Bioinformatics Analysis Tools for NGS Data

Quality control

Philipp Rescheneder, Moritz Smolka

April 27, 2016

Center for Integrative Bioinformatics Vienna
Max F. Perutz Laboratories

CIBIV
Center for Integrative Bioinformatics Vienna

Quality Control

- First step of any data analysis
- Always look at your data!
- Short sanity checks after each step

FastQC

- A quality control tool for high throughput sequence data.

FastQC

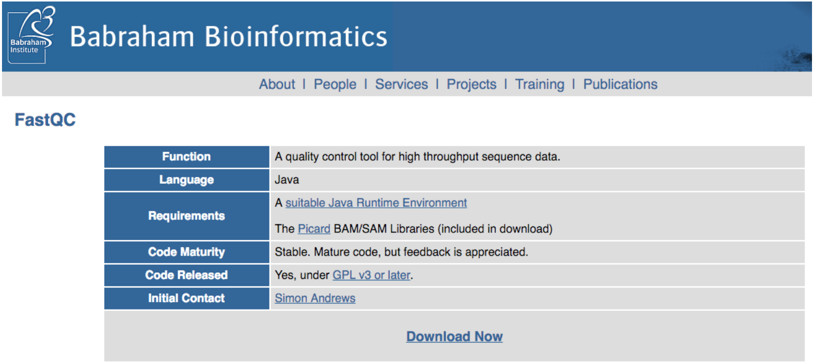
- A quality control tool for high throughput sequence data.
- Platform independent

FastQC

- A quality control tool for high throughput sequence data.
- Platform independent
- Interface: Command line or GUI

FastQC

- A quality control tool for high throughput sequence data.
- Platform independent
- Interface: Command line or GUI



Babraham Bioinformatics

Babraham Institute
About | People | Services | Projects | Training | Publications

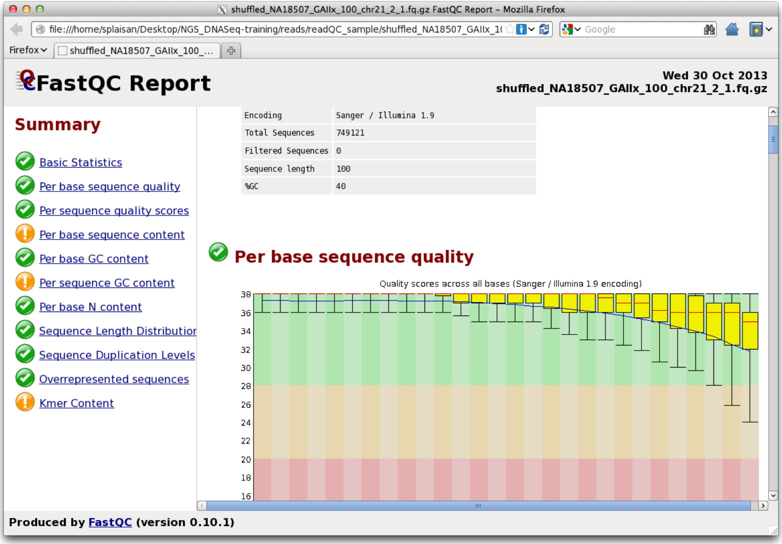
FastQC

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

Download Now

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

FastQC



shuffled_NA18507_GAIIx_100_chr21_2_1.fq.gz FastQC Report - Mozilla Firefox

file:///home/splaisan/Desktop/NGS_DNASeq-training/reads/readQC_sample/shuffled_NA18507_GAIIx_1(

FastQC Report

Wed 30 Oct 2013
shuffled_NA18507_GAIIx_100_chr21_2_1.fq.gz

Summary

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

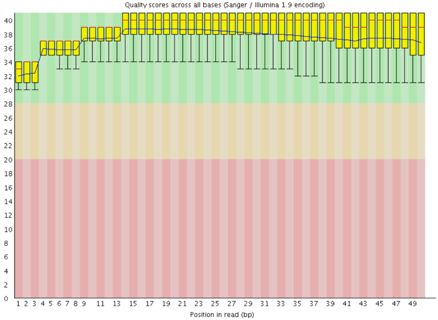
Encoding	Sanger / Illumina 1.9
Total Sequences	749121
Filtered Sequences	0
Sequence length	100
%GC	40

