Post-sequencing quality evaluation

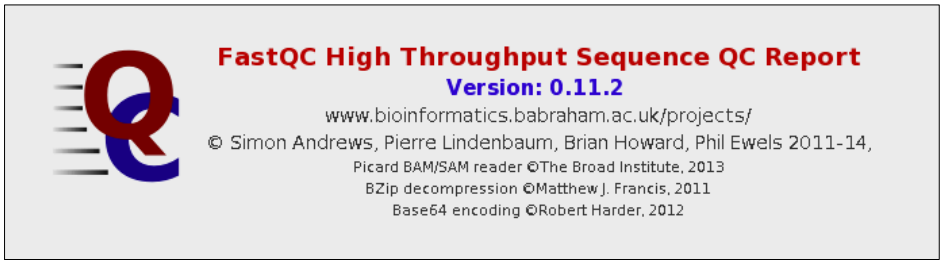
…or what to do when you get your reads from the sequencer

The fastq file contains information about sequence and quality

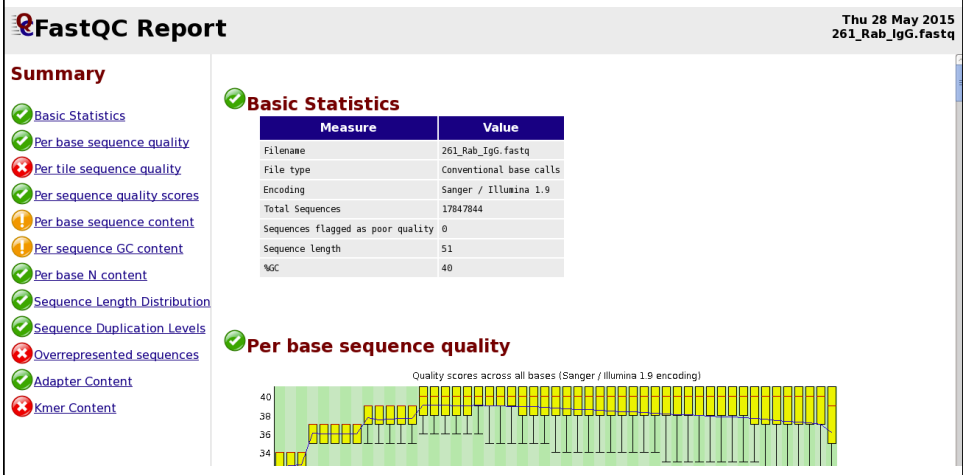
Sources of Library Read Quality Problems

- Sequencer problems
  - Read quality
- Library problems
  - GC content
  - Library complexity
  - Adaptor/primer contamination
  - Ribosomal RNA

Evaluating Quality



FastQC Report (html) basics



Assessing Sequencing Quality

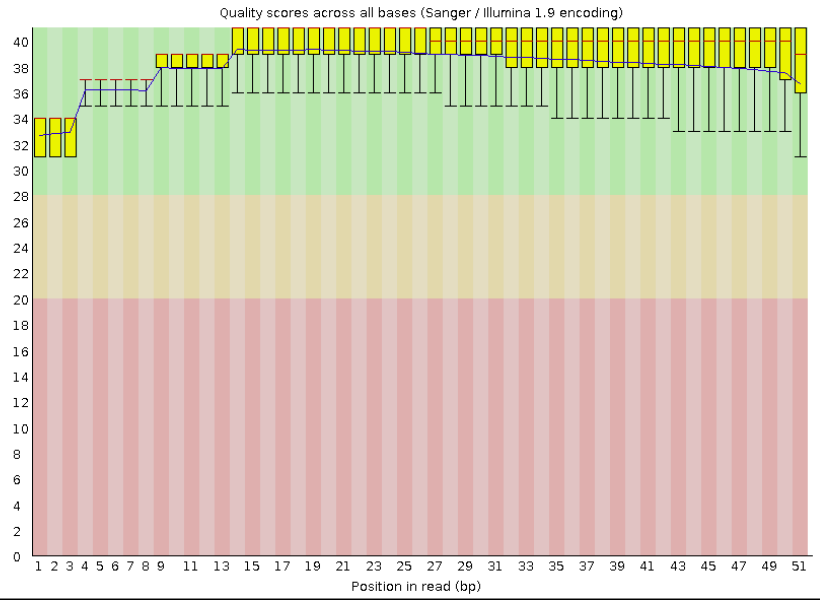
FastQC evaluates sequencer related quality in 3 different ways

- Per base sequence quality
  - Average quality for each base pair
- Per tile sequence quality
  - Average spatial quality on flow cell
- Per sequence quality score
  - Average quality per read

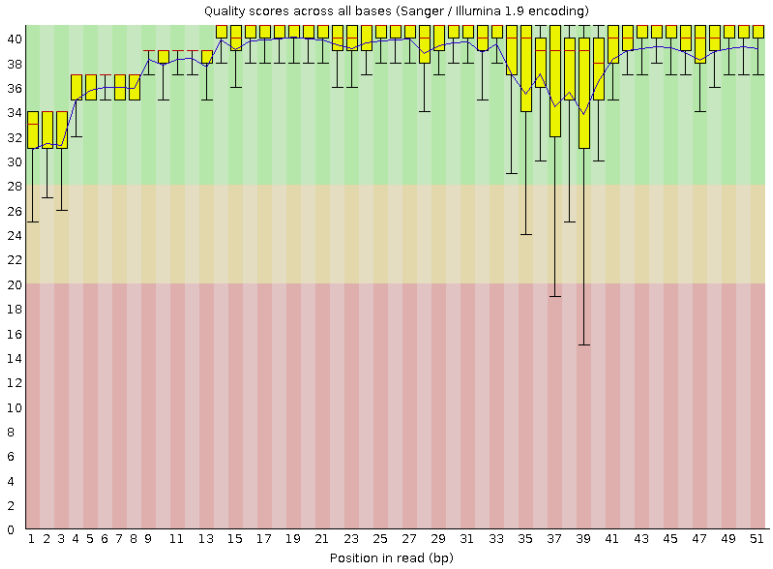
FastQC evaluates sequencer related quality in 3 different ways

