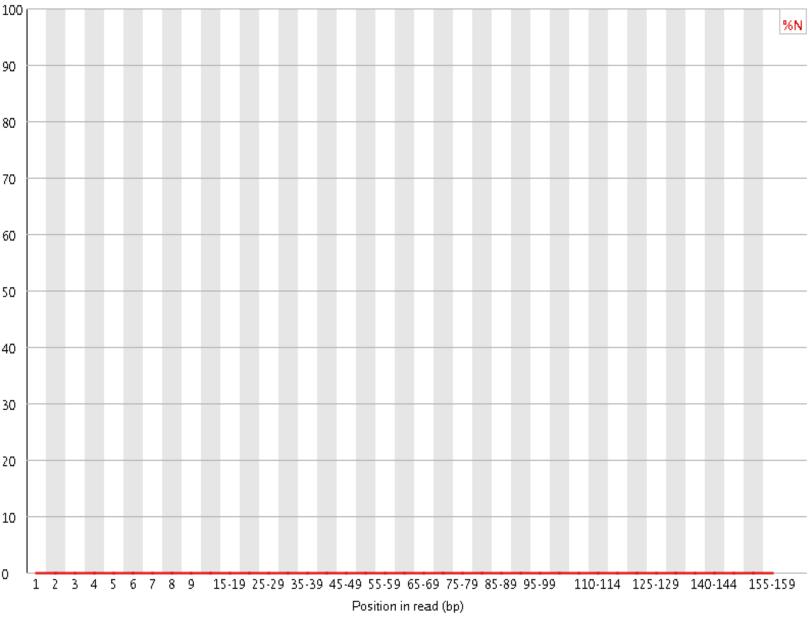


GC distribution over all sequences



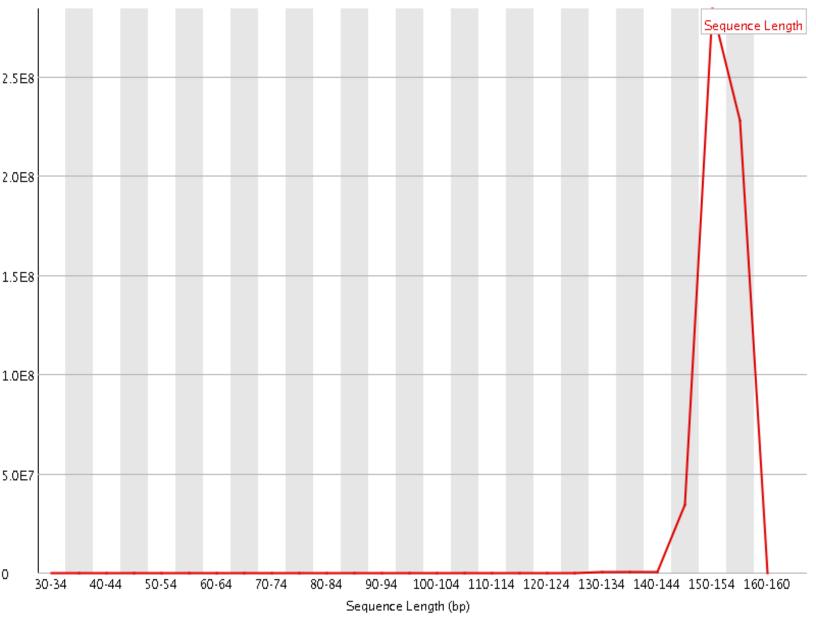


#### N content across all bases



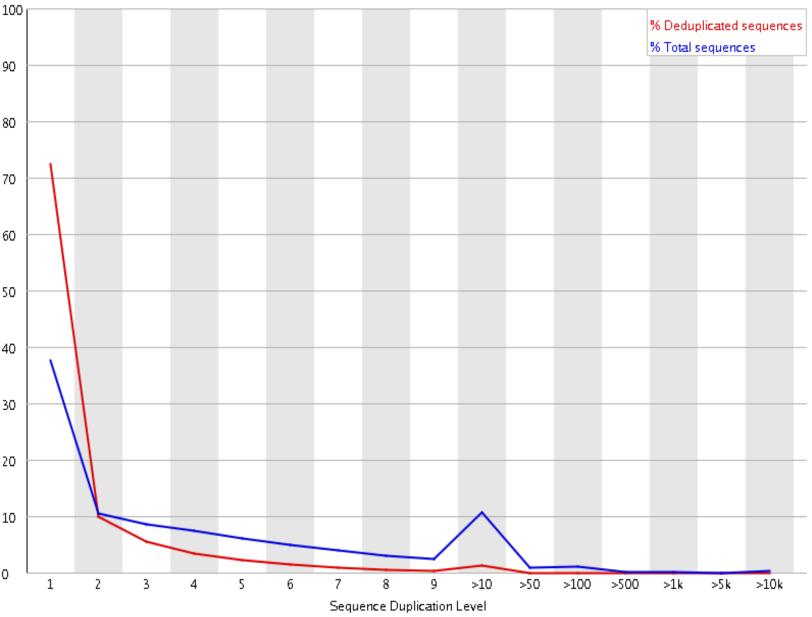
# Sequence Length Distribution

#### Distribution of sequence lengths over all sequences



### Sequence Duplication Levels

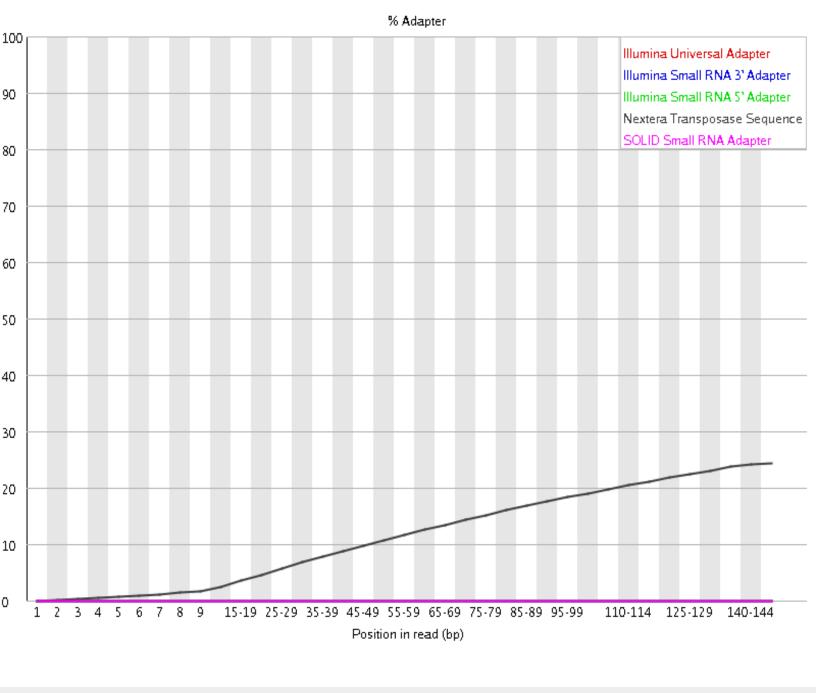
Percent of seqs remaining if deduplicated 52.18%



### Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	1464007	0.2646061010898556	No Hit





Produced by FastQC (version 0.11.7)

# **PastQC Report**

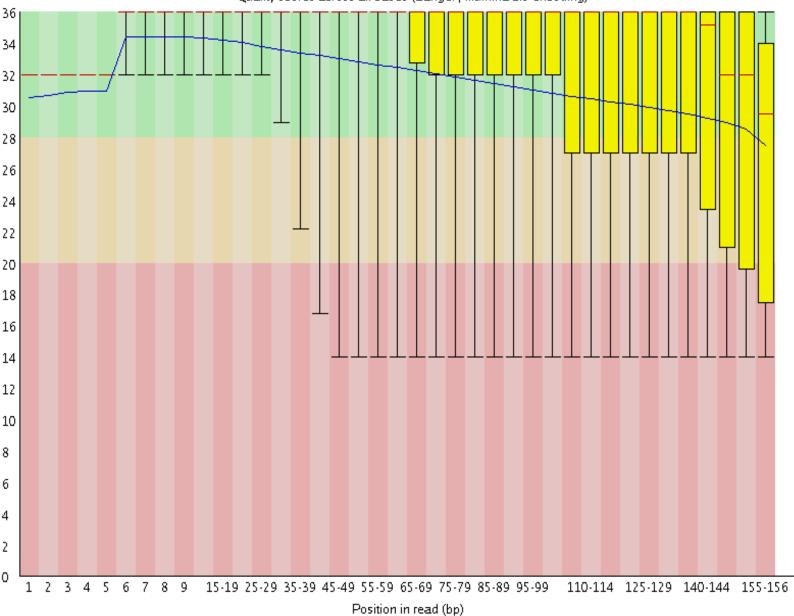
- Basic Statistics
  - Per base sequence quality
  - Per tile sequence quality
  - Per sequence quality scores
  - Per base sequence content
  - Per sequence GC content
- Per base N content
- Bequence Length Distribution
- Bequence Duplication Levels
- Overrepresented sequences
- Adapter Content