✓ Per base sequence quality

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

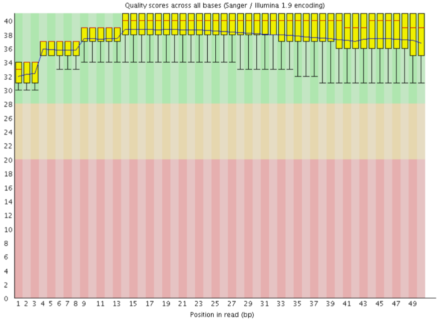
Produced by FastQC (version 0.10.1)

FastQC: QV distribution



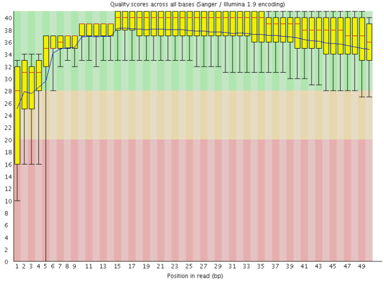
- Overview of the range of quality values across all bases at each position of the reads

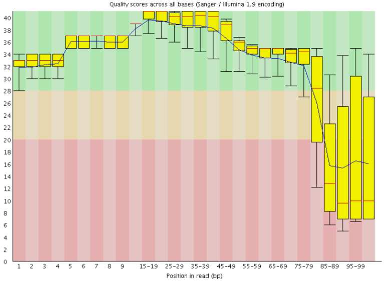
FastQC: QV distribution



- Overview of the range of quality values across all bases at each position of the reads
  - The central red line is the median value
  - The yellow box represents the inter-quartile range (25-75
  - The upper and lower whiskers represent the 10
  - The blue line represents the mean quality

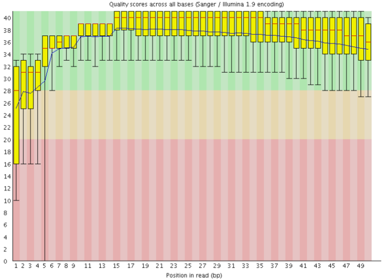
FastQC: QV distribution

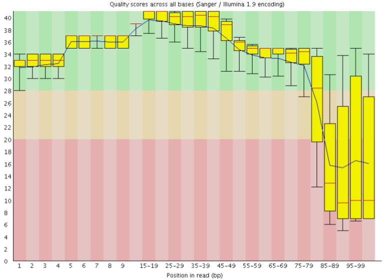




- Check for low qualities at 5’ or 3’ end

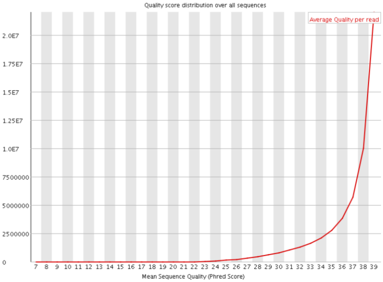
FastQC: QV distribution





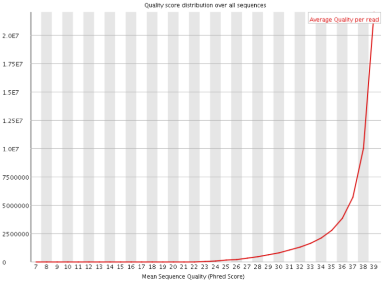
- Check for low qualities at 5’ or 3’ end
- Low quality ends can be removed using tools like fastx, cutadapt, trimmomatic (see adapter removal)

FastQC: Mean QV distribution



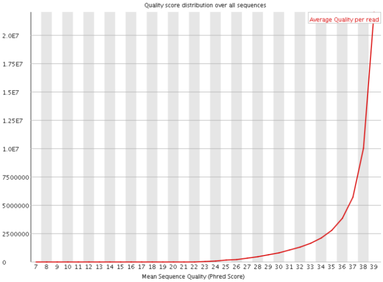
- Mean base quality per read

FastQC: Mean QV distribution



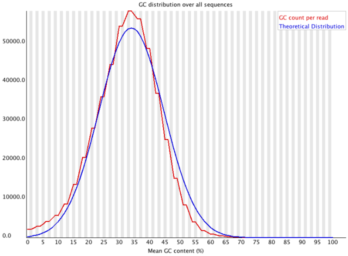
- Mean base quality per read
- The higher the better