- Per base sequence quality
  - Average quality for each base pair
- Per tile sequence quality
  - Average spatial quality on flow cell
- Per sequence quality score
  - Average quality per read

Per Base Sequence Quality



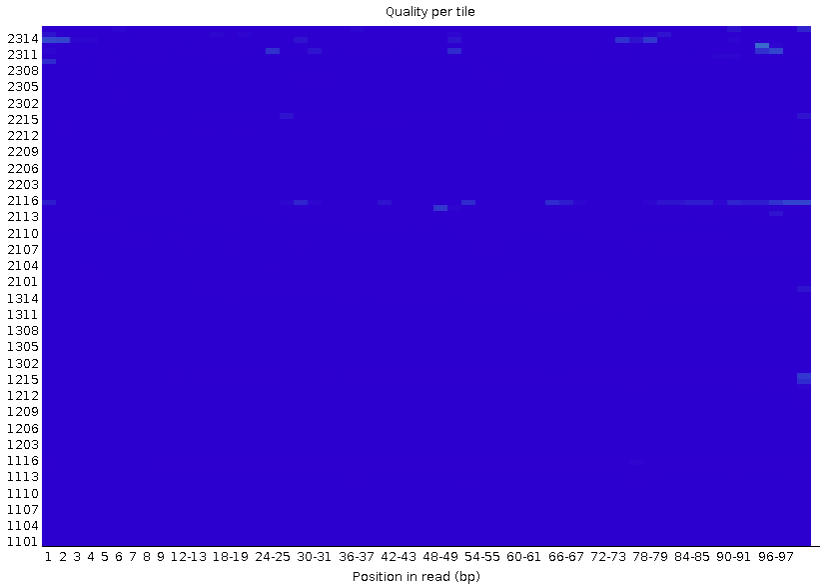
Mid-read drop in quality can effect mapping efficiency



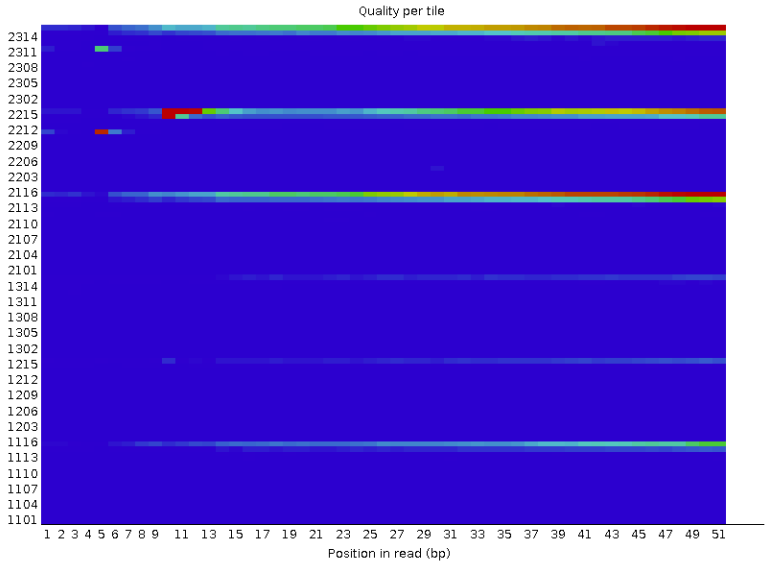
FastQC evaluates sequencer related quality in 3 different ways

- Per base sequence quality
  - Average quality for each base pair
- Per tile sequence quality
  - Average spatial quality on flow cell
- Per sequence quality score
  - Average quality per read

