Best practices in the analysis of RNA-seq and ChIP-seq data
27th – 31st, July 2015
University of Cambridge, Cambridge, UK

Quality assessment of NGS data

Ines de Santiago

CRUK Cambridge Research Institute

Ines.desantiago@cruk.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

CAMBRIDGE Cancer CENTRE

CANCER RESEARCH UK

Quality control analysis

All sequencing platform have errors

llumina HiSeq, the current market leader

illumina

SOLEXA: 2006
(2nd generation)

Library: polony PCR, revirsible terminator sequencing

PACIFIC BIOSCIENCES

SOLID systems
2007
(3rd generation)

Library: sequencing by ligation

life technologies

Ion torrent
2001
(2nd generation)

Detecting the protons released as nucleotides are incorporated during synthesis

Oxford NANOPORE Technologies

Minion 2012
(3rd generation)

http://rnaseq.agbioinfo.utk.edu/index.php/

Quality control

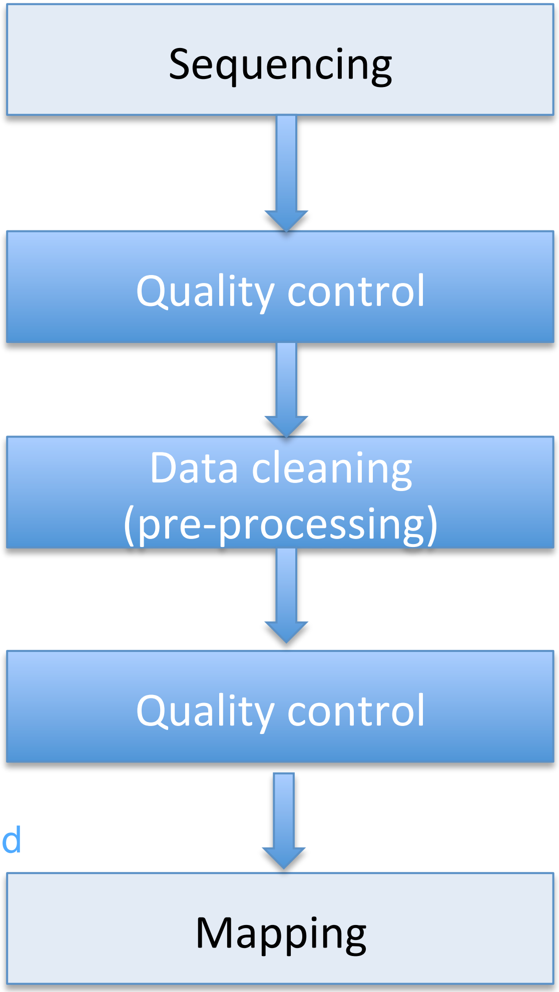
Modern high throughput sequencers can generate hundreds of millions of sequences in a single run.

Before analysing this sequence to draw biological conclusions you should always perform some simple quality control checks to ensure that the raw data looks good and there are no problems or biases in your data which may affect how you can usefully use it.

- It is important to check the quality of your sequenced reads!
- FASTQC: free program that reports quality profile of reads
- Pre-processing
  - Trim reads
  - exclude low quality reads
  - contaminations

Most sequencers will generate a QC report as part of their analysis pipeline, but this is usually only focused on identifying problems which were generated by the sequencer itself.

FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the starting library material.


Sequencing

Quality control

Data cleaning (pre-processing)

Quality control

Mapping

FastQC can be run in one of two modes. It can either run as a stand alone interactive application for the immediate analysis of small numbers of FastQ files, or it can be run in a non-interactive mode where it would be suitable for integrating into a larger analysis pipeline for the systematic processing of large numbers of files.

Checking read quality with FASTQC

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

The analysis in FastQC is performed by a series of analysis modules. The left hand side of the main interactive display or the top of the HTML report show a summary of the modules which were run, and a quick evaluation of whether the results of the module seem entirely normal (green tick), slightly abnormal (orange triangle) or very unusual (red cross).