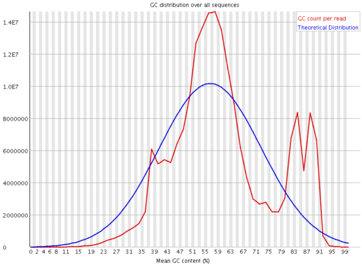
FastQC: Mean QV distribution



- Mean base quality per read
- The higher the better
- Errors here usually indicate a general loss of quality within a run

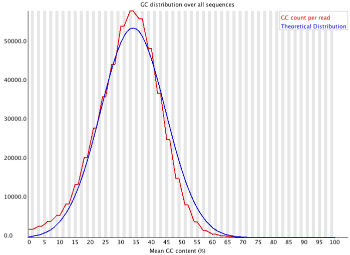
FastQC: GC distribution

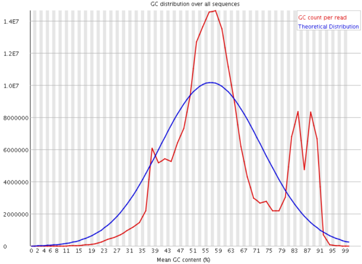




- Measures the GC content across the whole length of each read

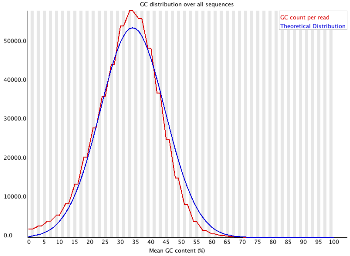
FastQC: GC distribution

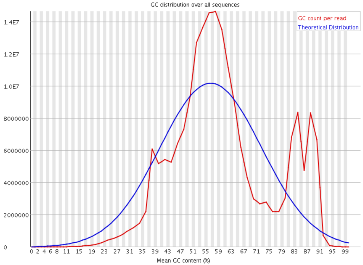




- Measures the GC content across the whole length of each read
- In normal library: normal distribution with peak that corresponds to the overall GC of the sequenced genome

FastQC: GC distribution





- Measures the GC content across the whole length of each read
- In normal library: normal distribution with peak that corresponds to the overall GC of the sequenced genome
- Sharp additional peaks results of specific contaminant (adapter dimers), broader peaks may represent contamination with different species

Contamination

- Most common contamination: human DNA

Contamination

- Most common contamination: human DNA
- DNA from other organisms in the lab

Contamination

- Most common contamination: human DNA
- DNA from other organisms in the lab
- RNA-Seq: rRNA contamination (usually depleted)

Contamination

- Most common contamination: human DNA
- DNA from other organisms in the lab
- RNA-Seq: rRNA contamination (usually depleted)
- Primer/Adapter dimers (probably caused by low amount of DNA/RNA)
- (Significantly) reduces the number of reads

Contamination

- Most common contamination: human DNA
- DNA from other organisms in the lab
- RNA-Seq: rRNA contamination (usually depleted)
- Primer/Adapter dimers (probably caused by low amount of DNA/RNA)
- (Significantly) reduces the number of reads
- Might cause problems during down stream analysis
    - Cause false positives during SNP calling
    - Severely alter expression values for RNA-Seq
    - Reduce the number of ”usable” reads to a point were only noise is picked up by downstream analysis

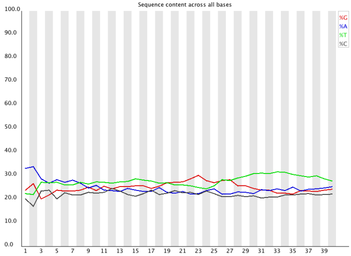
Contamination

- Most common contamination: human DNA
- DNA from other organisms in the lab
- RNA-Seq: rRNA contamination (usually depleted)
- Primer/Adapter dimers (probably caused by low amount of DNA/RNA)
- (Significantly) reduces the number of reads
- Might cause problems during down stream analysis
  - Cause false positives during SNP calling
  - Severely alter expression values for RNA-Seq
  - Reduce the number of ”usable” reads to a point were only noise is picked up by downstream analysis
- Strategy to identify contaminations:
  - Blast random subset of (unmapped) reads
  - Tools: Kraken (https://ccb.jhu.edu/software/kraken/)