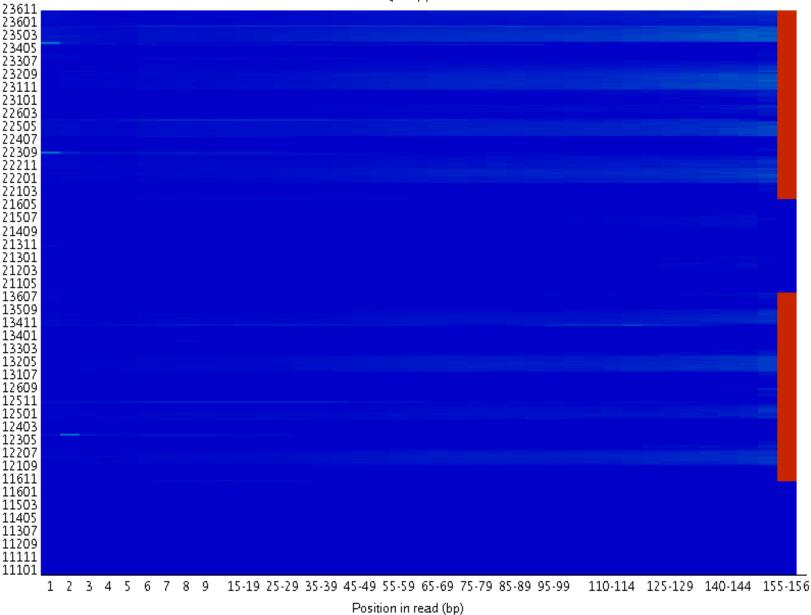
Measure	Value
Filename	stdin
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	338626576
Sequences flagged as poor quality	0
Sequence length	35-156
%GC	43



Quality scores across all bases (Sanger / Illumina 1.9 encoding)