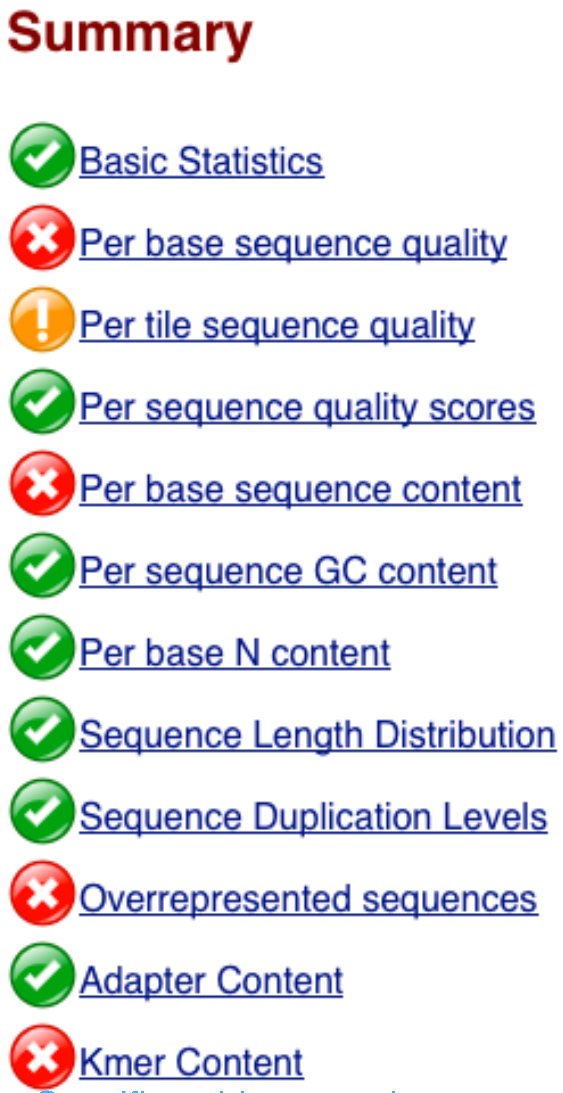
1. Run FASQC

```
fastqc sample.fastq
```

2. Open output file

sample_fastq.html

It is important to stress that although the analysis results appear to give a pass/fail result, these evaluations must be taken in the context of what you expect from your library. A 'normal' sample as far as FastQC is concerned is random and diverse. Some experiments may be expected to produce libraries which are biased in particular ways. You should treat the summary evaluations therefore as pointers to where you should concentrate your attention and understand why your library may not look random and diverse.



Summary

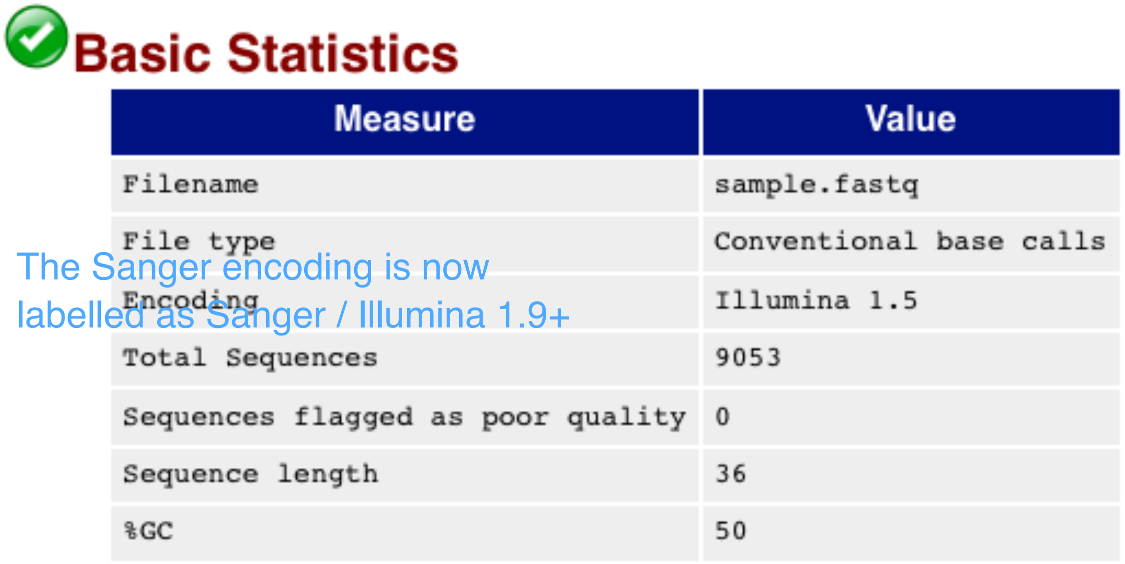
- ✓ Basic Statistics
- ✗ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✓ Adapter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