Contamination

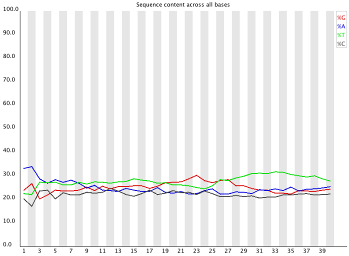
- Most common contamination: human DNA
- DNA from other organisms in the lab
- RNA-Seq: rRNA contamination (usually depleted)
- Primer/Adapter dimers (probably caused by low amount of DNA/RNA)
- (Significantly) reduces the number of reads
- Might cause problems during down stream analysis
  - Cause false positives during SNP calling
  - Severely alter expression values for RNA-Seq
  - Reduce the number of ”usable” reads to a point were only noise is picked up by downstream analysis
- Strategy to identify contaminations:
  - Blast random subset of (unmapped) reads
  - Tools: Kraken (https://ccb.jhu.edu/software/kraken/)
- If contaminating organisms are known, add to downstream analysis (e.g. remove while mapping)

FastQC: Per base sequence content



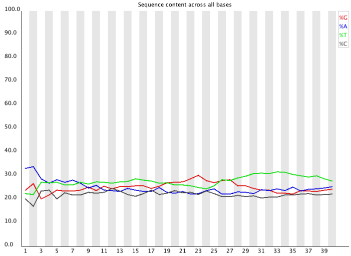
- Proportion of bases (A, T, G, C) per read position

FastQC: Per base sequence content



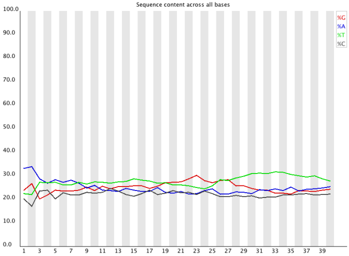
- Proportion of bases (A, T, G, C) per read position
- In random library, little difference between bases expected

FastQC: Per base sequence content



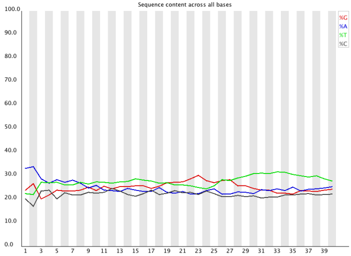
- Proportion of bases (A, T, G, C) per read position
- In random library, little difference between bases expected
- Causes:
  - Overrepresented sequences (e.g. adapter dimers, rRNA)

FastQC: Per base sequence content



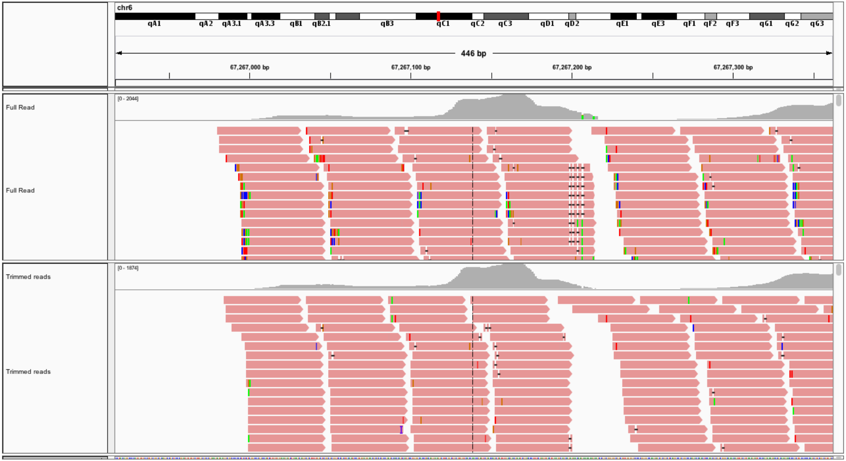
- Proportion of bases (A, T, G, C) per read position
- In random library, little difference between bases expected
- Causes:
  - Overrepresented sequences (e.g. adapter dimers, rRNA)
  - (Biased) random priming causes difference in first 12 bp (RNA-Seq)

FastQC: Per base sequence content

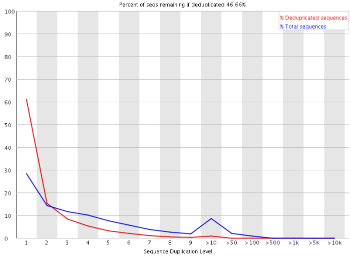


- Proportion of bases (A, T, G, C) per read position
- In random library, little difference between bases expected
- Causes:
  - Overrepresented sequences (e.g. adapter dimers, rRNA)
  - (Biased) random priming causes difference in first 12 bp (RNA-Seq)
  - Adapter trimming can cause differences at the end of the reads

