MEDI6215

University of Southampton

Introduction to Module 7

Bioinformatics, Interpretation and
Data Quality Assurance in Genome Analysis

Will Tapper

6th February 2017

Who are we?

Module leads:

Dr William Tapper
Senior Research Fellow in Genomic informatics
PhD, Genetic Epidemiology, University of Southampton
W.J.Tapper@soton.ac.uk
http://www.southampton.ac.uk/medicine/about/staff/wjt.page

Dr Christopher Woelk
Reader in Genomics and Bioinformatics
Director of Bioinformatics Core Facility
PhD, Viral Evolution, University of Oxford
C.H.Woelk@soton.ac.uk
http://www.southampton.ac.uk/medicine/about/staff/chw1f12.page

Learning outcomes

1. Analyse the principals of sequence data quality control, alignment, variant calling, annotation and filtering strategies to identify pathogenic mutations
2. Interrogate databases of genomic variation and integrate with clinical data to assess the pathogenic and clinical significance of genome results
3. Acquire computer skills and an understanding of statistical methods for analysing NGS data in diagnostics and research
4. Gain practical experience of the bioinformatics pipeline through the Genomics England programme
5. Justify and defend Professional Best Practice Guidelines in the diagnostic setting for the reporting of genomic variation

Course breakdown

Day 1: Monday 6th February

Lectures (9:00 – 11:45)

- Data basics: raw data, quality control, preparation for alignment
- Aligning data to the genome: methodology, assessment and QC
- Tools for viewing aligned data: IGV and UCSC etc

Workshop (13:00 – 17:00)

- An introduction to Galaxy
- Whole Exome data (WES01)
  - Assess raw data quality pre and post filtering
  - Align to reference genome, assess coverage and quality
  - Visualise and interrogate aligned data

Course breakdown

Day 2: Tuesday 7th February

Lectures (9:00 – 11:45)

- Integration of laboratory and clinical information
- Variant calling: identification of SNVs, indels and quality control
- Epigenomics

Workshop (13:00 – 17:00)

- Whole Genome data (NA12878)
  - Call variants and assess sensitivity and specificity
  - Quality control metrics and visualisation
- Call variants in Whole Exome data (WES01)

Course breakdown

Day 3: Monday 27th February

Lectures (9:00 – 11:45)

- Copy number, large indels and structural rearrangements
- Annotation: genes, variation databases, pathogenicity estimates
- Sensitivity and specificity of genomic tests

Workshop (13:00 – 17:00)

- Whole Exome data (WES01)
  - Annotation of variant call files
  - Generate a list of candidate genes
  - Prioritisation of pathogenic variants
  - Creating a running bioinformatic pipeline

Course breakdown

Day 4: Tuesday 28th February

Lectures (9:00 – 11:45)

- Best practice guideline for reporting clinical significance
- Principals of downstream functional analysis
- Genomics England Training Embassy

Workshop (13:00 – 17:00)

- Filtering strategies to identify pathogenic variants
- Set and begin assessment

Assessment

Analysis

Use Galaxy to analyse whole exome data and identify a causal variant that relates to the patients phenotype

Full report (≤1500 words, 75%)

- Alignment
- Variant calling
- Annotation
- Detect causal variant
- Quality control
- Bioinformatic pipeline

Diagnostic report (500 words, 25%)

Communicate NGS result with GP and discuss best practice guidelines for reporting genomic variation in a diagnostic setting