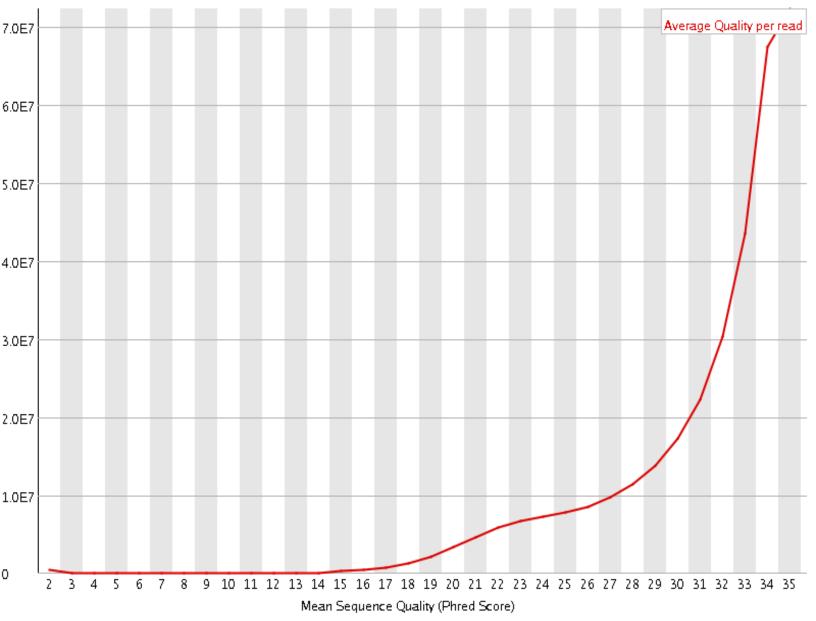






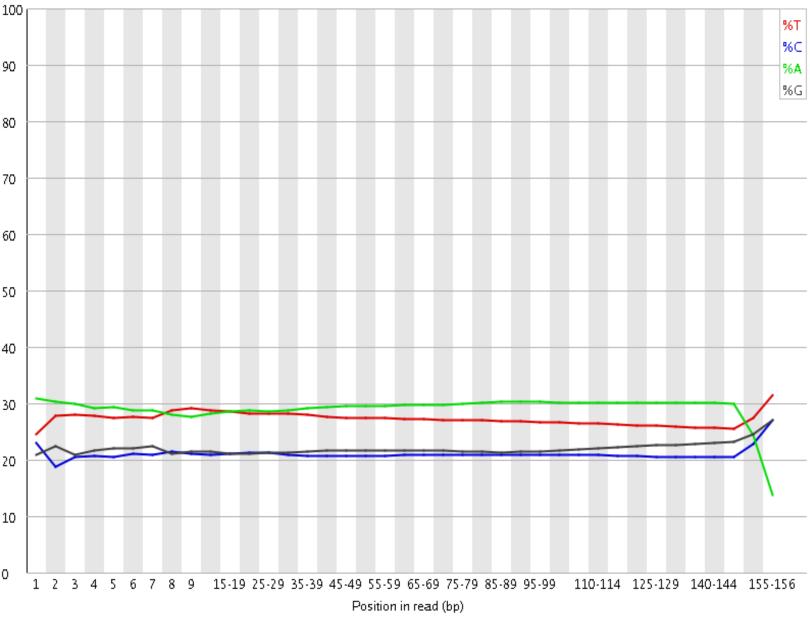


#### Quality score distribution over all sequences



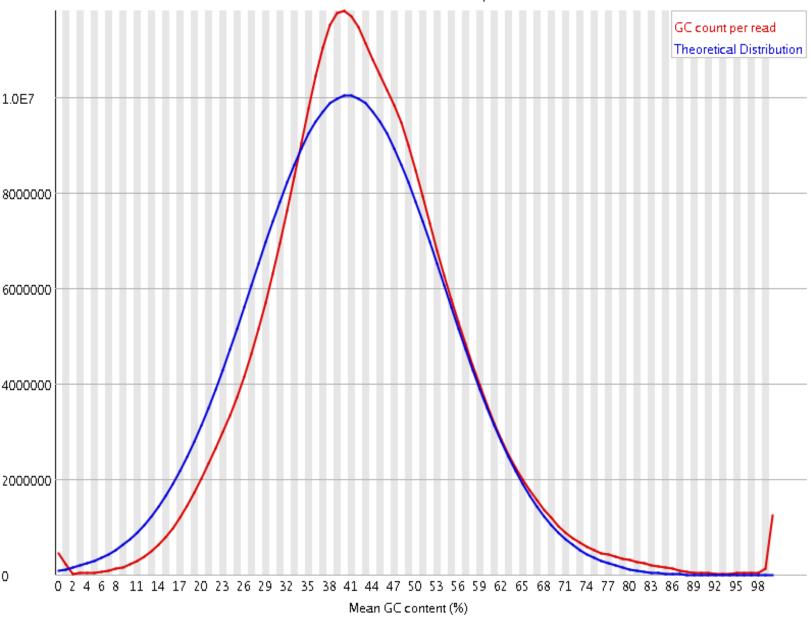


#### Sequence content across all bases



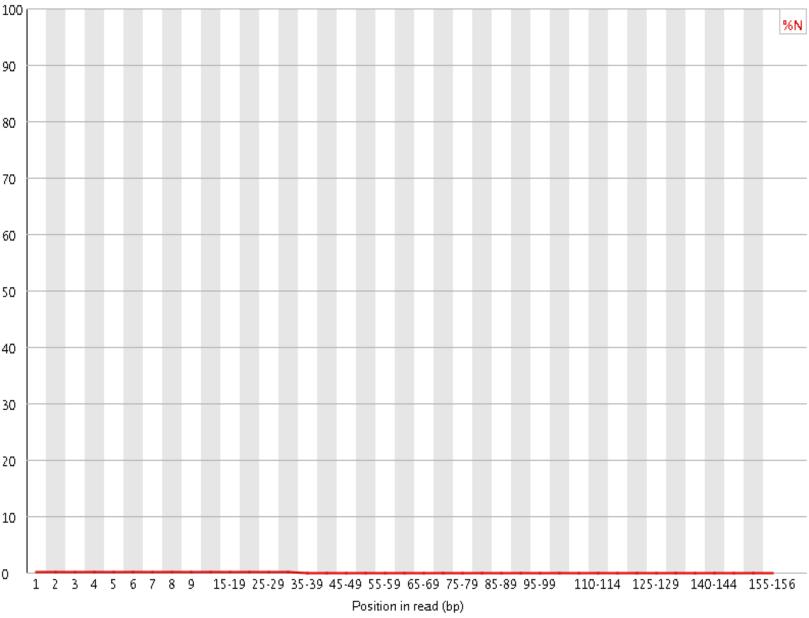


GC distribution over all sequences



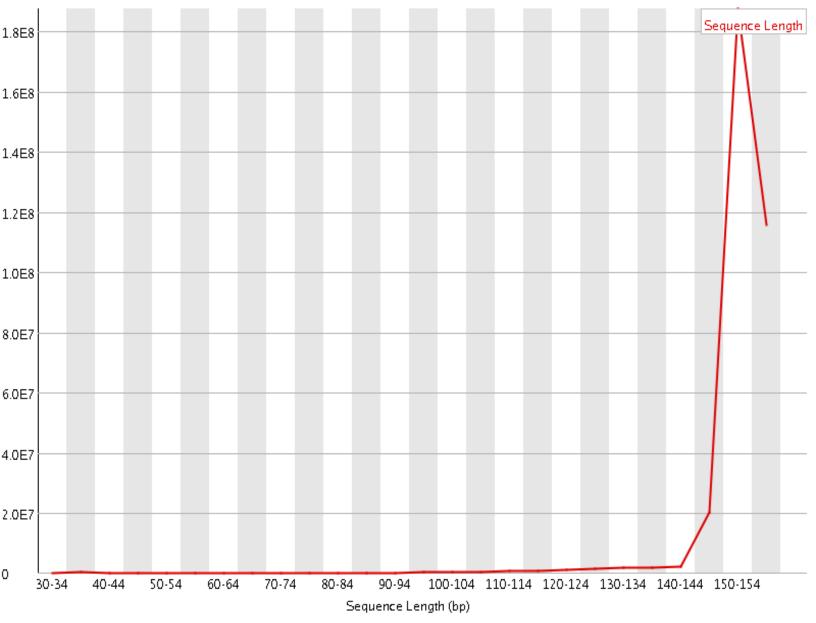


#### N content across all bases



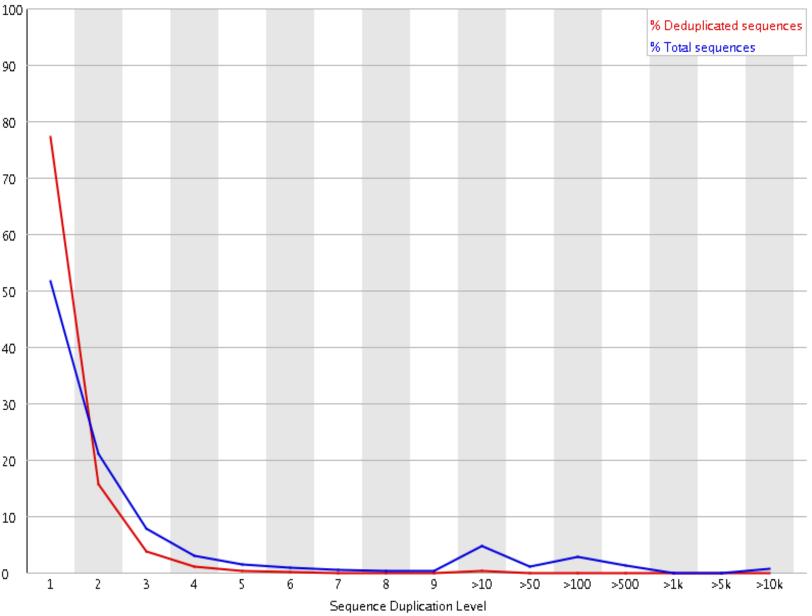
# Sequence Length Distribution

#### Distribution of sequence lengths over all sequences



### Sequence Duplication Levels

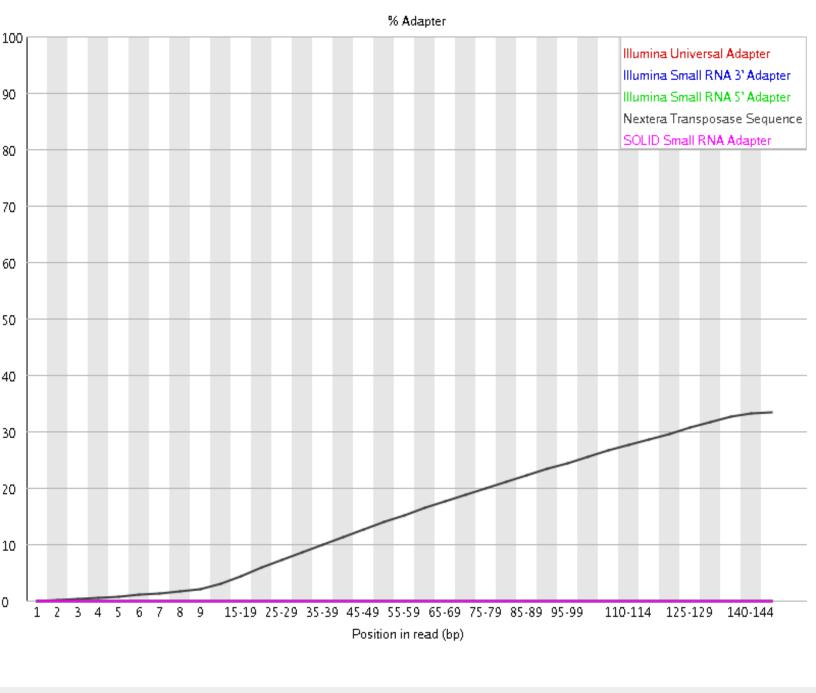
Percent of seqs remaining if deduplicated 66.96%



## Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	2384371	0.7041299085751616	No Hit