Example: Per base sequence content

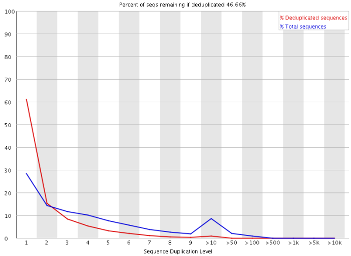


FastQC: Duplicates



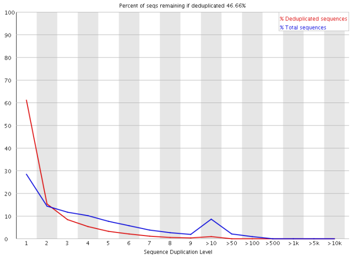
- PCR artifacts vs. biological duplicates

FastQC: Duplicates



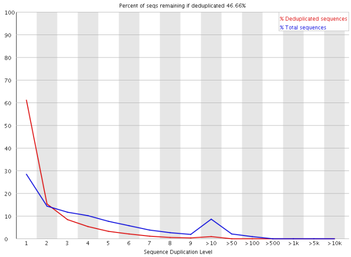
- PCR artifacts vs. biological duplicates
- Most sequences should fall into far left bins in red and blue line

FastQC: Duplicates



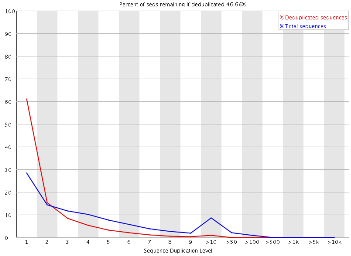
- PCR artifacts vs. biological duplicates
- Most sequences should fall into far left bins in red and blue line

FastQC: Duplicates



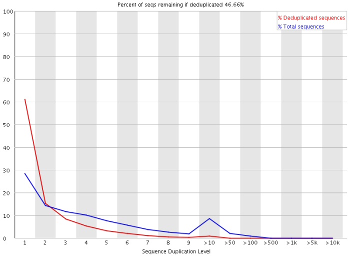
- PCR artifacts vs. biological duplicates
- Most sequences should fall into far left bins in red and blue line
- Datasets with high coverage will flatten the lines

FastQC: Duplicates



- PCR artifacts vs. biological duplicates
- Most sequences should fall into far left bins in red and blue line
- Datasets with high coverage will flatten the lines
- Contaminations tend to produce spikes toward the right of the plot (read line)

FastQC: Duplicates



- PCR artifacts vs. biological duplicates
- Most sequences should fall into far left bins in red and blue line
- Datasets with high coverage will flatten the lines
- Contaminations tend to produce spikes toward the right of the plot (read line)
- RNA-Seq: highly abundant transcripts might cause peaks in the higher duplication bins

Duplicates removal

- Can be done on raw reads using fastx (reduces mapping runtime)

Duplicates removal

- Can be done on raw reads using fastx (reduces mapping runtime)
- Hard to decided whether duplicates were caused by PCR or biological/experimental reasons

Duplicates removal

- Can be done on raw reads using fastx (reduces mapping runtime)
- Hard to decided whether duplicates were caused by PCR or biological/experimental reasons
- Mapping all reads first and removing duplicates after manual inspection is more sensible