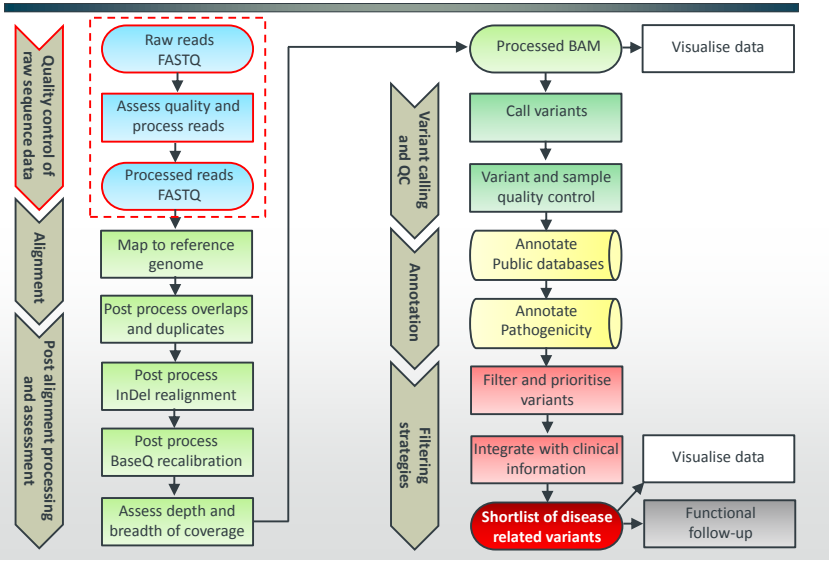
Lecture 1:

Data basics

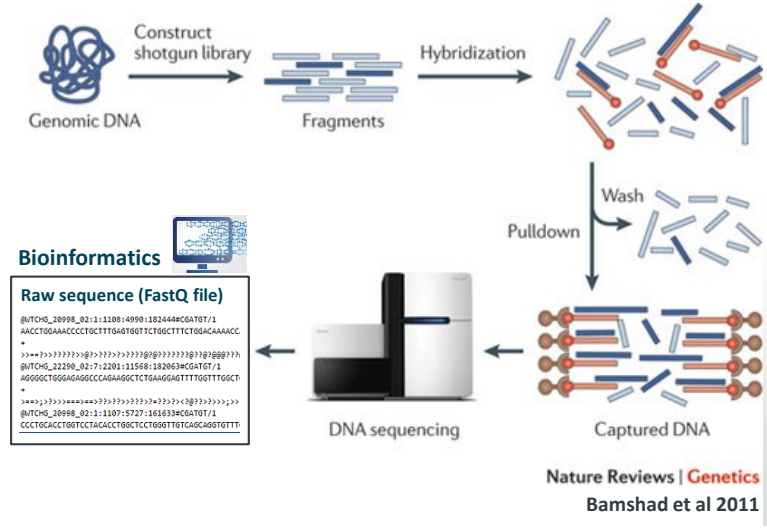
Lecture outline

- Read types (single end, paired end, mate pair) and applications
- Raw sequence data - FastQ files
- Phred scores and the probability of sequencing error
- Methods to assess the quality of raw sequence data
- Quality control:
  - Trim low quality bases from 3'
  - Remove reads with low average quality
  - 5' clipping
  - Remove adapters

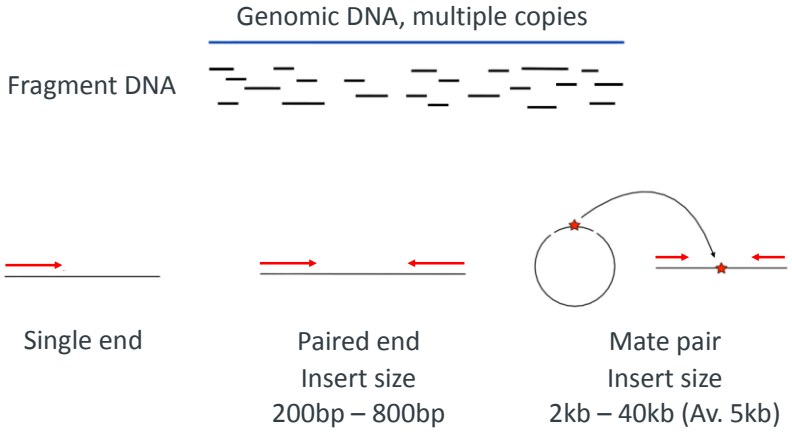
Analysis workflow



Where is the FastQ data from?