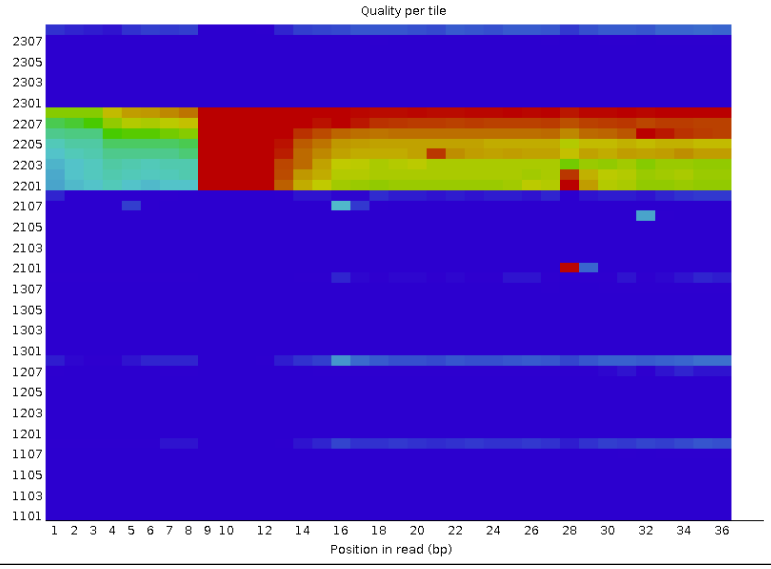
Per tile sequence quality



Small loss of tile quality



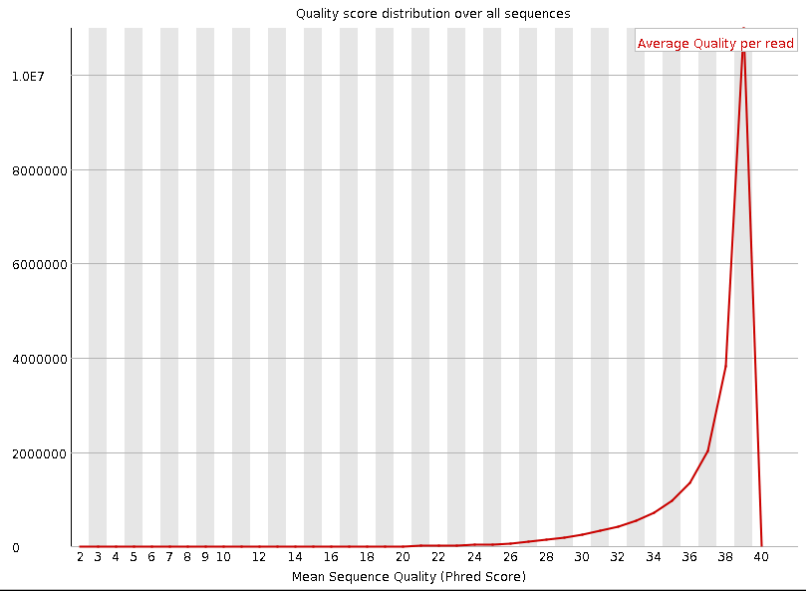
Large loss of flow cell quality



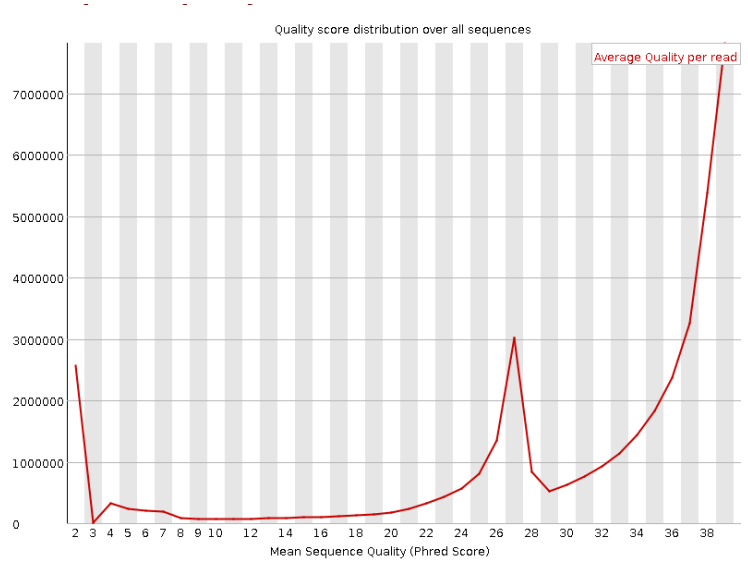
FastQC evaluates sequencer related quality in 3 different ways

- Per base sequence quality
  - Average quality for each base pair
- Per tile sequence quality
  - Average spatial quality on flow cell
- Per sequence quality score
  - Average quality per read

Average quality per read



Drop in quality for a portion of reads



Assessing Library Quality

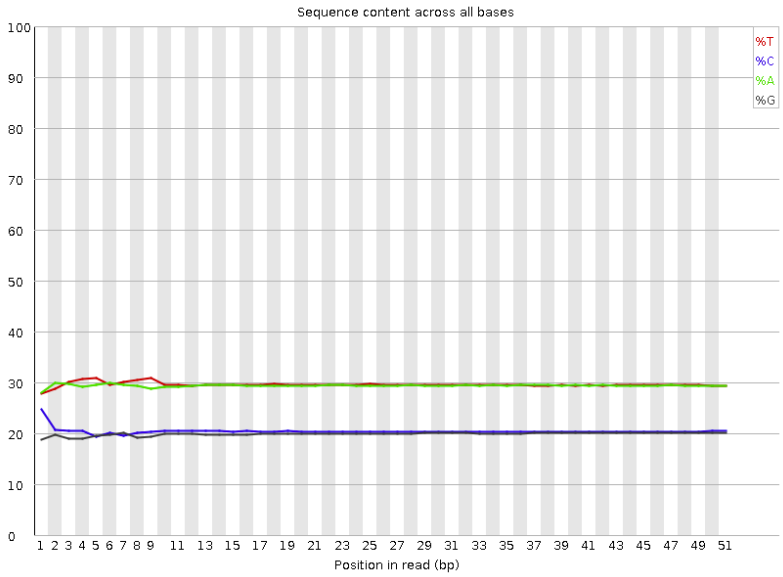
Nucleotide content of the reads

- Per base sequence content
  - % nucleotide representation at each bp
- Per sequence GC content
  - Distribution of % GC content per read
- Per base N content
  - % uncalled assigned nucleotides (N) per position
- Sequence length distribution

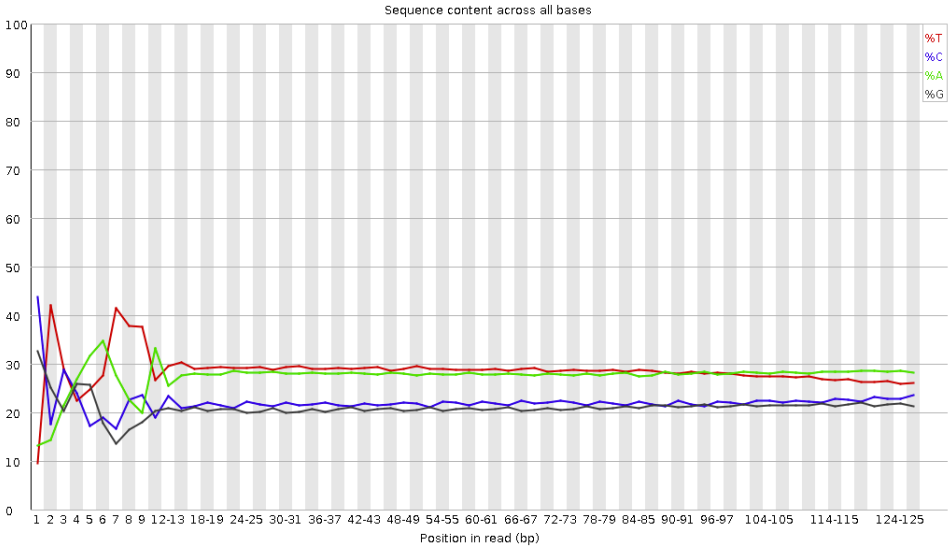
Nucleotide content of the reads

- Per base sequence content
  - % nucleotide representation at each bp
- Per sequence GC content
  - Distribution of % GC content per read
- Per base N content
  - % uncalled assigned nucleotides (N) per position