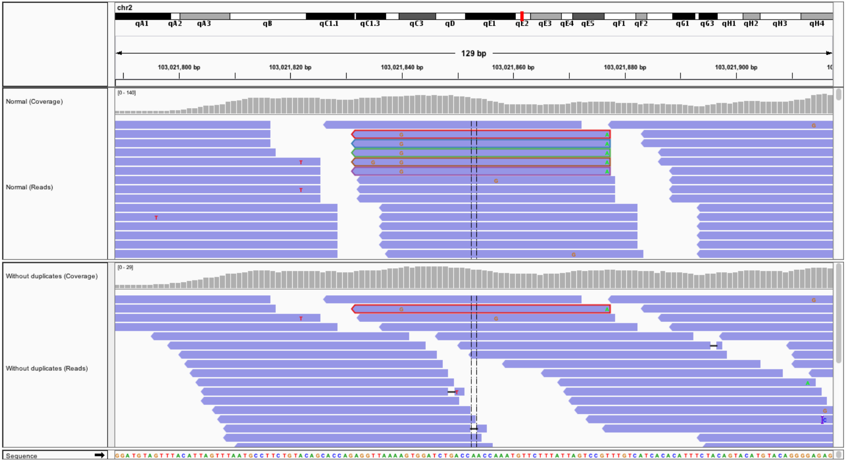
Duplicates removal

- Can be done on raw reads using fastx (reduces mapping runtime)
- Hard to decided whether duplicates were caused by PCR or biological/experimental reasons
- Mapping all reads first and removing duplicates after manual inspection is more sensible
- Picard-tools identifies duplicates after mapping

Duplicates removal

- Can be done on raw reads using fastx (reduces mapping runtime)
- Hard to decided whether duplicates were caused by PCR or biological/experimental reasons
- Mapping all reads first and removing duplicates after manual inspection is more sensible
- Picard-tools identifies duplicates after mapping
- For some data experiments removing duplicates can greatly effect down stream analysis!

Example: sequence duplication

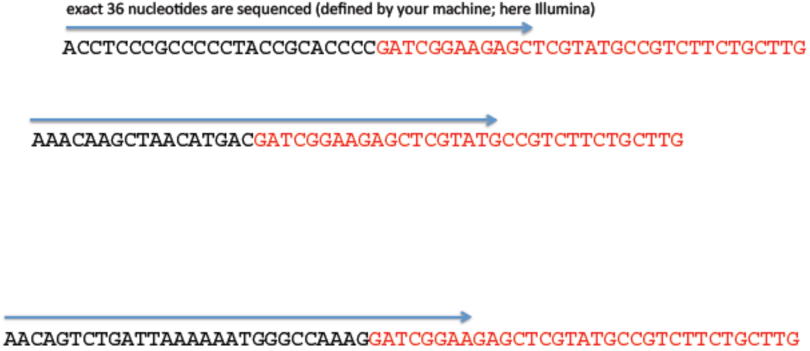


Library complexity

- Library complexity refers to the number of unique fragments present in a given library
- Complexity is affected by:
  - Amount of starting material
  - Amount of DNA lost during cleanups and size selection
  - Amount of duplication introduced via PCR
- For most libraries that only need to be run across a few lanes, the standard protocol provides libraries with ample complexity
- However, certain projects require very deep coverage from a single sample - i.e. SNP discovery, mammalian assembly, cancer resequencing
- When dozens of lanes are required, library complexity becomes very important

Adapter contamination

- If there are fragment smaller than the read length,



exact 36 nucleotides are sequenced (defined by your machine; here Illumina)

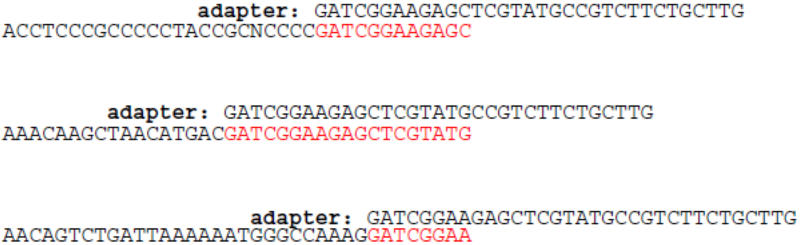
ACCTCCCGCCCCCTACCGCACCCCGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG

AAACAAGCTAACATGACGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG

AACAGTCTGATTAAAAAATGGGCCAAAGGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG

Adapter contamination

- Align adapters to reads



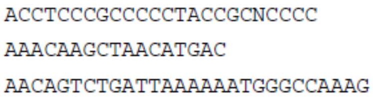
```
               adapter: GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
ACCTCCCGCCCCCTACCGCNCCCCGATCGGAAGAGC

        adapter: GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
AAACAAGCTAACATGACGATCGGAAGAGCTCGTATG

                   adapter: GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
AACAGTCTGATTAAAAAATGGGCCAAAGGATCGGAA
```