Specific guidance on how to interpret the output of each module can be found in the modules section of the help.

FASTQC: Report

Generates some simple composition statistics for the file analysed.
Filename: The original filename of the file which was analysed
File type: Says whether the file appeared to contain actual base calls or colorspace data which had to be converted to base calls
Encoding: Says which ASCII encoding of quality values was found in this file.

1) Basic statistics
2) Per base sequence quality
3) Per tile sequence quality
4) Per sequence quality scores
5) Per base sequence content
6) Per sequence GC content
7) Per base N content
8) Sequence Length Distribution
9) Sequence duplication levels
10) Over-represented sequences
11) Adapter/Kmer content



✓ Basic Statistics

Measure	Value
Filename	sample.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	9053
Sequences flagged as poor quality	0
Sequence length	36
%GC	50

The Sanger encoding is now labelled as Sanger / Illumina 1.9+

Never raises errors or warnings

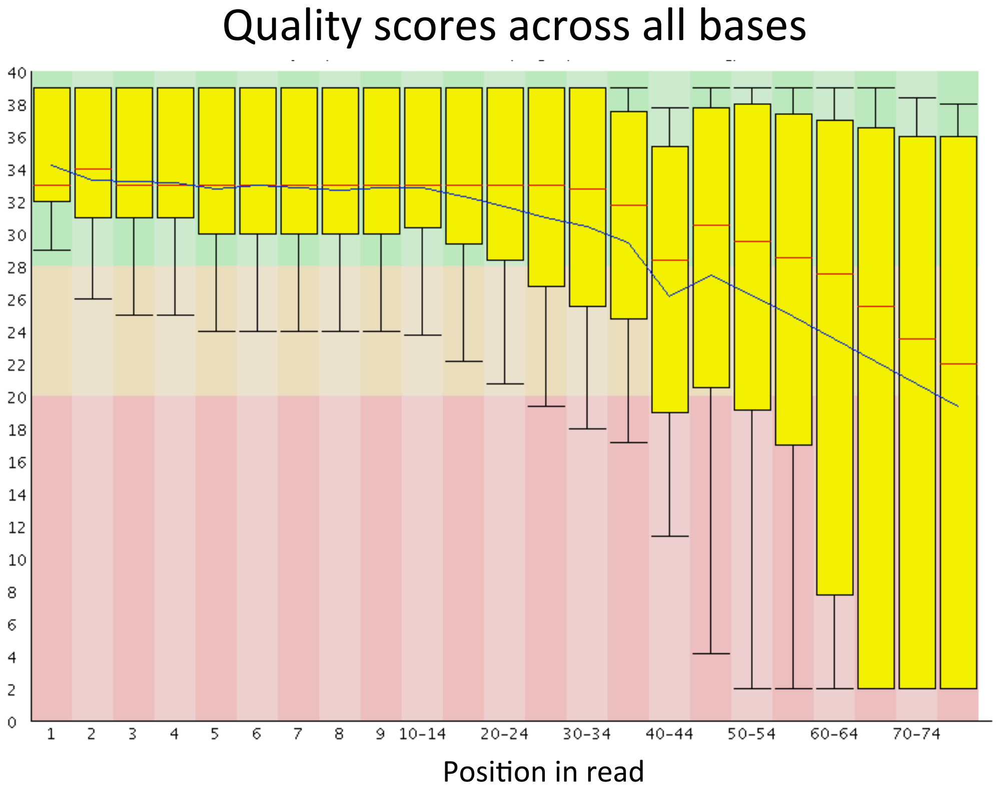
Total Sequences: A count of the total number of sequences processed.
Filtered Sequences: If running in Casava mode sequences flagged to be filtered will be removed from all analyses. The number of such sequences removed will be reported here. The total sequences count above will not include these filtered sequences and will the number of sequences actually used for the rest of the analysis.
Sequence Length: Provides the length of the shortest and longest sequence in the set. If all sequences are the same length only one value is reported.
%GC: The overall %GC of all bases in all sequences