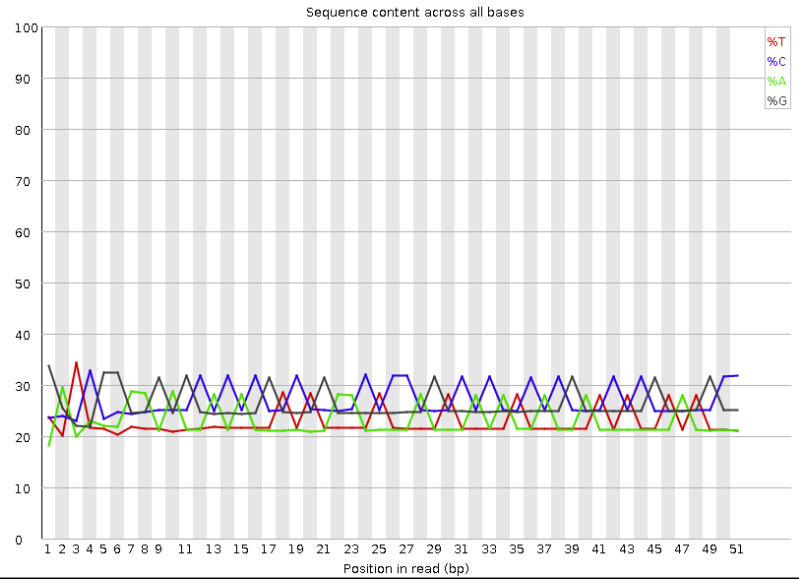
Per base sequence content



Random hexamer bias



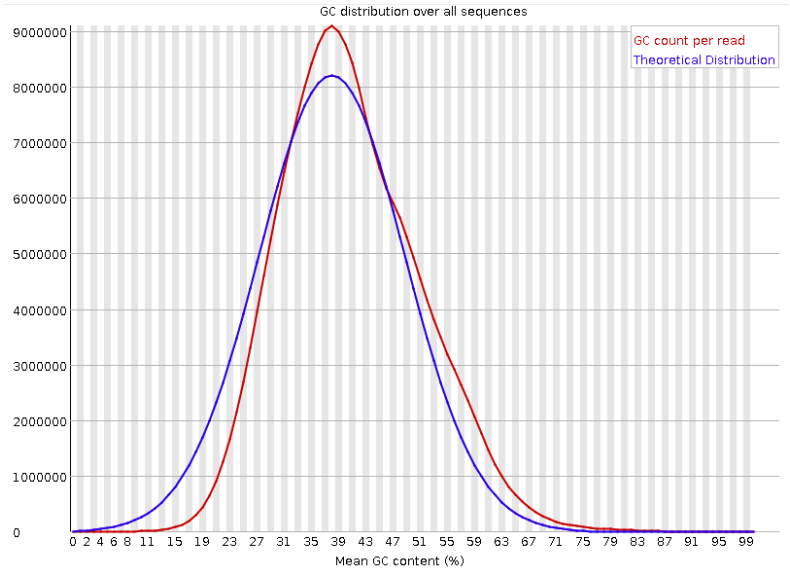
Adaptor/adaptor product



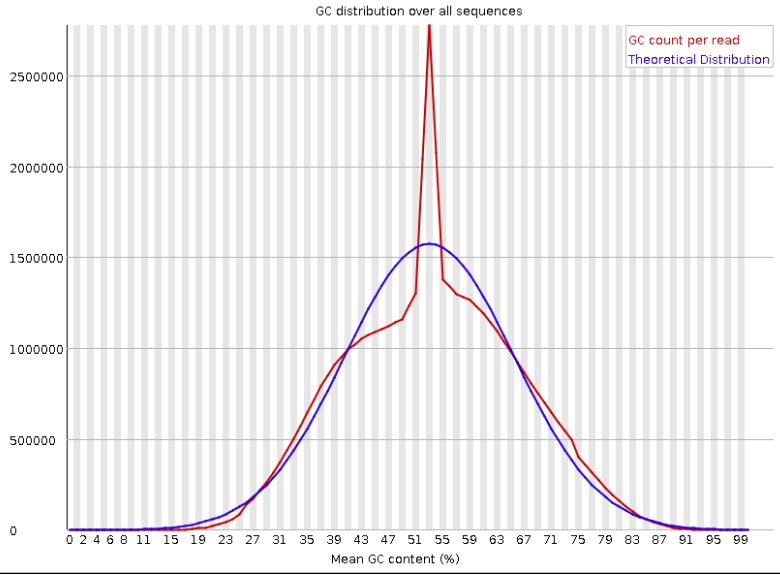
Nucleotide content of the reads

- Per base sequence content
  - % nucleotide representation at each bp
- Per sequence GC content
  - Distribution of % GC content per read
- Per base N content
  - % uncalled assigned nucleotides (N) per position
- Sequence length distribution