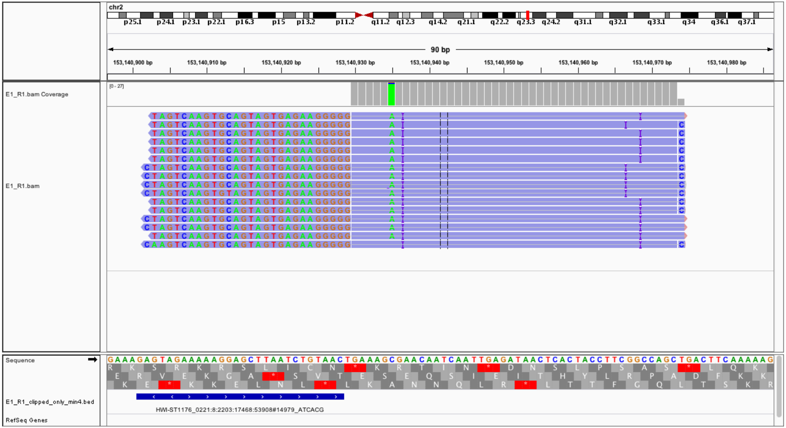
Adapter contamination

- If overlap large enough, remove adapter and everything that follows



```
ACCTCCCGCCCCCTACCGCNCCCC
AAACAAGCTAACATGAC
AACAGTCTGATTAAAAAATGGGCCAAAG
```

Example: Adapter contamination


chr2
p25.1 p24.1 p23.1 p22.1 p16.3 p15 p13.2 p11.2 q11.2 q12.3 q14.2 q21.1 q22.2 q23.3 q24.2 q31.1 q32.1 q33.1 q34 q36.1 q37.1
90 bp
153,140,900 bp
153,140,910 bp
153,140,920 bp
153,140,930 bp
153,140,940 bp
153,140,950 bp
153,140,960 bp
153,140,970 bp
153,140,980 bp
E1_R1.bam Coverage
[0 - 27]
E1_R1.bam
TAGTCAAAGTGCAGTAGTGAGAAGGGGG
CTAGTCAAAGTGCAGTAGTGAGAAGGGGG
CAAGTCAAAGTGCAGTAGTGAGAAGGGGG
Sequence
GAAAGAGTAGAAAAAAGGAGCTTAATCTGTAACTGAAAGCGAACAATCAATTGAGATAACTCACTACCTTCGGCCAGCTGACTTCAAAAAG
E1_R1_clipped_only_min4.bed
HWI-ST1176_0221:8:2203:17468:53908#14979_ATCACG
RefSeq Genes

Finding the correct adapter sequence

If adapters are not mentioned in the study summary:

- Illumina documentation

Finding the correct adapter sequence

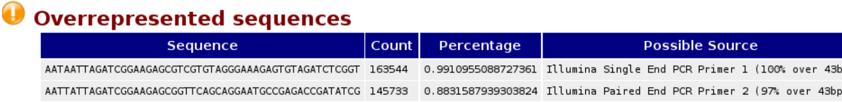
If adapters are not mentioned in the study summary:

- Illumina documentation
- FASTQC Overrepresented sequences

Finding the correct adapter sequence

If adapters are not mentioned in the study summary:

- Illumina documentation
- FASTQC Overrepresented sequences



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AATAATTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGT	163544	0.9910955088727361	Illumina Single End PCR Primer 1 (100% over 43b
AATTATTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATATCG	145733	0.8831587939303824	Illumina Paired End PCR Primer 2 (97% over 43bp

Removing adapter sequence

- Several tools available

Removing adapter sequence

- Several tools available
- Example: cutadapt (`https://code.google.com/p/cutadapt/`)

```
$ cutadapt -m 15 -a GATCGGAAGAGCACACGTCTGAACTCCAGT-
CACACAGTGATCTCGTATGCCGTCTTCTGCTTG SRR501544.fastq >
SRR501544_cutadapt.fastq
```

  - -a Sequence of an adapter that was ligated to the 3’ end
  - -m Discard trimmed reads that are shorter than LENGTH

Removing adapter sequence

- Several tools available
- Example: cutadapt (https://code.google.com/p/cutadapt/)

```
$ cutadapt -m 15 -a GATCGGAAGAGCACACGTCTGAACTCCAGT-
CACACAGTGATCTCGTATGCCGTCTTCTGCTTG SRR501544.fastq >
SRR501544_cutadapt.fastq
```

  - -a Sequence of an adapter that was ligated to the 3’ end
  - -m Discard trimmed reads that are shorter than LENGTH
- Cutadapt can also remove low-quality ends (-q <quality cut off>)