Read types



Special applications

Single end

- The best option for degraded DNA samples (FFPE or ancient DNA)
- Fast, cheap, sufficient for counting experiments (eg RNAseq)

Paired end

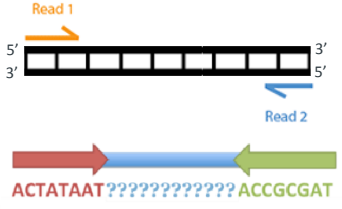
- Better alignment and variant calling
- Good for small to medium insertions and deletions (indels)
- Some structural variation
- Scaffolding in *de novo* genome assembly

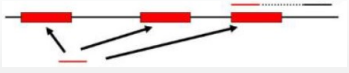
Mate pairs

- Scaffolding in *de novo* genome assembly
- Best option for structural variation

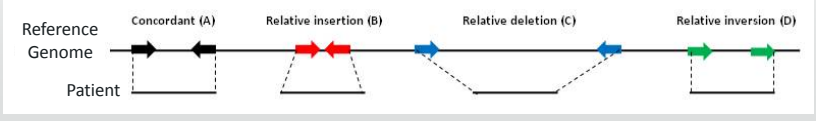
Paired end sequencing

- Sequence both ends (<150bp) of a larger fragment of DNA (200-800bp)
- Provides 2 sequences with a gap between of known length but unknown sequence
- Better alignment, especially near repeat DNA, can map across repeats and recover unmapped reads
- Also useful in identifying structural variation





Single read maps to multiple positions of repeat DNA



FASTQ files

Single end

One FASTQ file per lane of sequencing

Paired end

Read 1

Read 2

2 x FASTQ files per lane Synchronised

Mate pair

Read 1

Read 2

2 x FASTQ files per lane Synchronised

Paired end

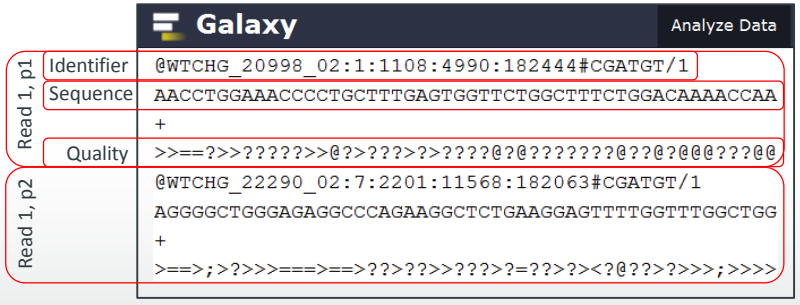
Read 1. Pair 1	Read 2. Pair 1
Read 1. Pair 2	Read 2. Pair 2
Read 1. Pair 3	Read 2. Pair 3
.	.
.	.
.	.

WES01_chr22_R1.fastq.gz

WES01_chr22_R2.fastq.gz

FASTQ files: Raw unaligned reads

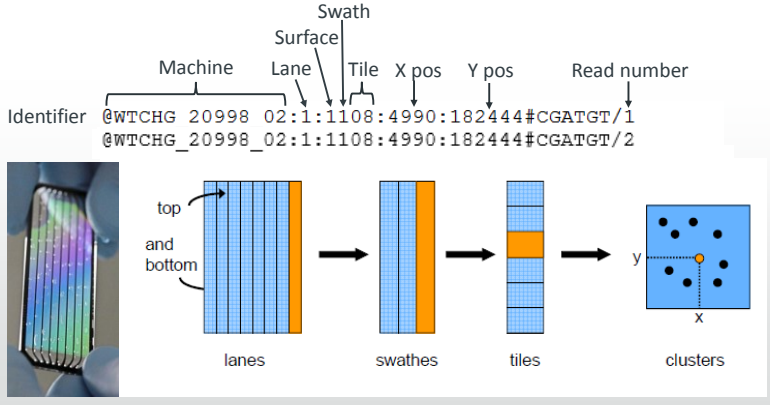
- Simple extension from traditional FASTA format



- Official specification: http://maq.sourceforge.net/fastq.shtml

FASTQ files: Read identifier

- Identifier describes location of reads on the flow cell
- Identifiers are used to match reads that come from the same DNA fragment (ie read 1 and 2 from pair 1)



A quick guide to Phred scaling

- Each bp in the FASTQ file has a quality score on the Phred scale which describes the probability of sequencing error (p-value)
- *Phred value Q* = -10 x $\log_{10}$(p-value), p-value = $10^{\wedge(-Q/10)}$
- Q30 = 0.1% error, 99.9% confidence [-10 x $\log_{10}$(**0.001**)]
- Q20 = 1% error, 99% confidence [-10 x $\log_{10}$(**0.01**)]