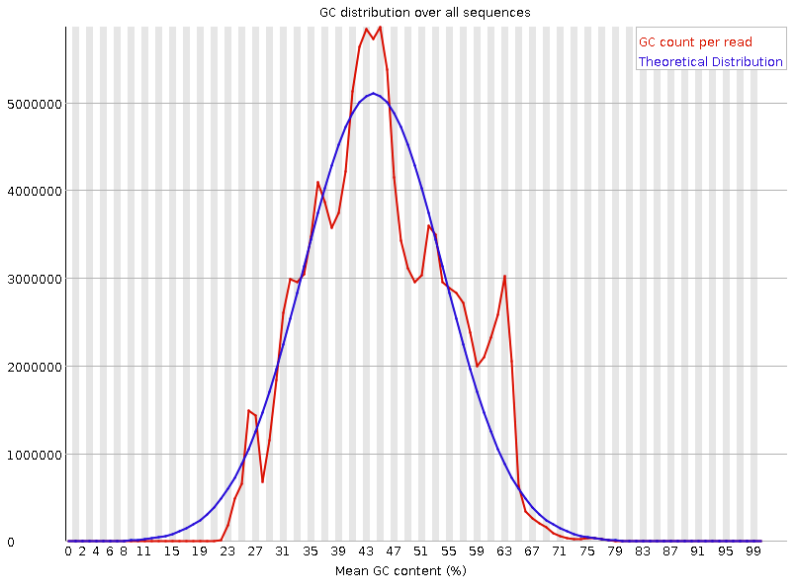
GC content per read



Adaptor/Adaptor product



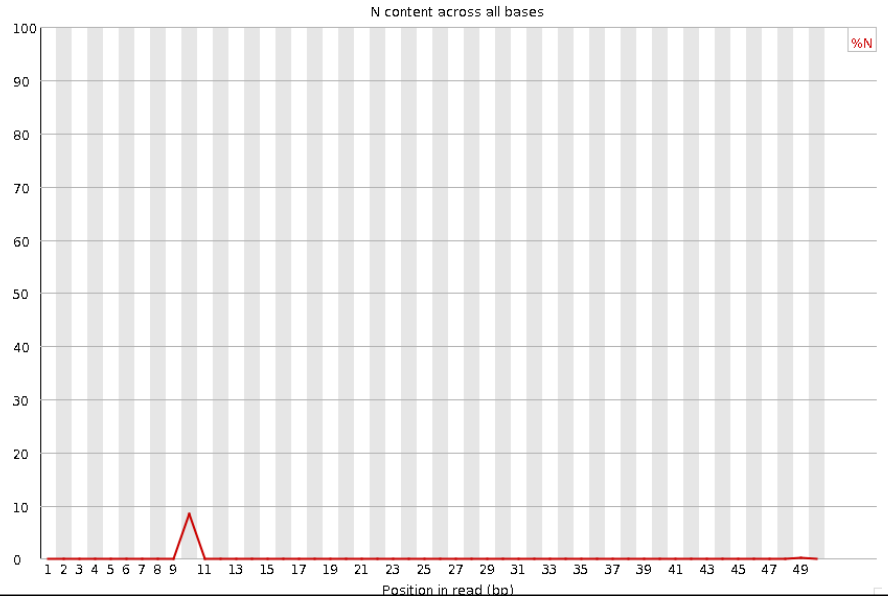
Biologically overrepresented sequence



Nucleotide content of the reads

- Per base sequence content
  - % nucleotide representation at each bp
- Per sequence GC content
  - Distribution of % GC content per read
- Per base N content
  - % uncalled assigned nucleotides (N) per position
- Sequence length distribution

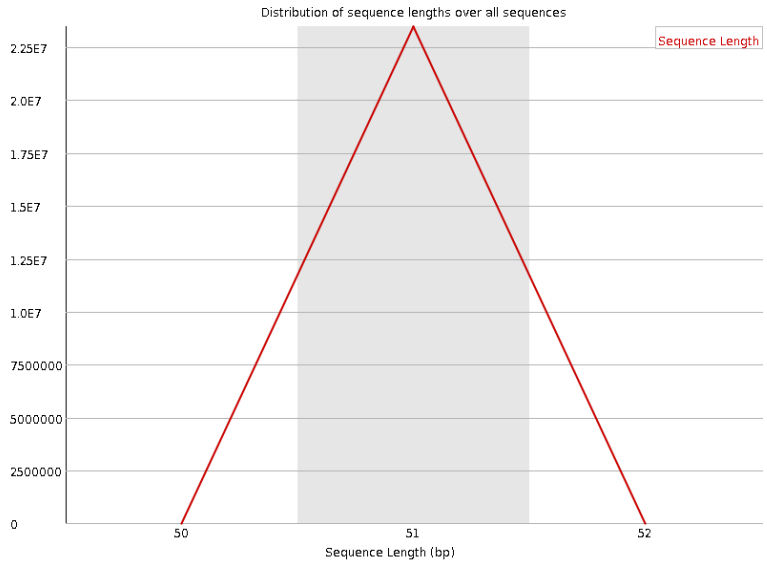
Per base N content



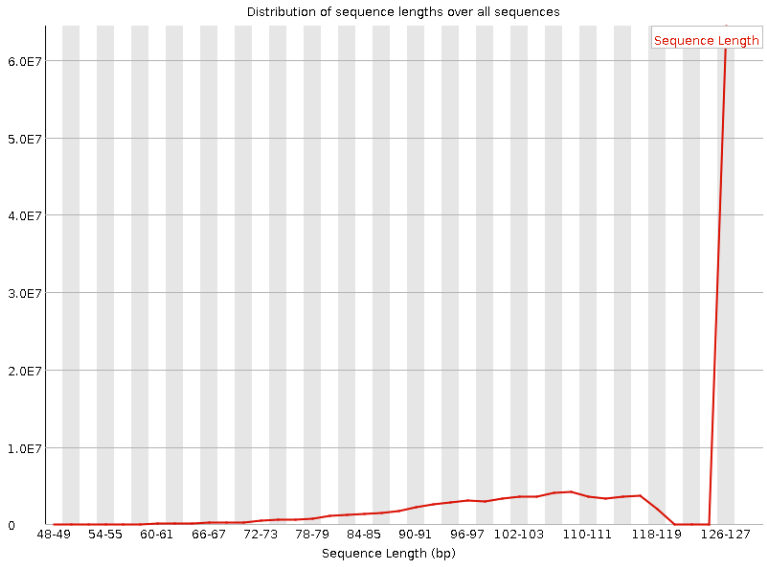
Nucleotide content of the reads

- Per base sequence content
  - % nucleotide representation at each bp
- Per sequence GC content
  - Distribution of % GC content per read
- Per base N content
  - % uncalled assigned nucleotides (N) per position
- Sequence length distribution