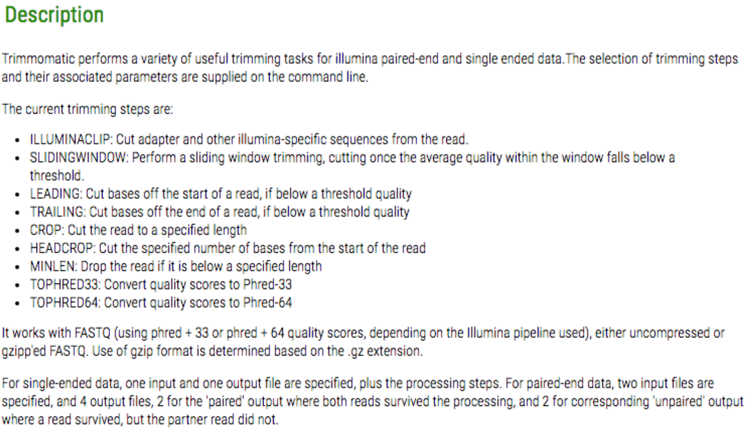
Removing adapter sequence

- Several tools available
- Example: cutadapt (https://code.google.com/p/cutadapt/)

```
$ cutadapt -m 15 -a GATCGGAAGAGCACACGTCTGAACTCCAGT-
CACACAGTGATCTCGTATGCCGTCTTCTGCTTG SRR501544.fastq >
SRR501544_cutadapt.fastq
```

  - -a Sequence of an adapter that was ligated to the 3’ end
  - -m Discard trimmed reads that are shorter than LENGTH
- Cutadapt can also remove low-quality ends (-q <quality cut off>)
- Cutadapt doesn’t handle paired-end data well. Use trimmomatic or other tools instead

Trimmomatic



Description

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data.The selection of trimming steps and their associated parameters are supplied on the command line.

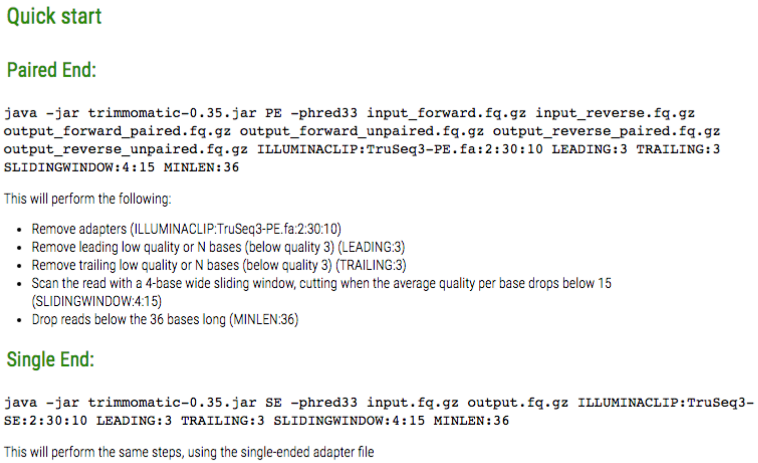
The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

It works with FASTQ (using phred + 33 or phred + 64 quality scores, depending on the Illumina pipeline used), either uncompressed or gzipp'ed FASTQ. Use of gzip format is determined based on the .gz extension.

For single-ended data, one input and one output file are specified, plus the processing steps. For paired-end data, two input files are specified, and 4 output files, 2 for the 'paired' output where both reads survived the processing, and 2 for corresponding 'unpaired' output where a read survived, but the partner read did not.

Trimmomatic: quick start



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
output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

This will perform the following:

- Remove adapters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10)
- Remove leading low quality or N bases (below quality 3) (LEADING:3)
- Remove trailing low quality or N bases (below quality 3) (TRAILING:3)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)
- Drop reads below the 36 bases long (MINLEN:36)

Single End:

```
java -jar trimmomatic-0.35.jar SE -phred33 input.fq.gz output.fq.gz ILLUMINACLIP:TruSeq3-
SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

This will perform the same steps, using the single-ended adapter file

Summary

- Understand plots that you see
- Don’t just look at the first plot and move on
- Methods that rely on counting reads (RNA-Seq, ChIP-Seq) are sensitive to duplication rates
- Methods that rely on assembling unknown genomes/transcriptomes are sensitive to base calling errors
- It is possible to overcorrect!