Character	ASCII	Phred (Q)	P-value
?	63	30	0.001
5	53	20	0.01

- Importance? Error probability is used in variant calling
- Why not use p-values? Save space: Exome, ~100 Million reads

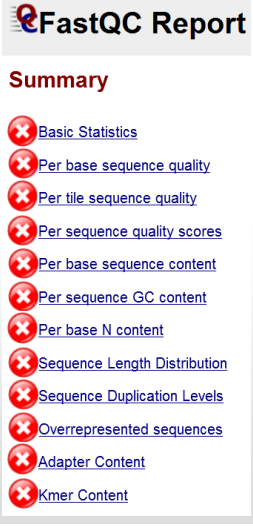
10 Billion bases = 10 Gb

Assess sequence quality and pre-process

- Sequencer output: Reads + quality
- How many reads?
- Is the sequence quality ok?
- Are there any problems & fixes?

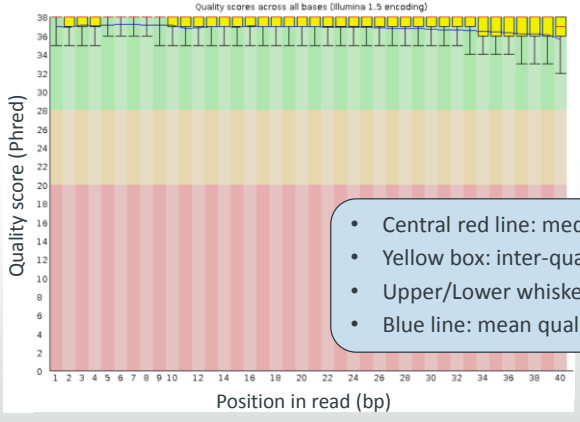
Basic Statistics

Measure	Value
Filename	WTCHG_36466_05_2_sequence.txt.gz
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	943293
Filtered Sequences	0
Sequence length	100
%GC	47



Per base sequence quality

- Range of quality scores over all reads at each position



Good data

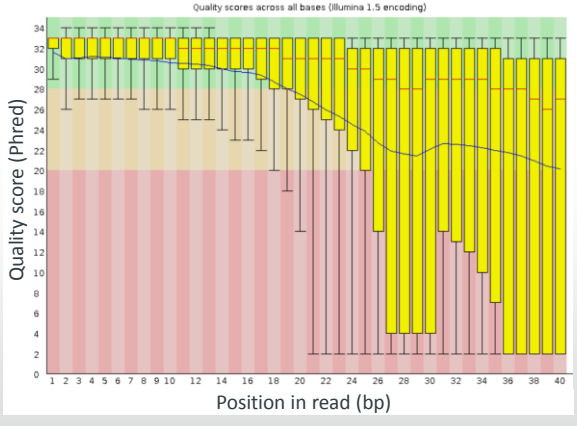
Consistent

High quality along read

- Central red line: median value
- Yellow box: inter-quartile range (25-75%)
- Upper/Lower whiskers: 10% and 90% points
- Blue line: mean quality

Per base sequence quality

- Range of quality scores over all reads at each position



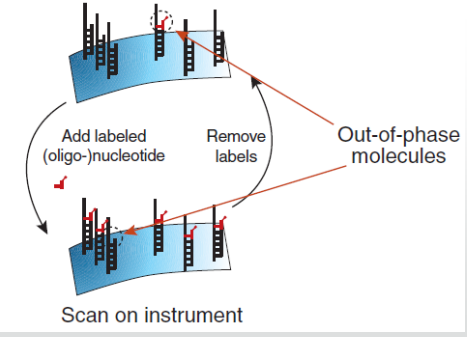
Bad data

- Big range
- Average quality drops towards end of reads

- Trim from 3' to quality of Phred 20

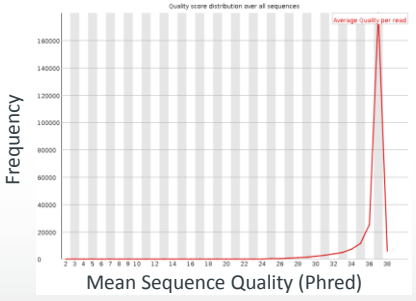
Why does quality drop towards the end?