NNNNNNNNNNNNNNNNNNNNNNNNNNN	414840	0.1225066280680817	No Hit





Produced by FastQC (version 0.11.7)

# **PastQC Report**

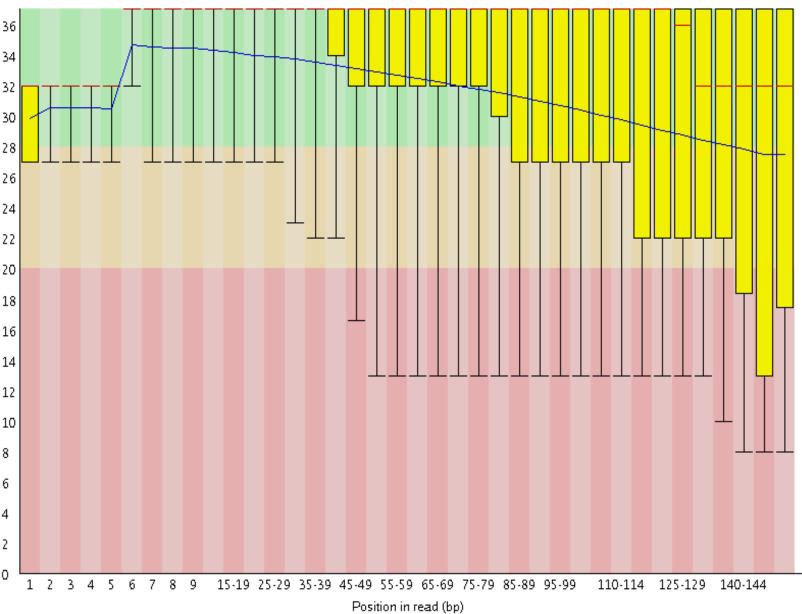
- Basic Statistics
  - Per tile sequence quality
  - Per sequence quality scores
  - Per base sequence content
  - Per sequence GC content
- Per base N content
- Bequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content



Measure	Value
Filename	stdin
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	366199264
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	40

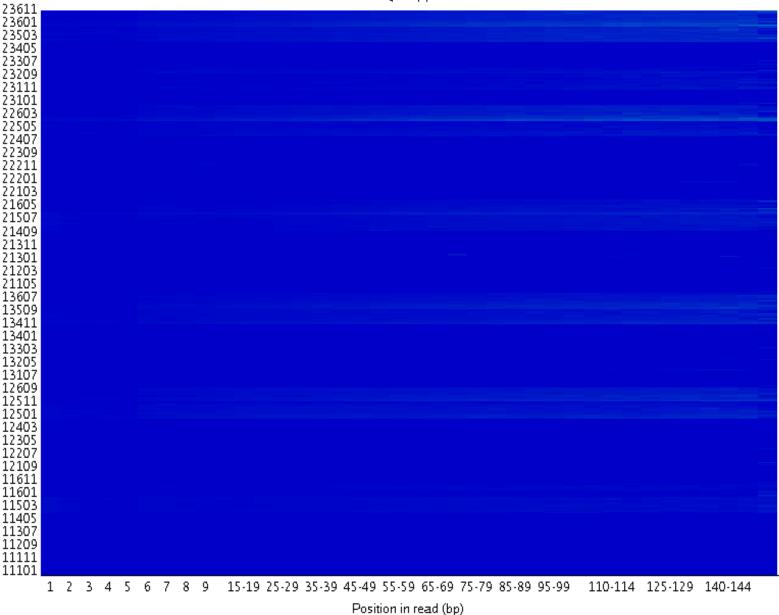


Quality scores across all bases (Sanger / Illumina 1.9 encoding)