Sequence length distribution



Length post-trimming



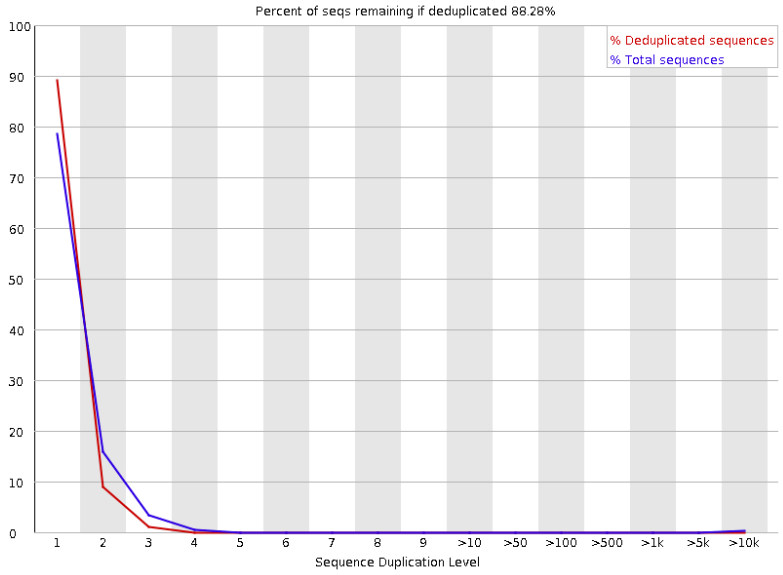
Library Content

- Sequence duplication levels
  - Total sequences vs. De-duplicated sequence
- Overrepresented sequences
  - Large polymer sequences
- Adaptor content
  - % adapter per nucleotide
- Kmer Content
  - Overrepresented 5-mers

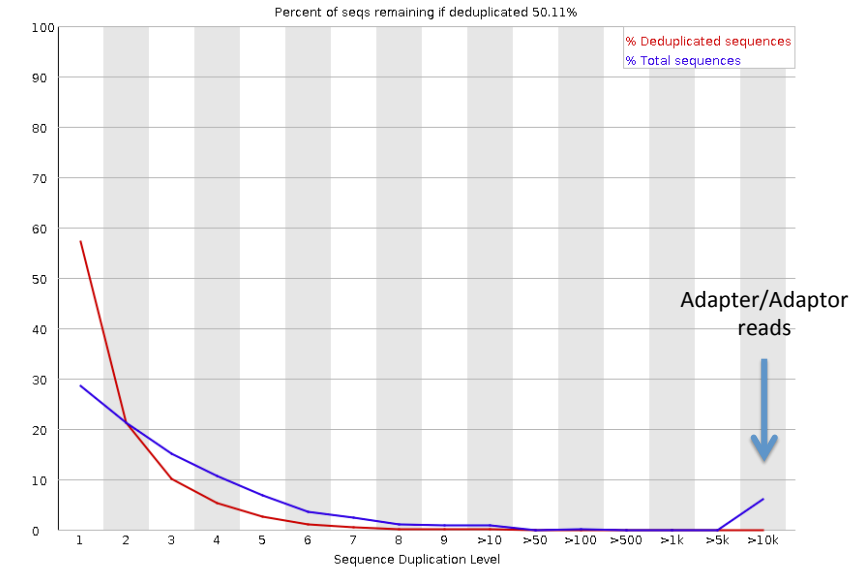
Library Content

- Sequence duplication levels
  - Total sequences vs. Deduplicated sequence
- Overrepresented sequences
  - Large polymer sequences
- Adaptor content
  - % adapter per nucleotide
- Kmer Content
  - Overrepresented 5-mers

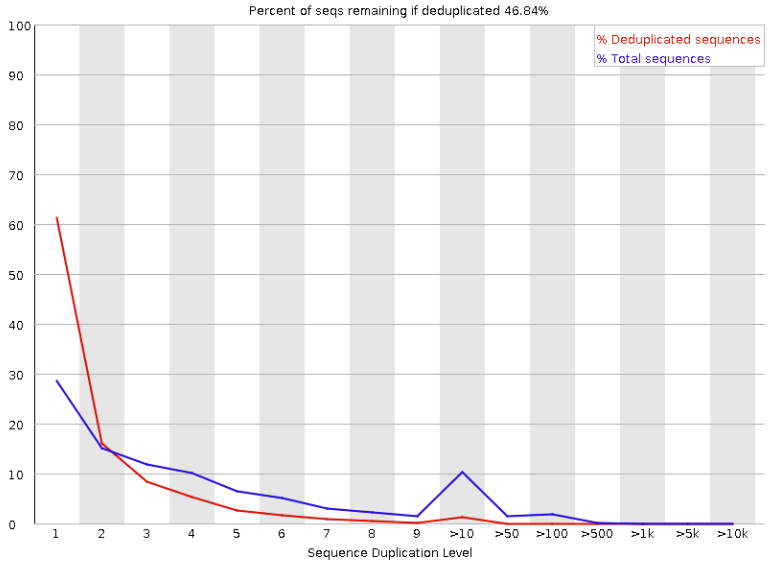
Sequence duplication levels



Low complexity



Biological sequence duplication



Library Content

- Sequence duplication levels
  - Total sequences vs. Deduplicated sequence
- Overrepresented sequences
  - Large polymer sequences
- Adaptor content
  - % adapter per nucleotide
- Kmer Content
  - Overrepresented 5-mers