- Typical for ensemble-based sequencing by synthesis (eg Illumina)
- Sequence determined from average over all copies in a cluster
- Cluster becomes desynchronised, reduces accuracy of average



Per Sequence Quality Distribution

- Average quality scores over all reads



Good data

Most reads have high quality

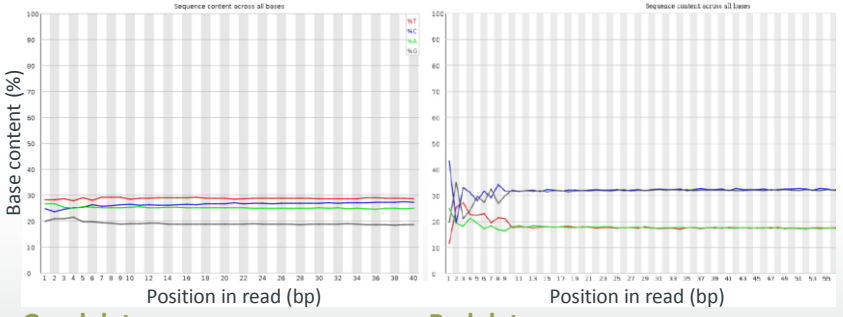
Bad data

Not a uniform distribution

- Remove reads with quality < Q20 over 90% of read

Per Base Sequence Content

- The proportion of each base at each position in the read



Good data

Little difference between bases

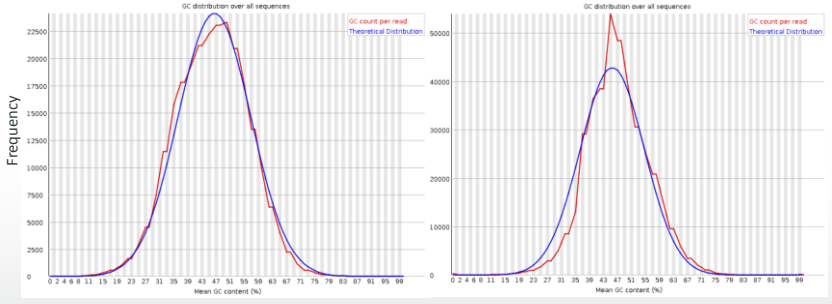
Bad data

Sequence position bias

- Clip a certain number of bases from the 5’ end of reads

Per Sequence GC Content

- Observed GC content per read (red) vs modelled normal distribution (blue)



Good data

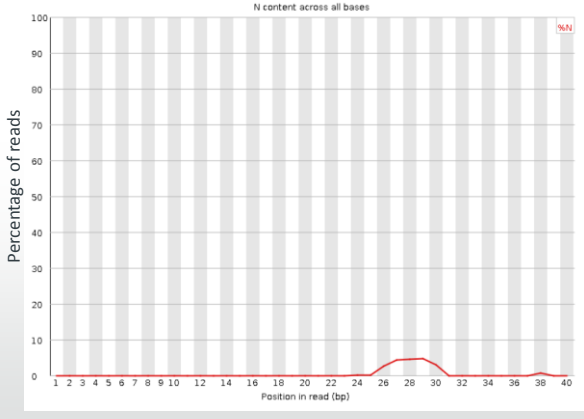
GC content is normally distributed and agrees with the model

Bad data

- Doesn’t fit modelled distribution
- Contaminated library?
- If distributed is shifted it suggests a systematic bias

Per Base N Content

- Percentage of uncalled (N) bases at each position

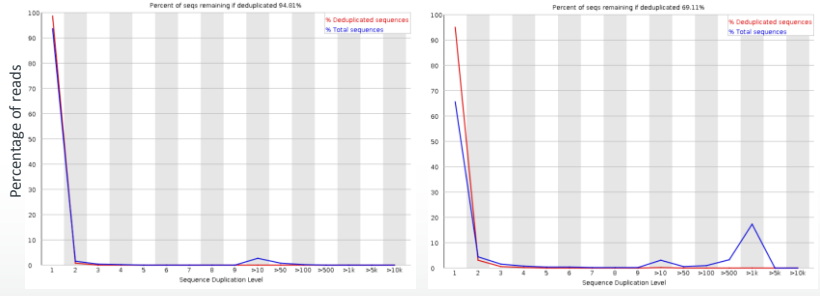


Bad data

N content should be low and uniform across read length

Sequence Duplication Levels

- Proportion of data made from sequences at different duplication levels



Good data

Majority of sequence is unique (left side of plot)