(2) FASTQC: Per base sequence content

- Poor quality at the end of reads

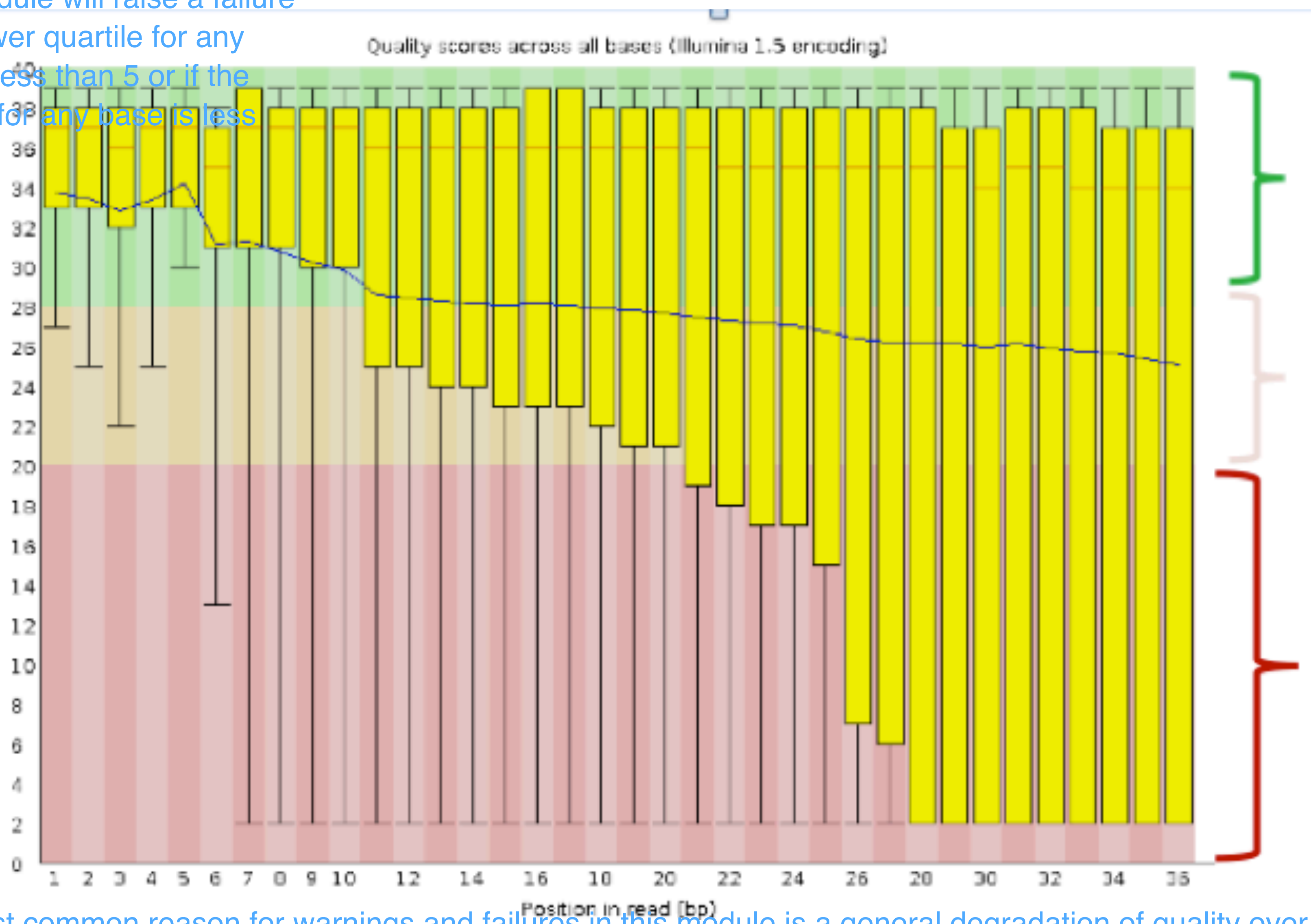
The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