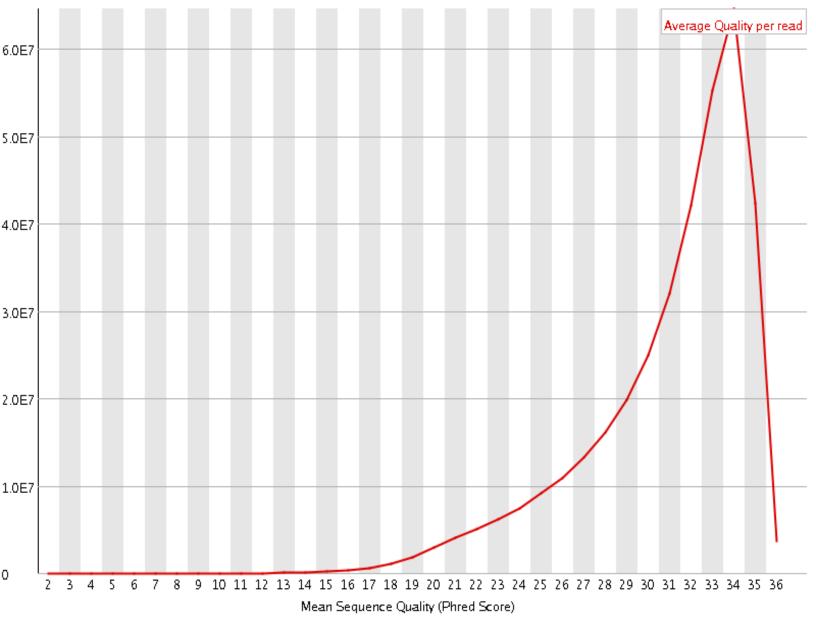






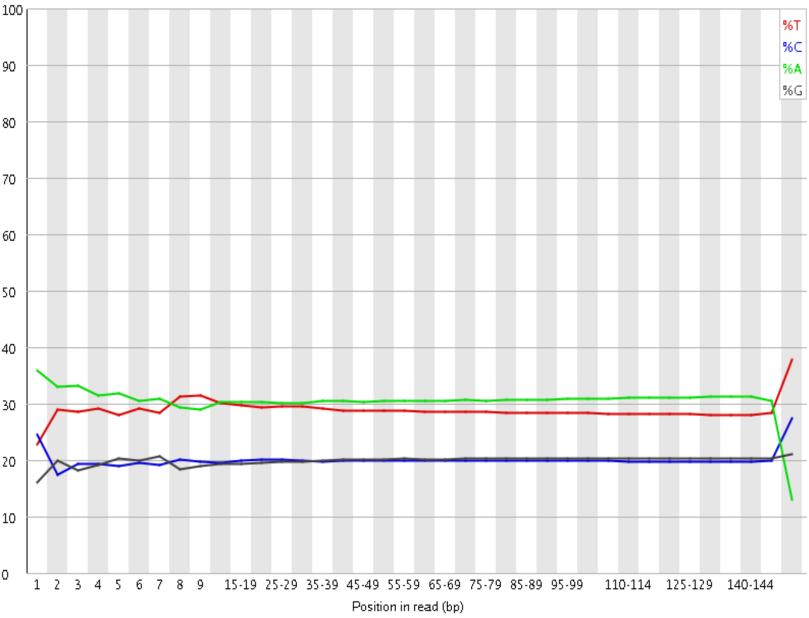


#### Quality score distribution over all sequences



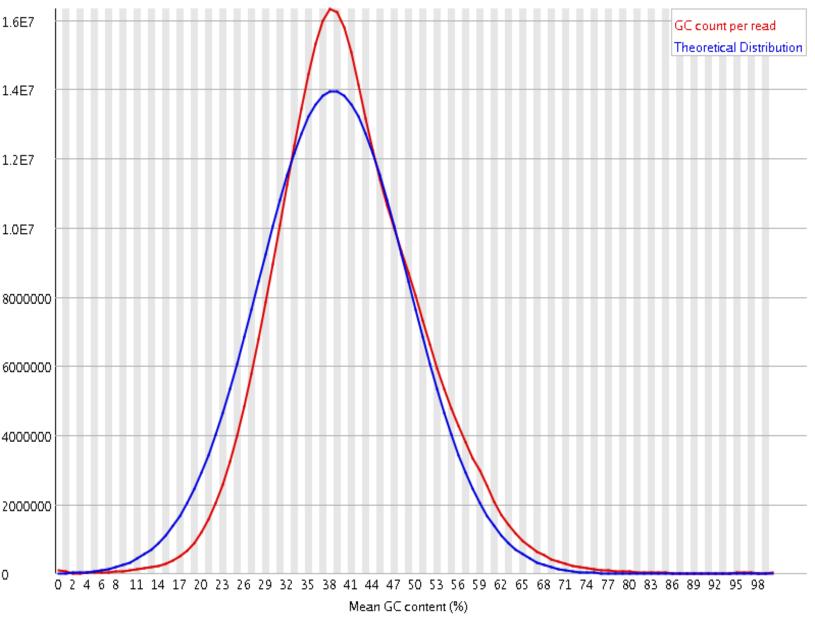


#### Sequence content across all bases



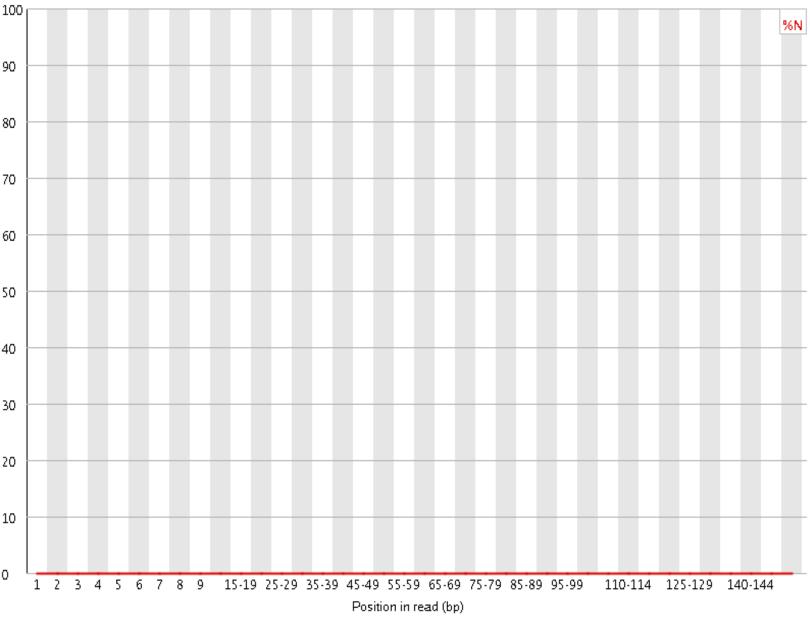


GC distribution over all sequences



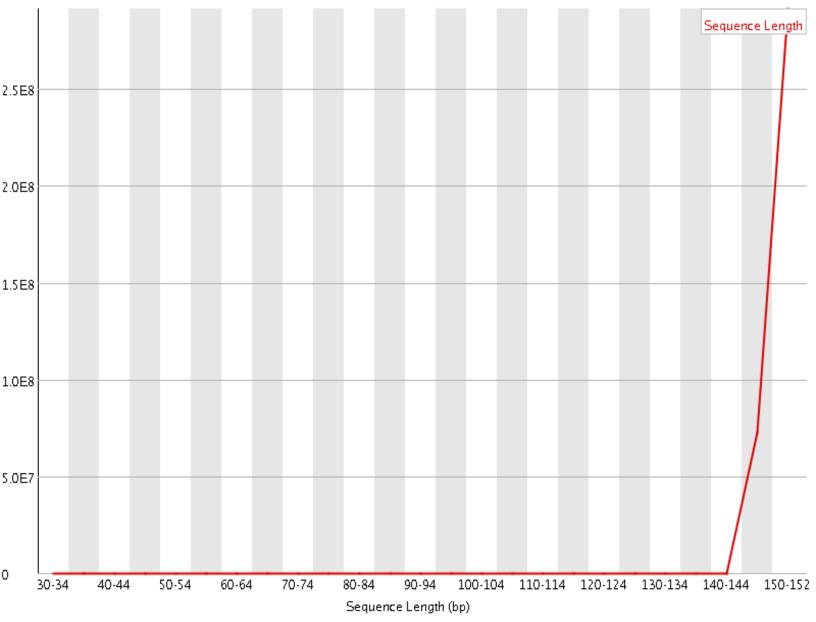


#### N content across all bases



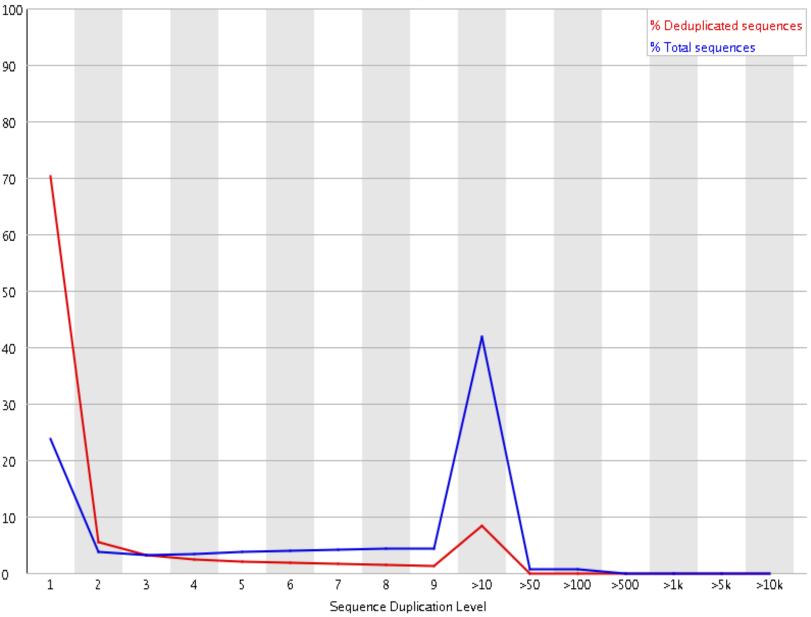
## Sequence Length Distribution

Distribution of sequence lengths over all sequences