Overrepresented Sequences

Ribosomal RNA

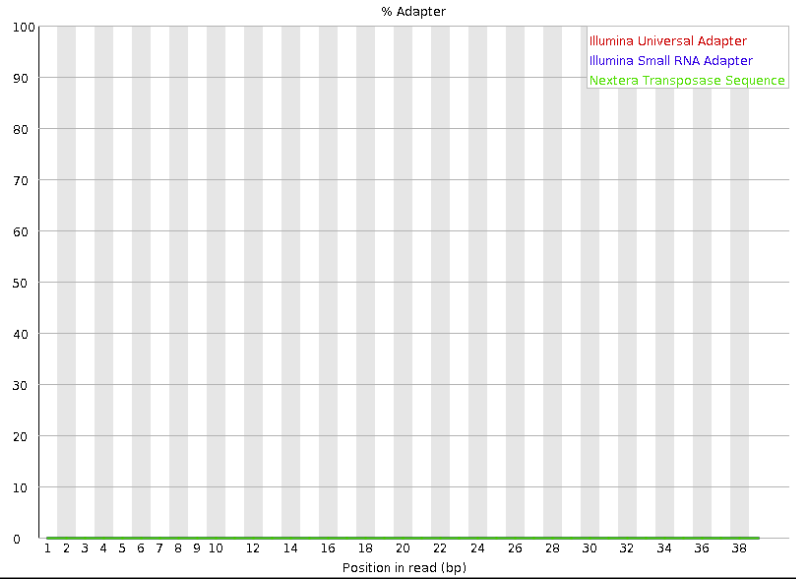
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATCA	693426	1.269663831294187	No Hit
CTCGCTATGTTGCCCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATC	644747	1.1805325243579463	No Hit
CAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATCAG	529701	0.9698831614337537	No Hit
GCTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGAT	464182	0.8499177944550657	No Hit
CTGGAGTCTTGGAAGCTTGACTACCCTACGTTCTCCTACAAATGGACCTT	428466	0.7845217559469866	No Hit
CTCAGACCGCGTTCTCTCCCTCTCACTCCCCAATACGGAGAGAAGAACGA	404432	0.7405154733424628	No Hit
CTCGCTATGTTGCTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATC	399379	0.7312634243285384	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGC	364352	0.667128945640486	TruSeq Adapter, Index 5 (100% over 50bp)
GCCCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGAT	339293	0.6212458868160389	No Hit
GGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATCAGCA	337332	0.6176552934821172	No Hit
CCACAAATTATGCAGTCGAGTTTCCCACATTTGGGGAAATCGCAGGGGTC	305105	0.5586476181265383	No Hit
CGGGGTCTCGCTATGTTGCCCAGGCTGGAGTGCAGTGGCTATTCACAGGC	303105	0.5549856157462001	No Hit

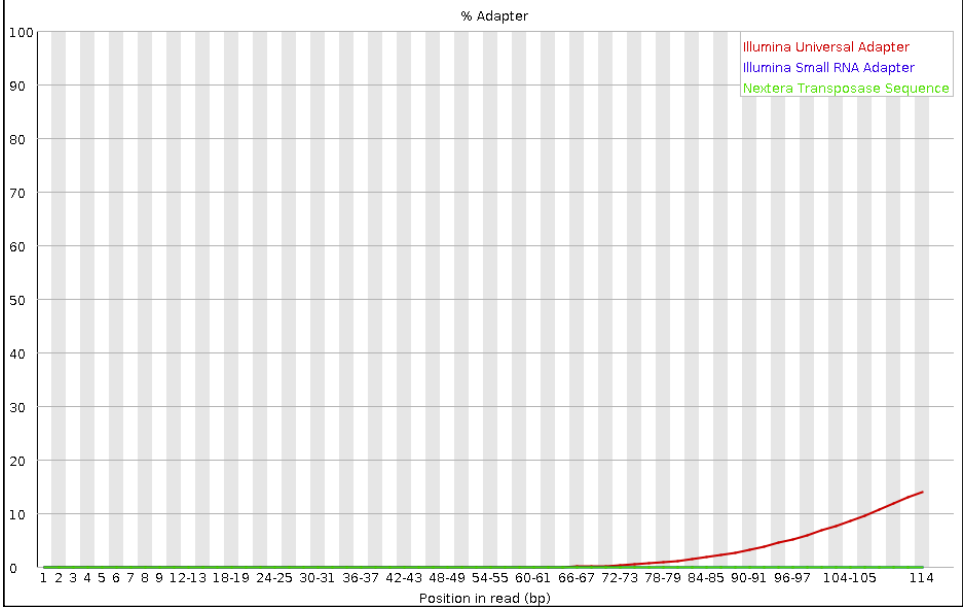
Library Content

- Sequence duplication levels
  - Total sequences vs. Deduplicated sequence
- Overrepresented sequences
  - Large polymer sequences
- Adaptor content
  - % adapter per nucleotide
- Kmer Content
  - Overrepresented 5-mers

Adaptor Content



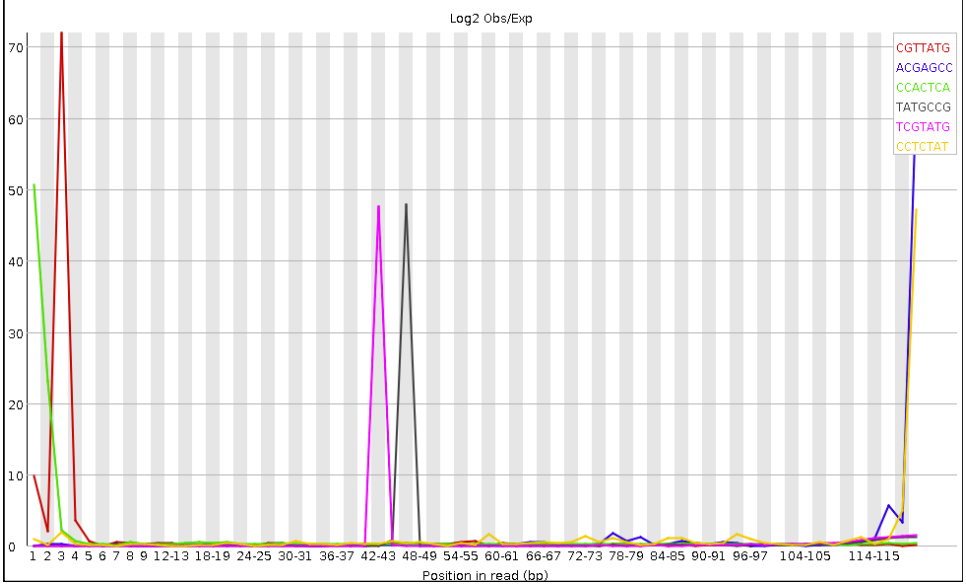
Too-small library fragments



Library Content

- Sequence duplication levels
  - Total sequences vs. Deduplicated sequence
- Overrepresented sequences
  - Large polymer sequences
- Adaptor content
  - % adapter per nucleotide
- Kmer Content
  - Overrepresented 5-mers

K-mer (5-mer) content



Don't worry be happy!!

Just because your library doesn't look "perfect" doesn't mean it is BAD.