Bad data

- Low proportion of unique sequence
- Spikes of duplicated sequences (RHS)
- Over-sequencing, general enrichment, both lines slope from left to right

Overrepresented Sequences

- Lists sequences that account for > 0.1% of the total
- Sequences accounting for >1% of total are either biologically important or indicative of contamination eg adapters
- Adapters are a common contaminant that occurs due to some DNA fragments being shorter than the read length

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATATCGTATGC	1547768	38.192098035156306	TruSeq Adapter, Index 1 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGC	146635	3.61830603513262	TruSeq Adapter, Index 1 (100% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCAAGATATCGTATGC	6639	0.16382128255358863	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATTTCGTATGC	6462	0.15945370204267054	TruSeq Adapter, Index 1 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTACGATATCGTATGC	5433	0.1340625136486891	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATAACGATATCGTATGC	5147	0.1270052931621209	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACCACGATATCGTATGC	4703	0.11604932849066535	TruSeq Adapter, Index 1 (97% over 41bp)

- Use programs to remove adapters

What is a *k*-mer?

- Create a sliding window of size *k*, move it over all your reads and count the occurrence of *k*-mers

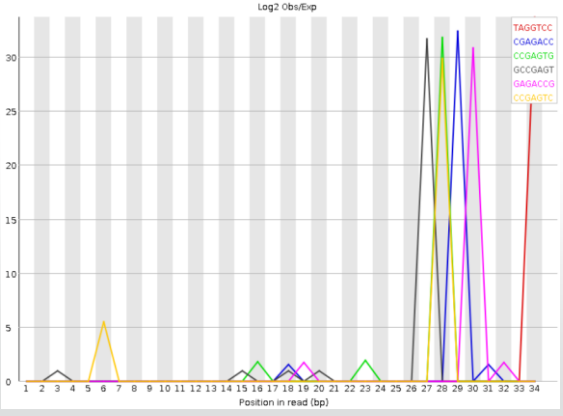
Eg. *k*=5

DNA: ACGTGTAACGTGACGTTGGA

ACGTG
CGTGT
GTGTA

k-mer content

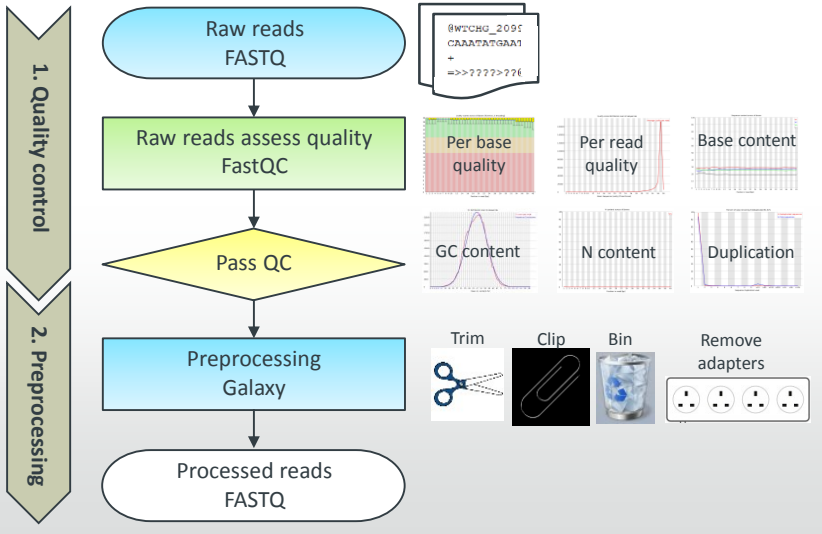
- Observed: counts enrichment of every 7-mer
- Expected: based on base content of the library



Bad data

Any k-mer with more than 10 fold enrichment at any base position

Summary




UNIVERSITY OF
Southampton
Lecture 2:
Alignment