## Sequence Duplication Levels

Percent of seqs remaining if deduplicated 33.98%

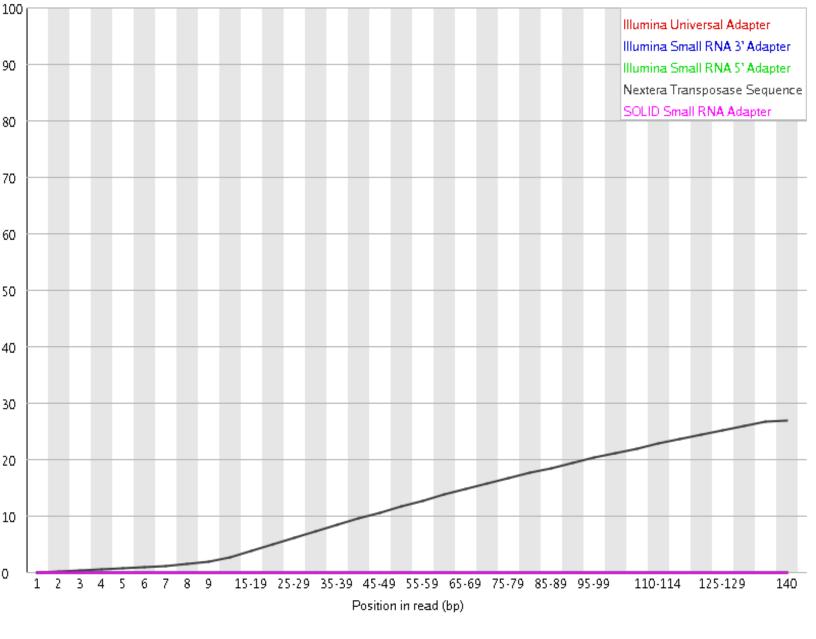


### Overrepresented sequences

No overrepresented sequences







Produced by FastQC (version 0.11.7)

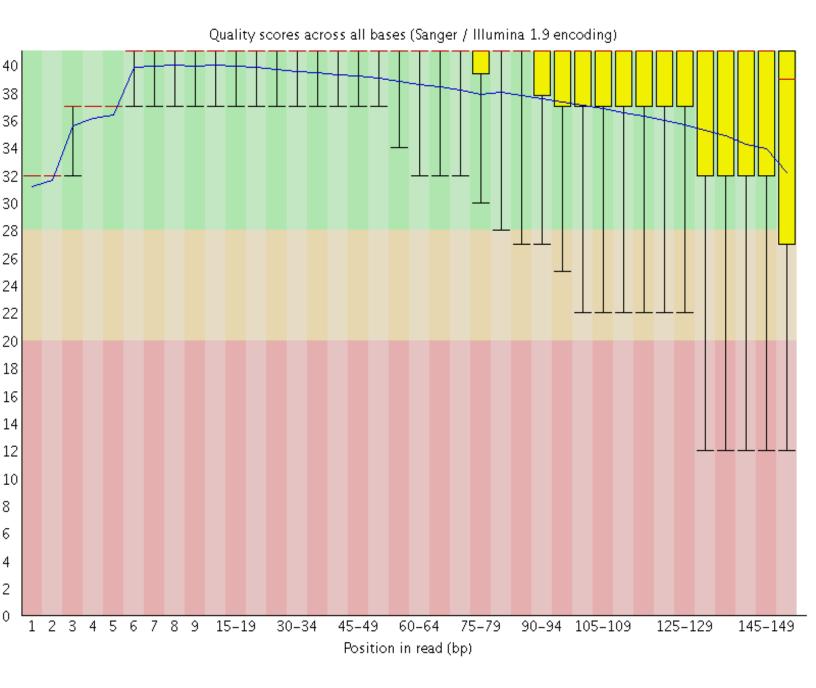
# **PastQC Report**

- 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



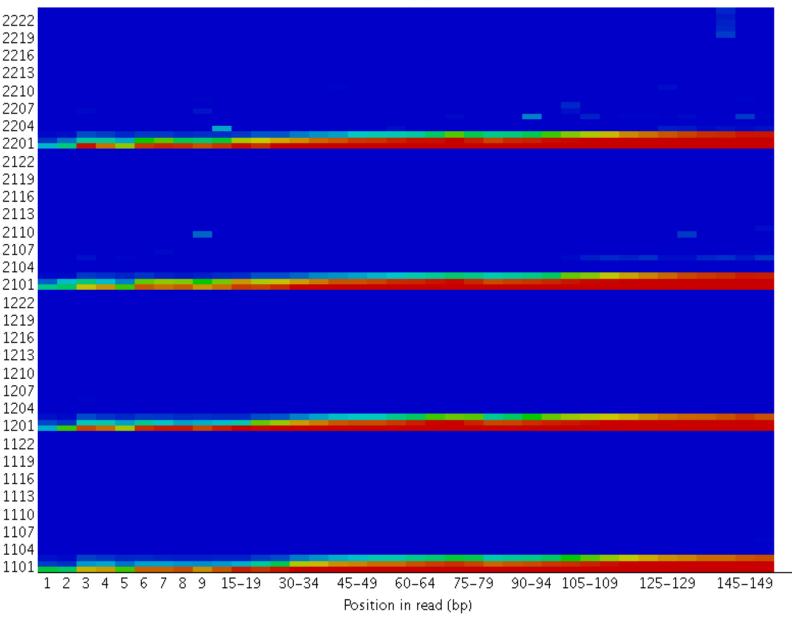
Measure	Value
Filename	<pre>Venter_S1_merge_R1.fastq.gz</pre>
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	789239544
Sequences flagged as poor quality	0
Sequence length	151
%GC	44