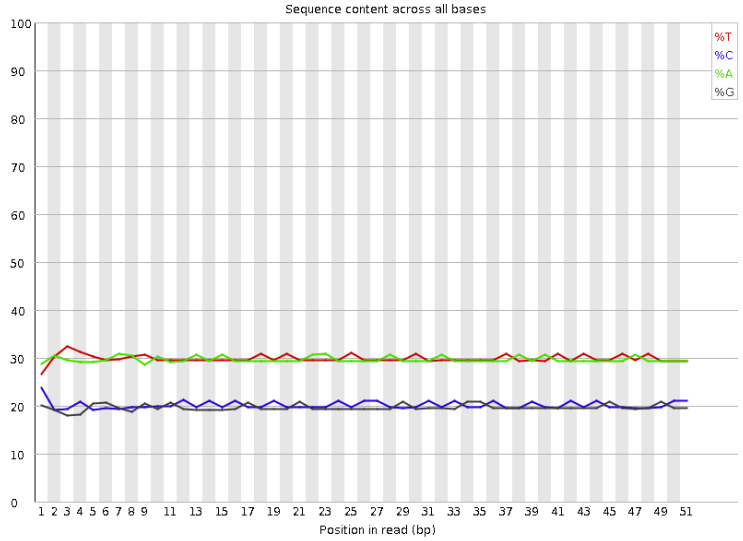
- Trim reads
  - Low quality or adapter reads
- Remove duplicates
  - ONLY IF NECESSARY
- Mapping takes into account base quality
- Get more coverage

Ultimately it is a judgement call!!!

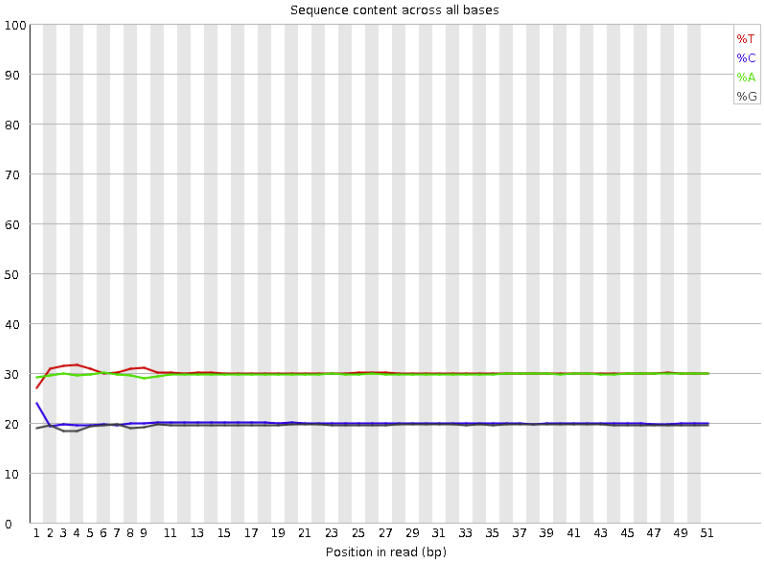
Trimming reads

- You can trim reads to remove adapters and low quality sequence
- The short read workshop has a video on this process.
- Here is a preview of how the library can change

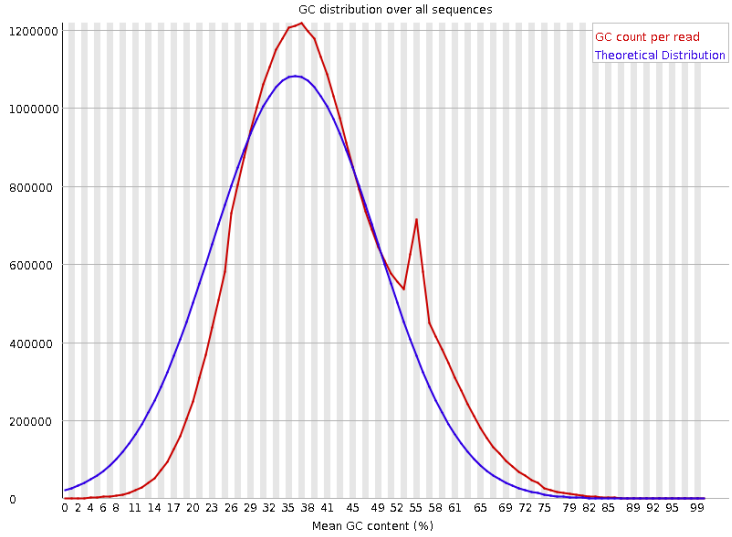
Per base sequence content- before trimming



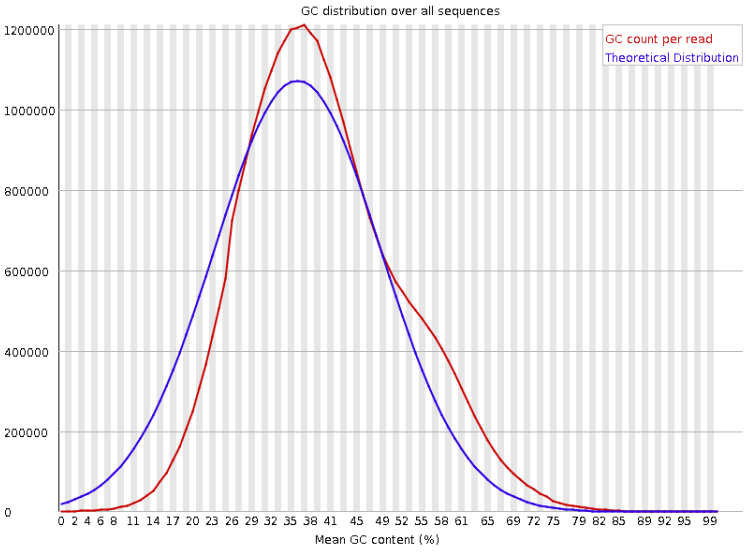
Per base sequence content- after trimming



Per sequence GC content- Before trimming



Per sequence GC content- after trimming



Overrepresented sequences

Before trimming

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCC	221948	1.2435563645670 59	TruSeq Adapter, Index 11 (100% over 51bp)

After trimming

Overrepresented sequences

No overrepresented sequences