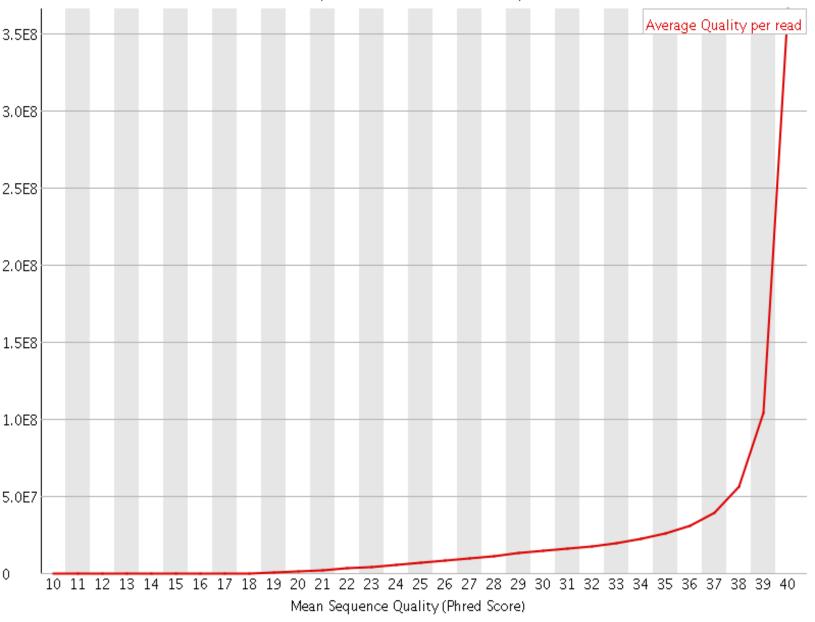


Quality per tile



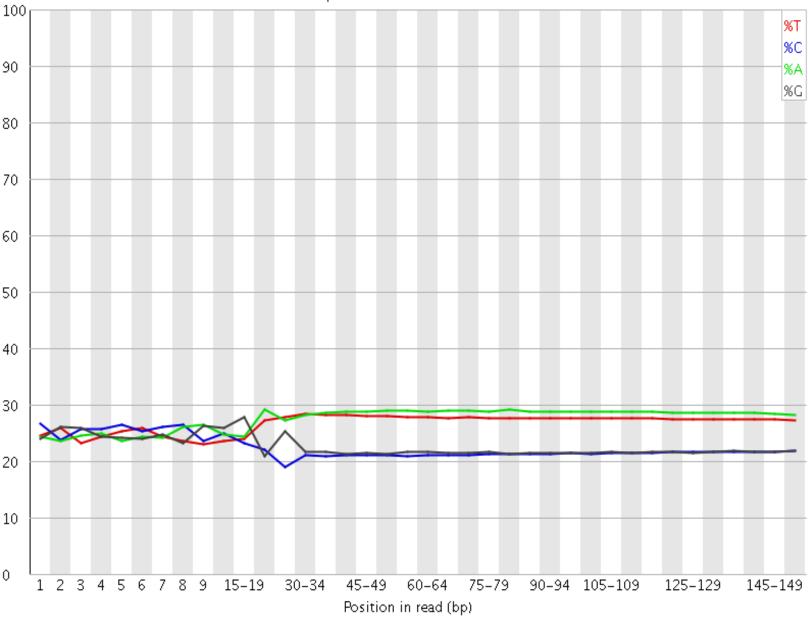


### Quality score distribution over all sequences



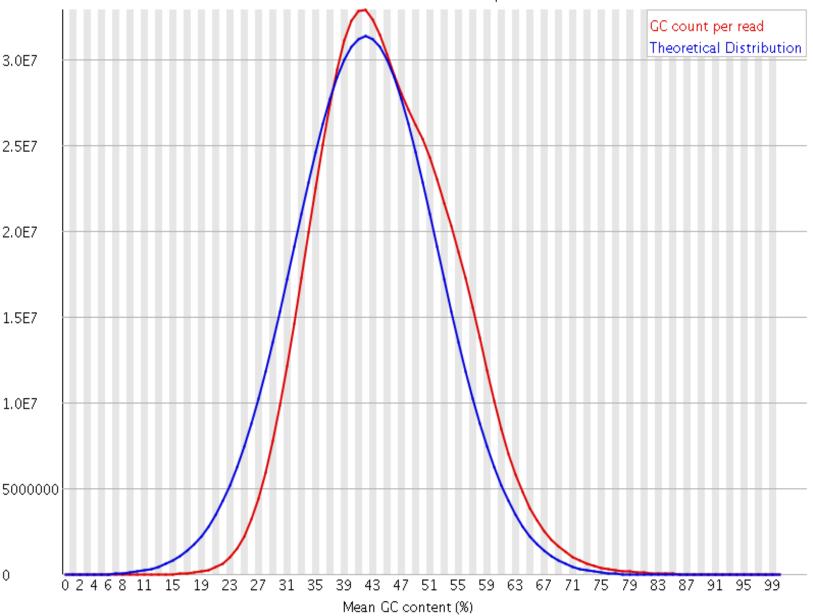


Sequence content across all bases



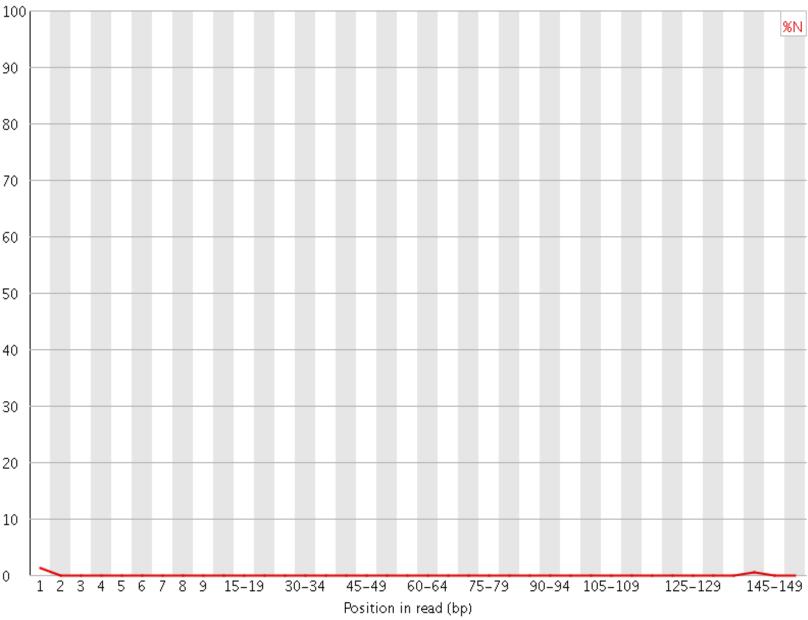


GC distribution over all sequences

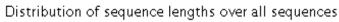


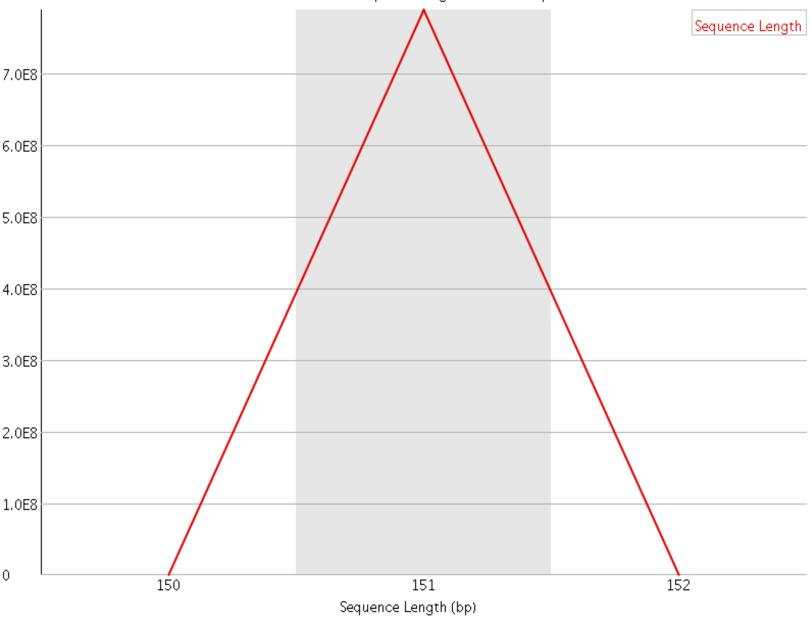


### N content across all bases

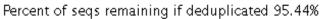


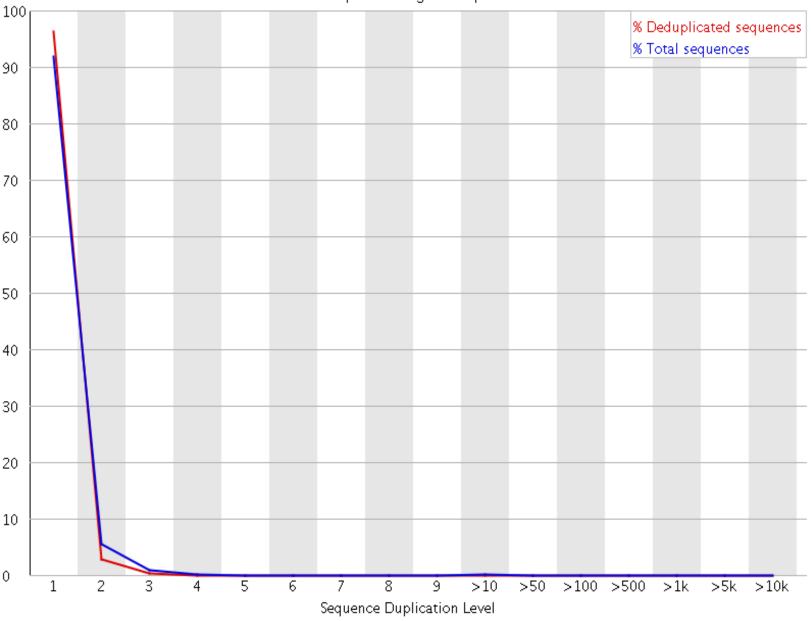








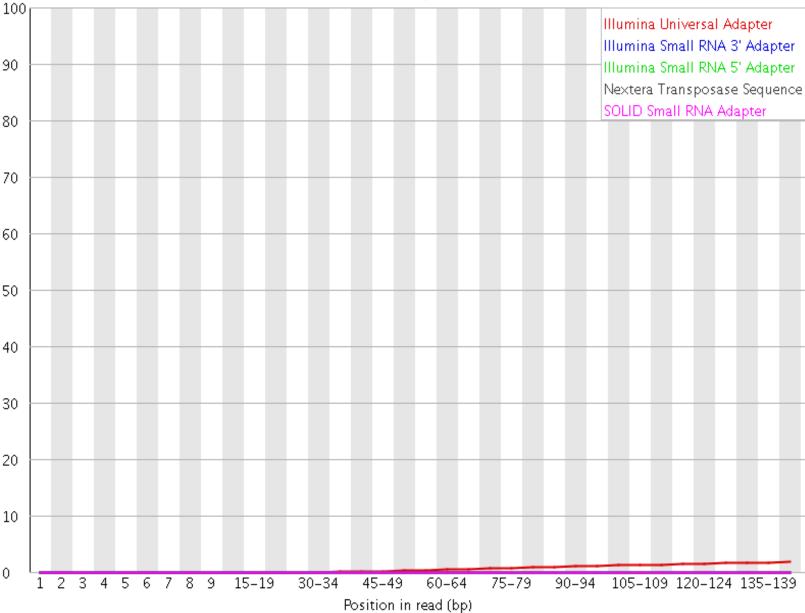




# Overrepresented sequences

No overrepresented sequences





Produced by FastQC (version 0.11.7)

0

### % Adapter

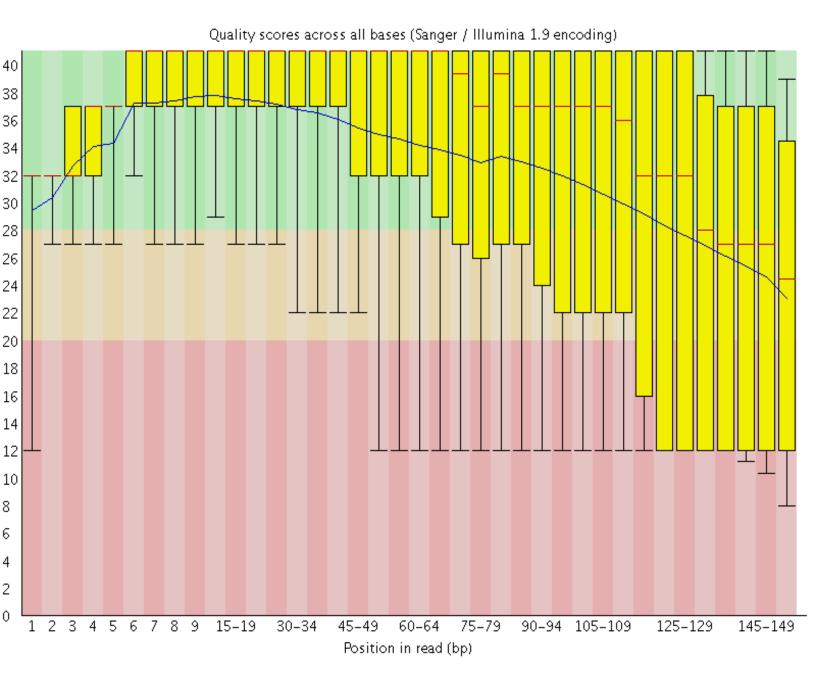
# **PastQC Report**

- 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

### Basic Statistics

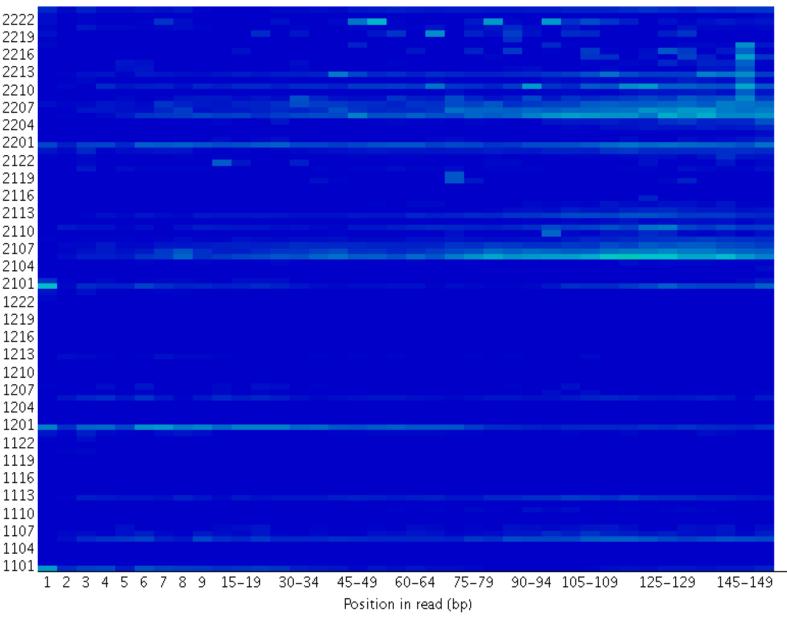
Measure	Value
Filename	Venter_S1_merge_R2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	789239544
Sequences flagged as poor quality	0
Sequence length	151
%GC	43





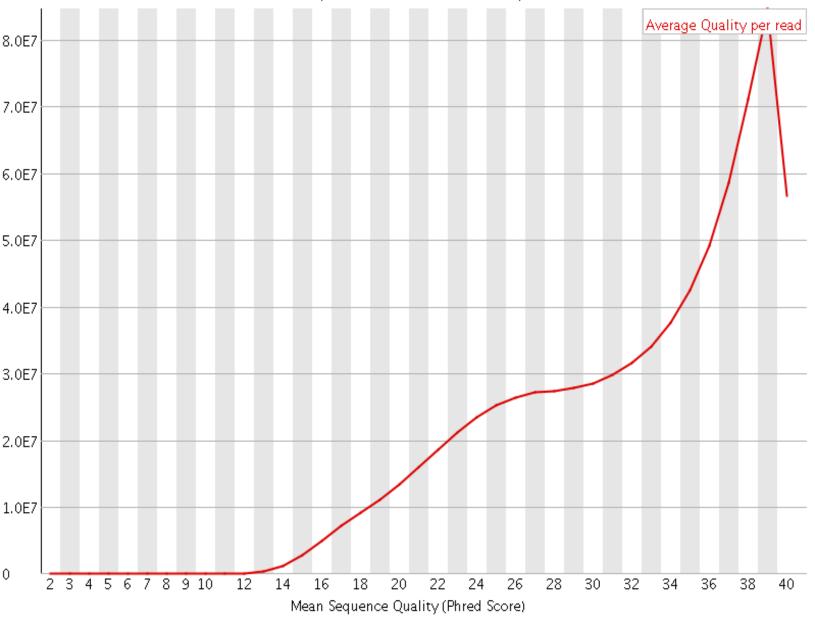


Quality per tile



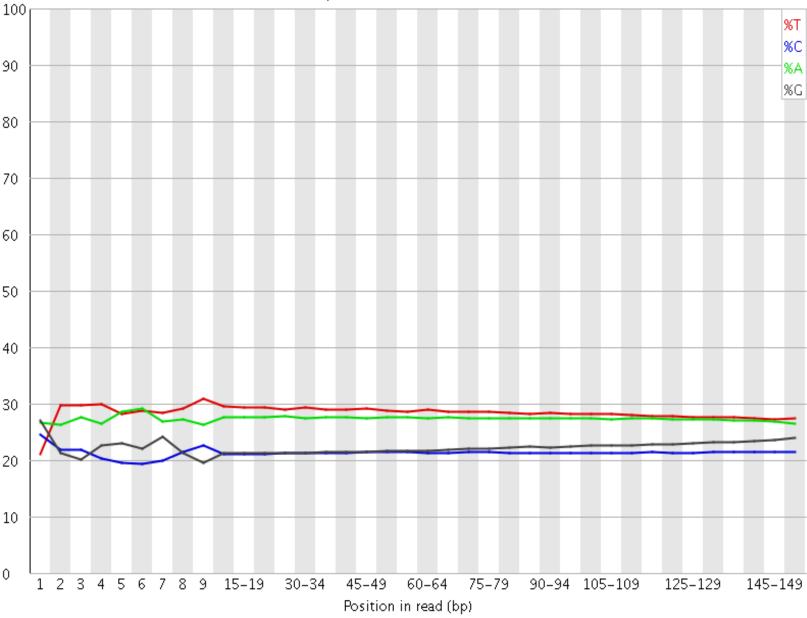


Quality score distribution over all sequences



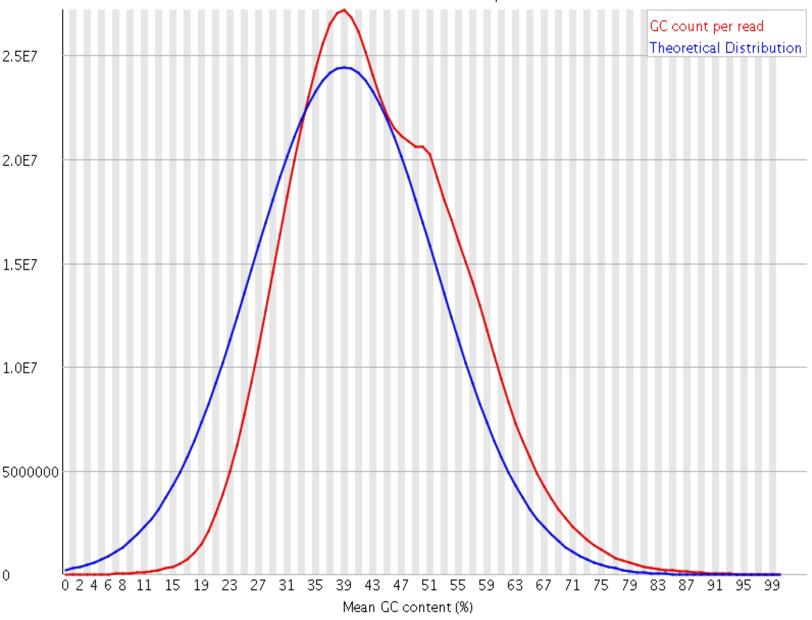


### Sequence content across all bases



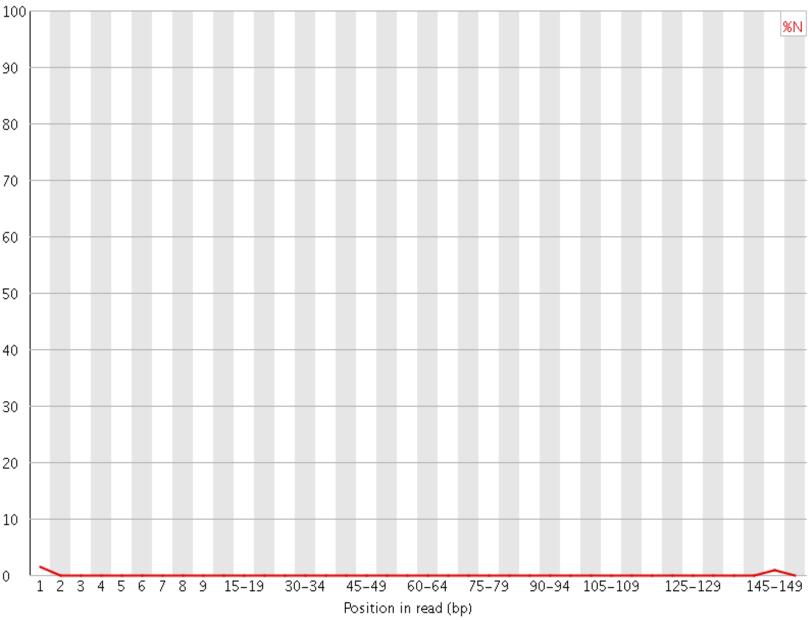


GC distribution over all sequences

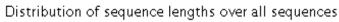


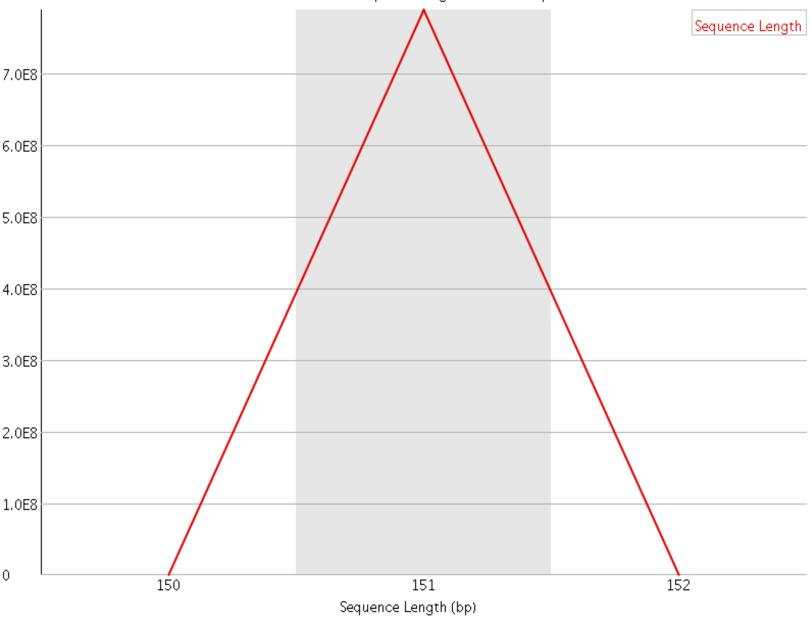


### N content across all bases

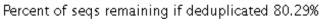


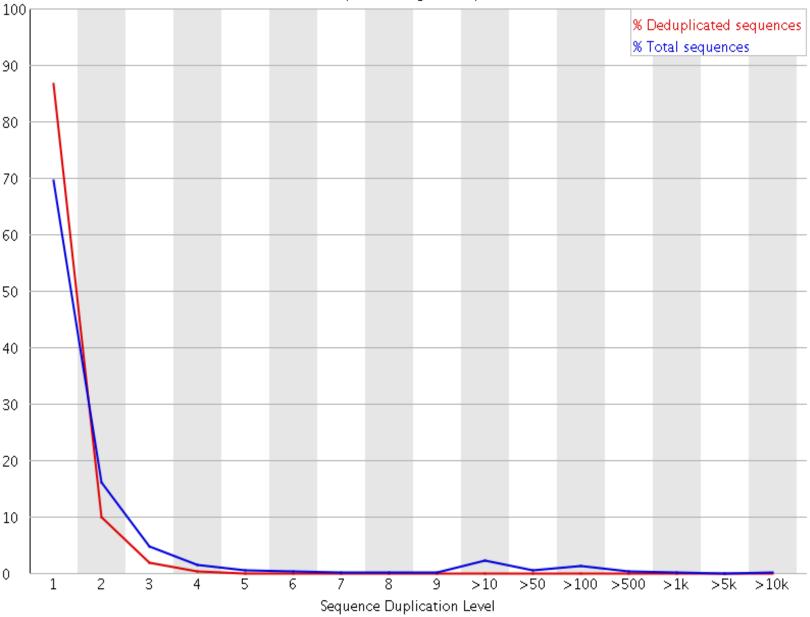








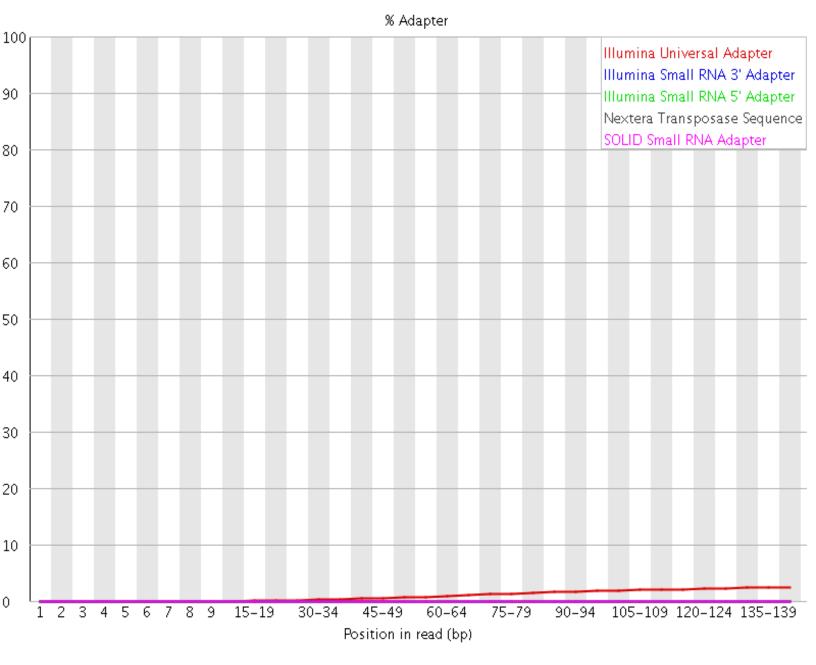




## Overrepresented sequences

No overrepresented sequences





Produced by FastQC (version 0.11.7)