It should be mentioned that there are number of different ways to encode a quality score in a FastQ file. FastQC attempts to automatically determine which encoding method was used, but in some very limited datasets it is possible that it will guess this incorrectly (ironically only when your data is universally very good!). The title of the graph will describe the encoding FastQC thinks your file used.



(2) FASTQC: Per base sequence content

This module will raise a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20



very good quality calls

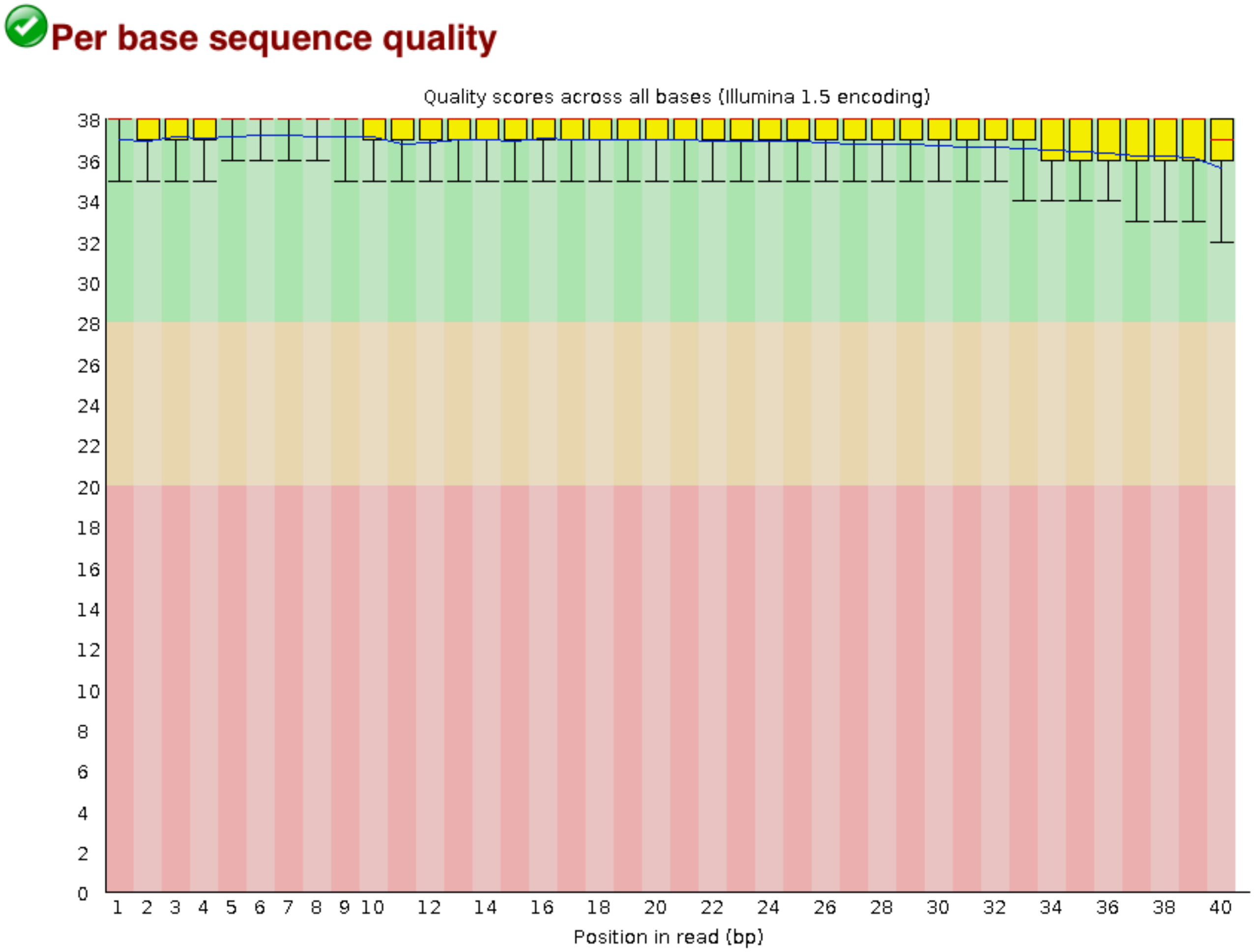
reasonable quality

poor quality

The most common reason for warnings and failures in this module is a general degradation of quality over the duration of long runs. In general sequencing chemistry degrades with increasing read length and for long runs you may find that the general quality of the run falls to a level where a warning or error is triggered.

(2) FASTQC: Per base sequence content

Good Illumina data:

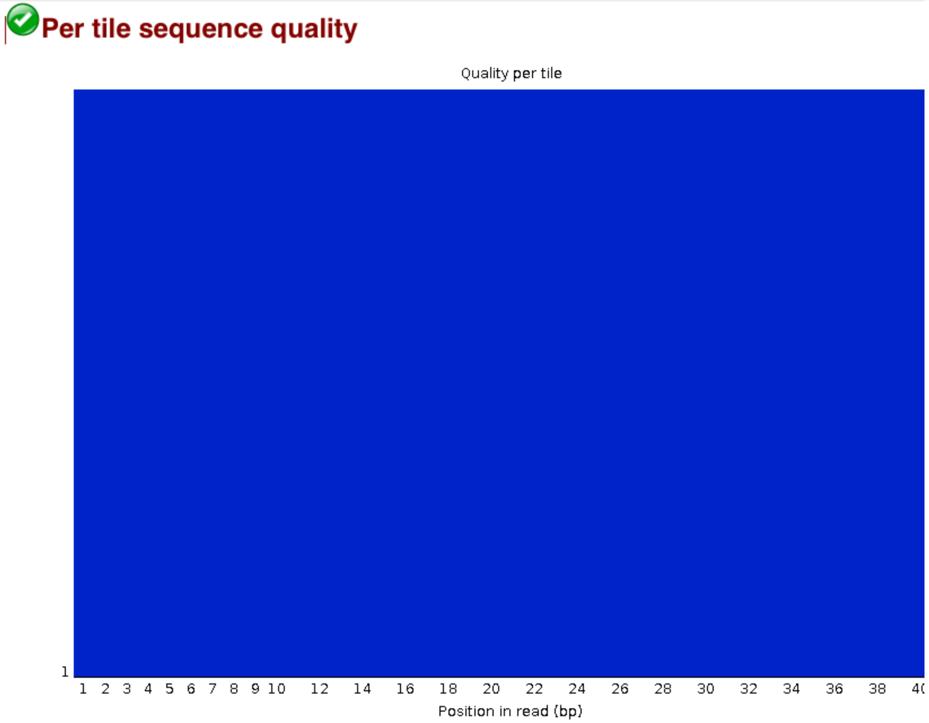


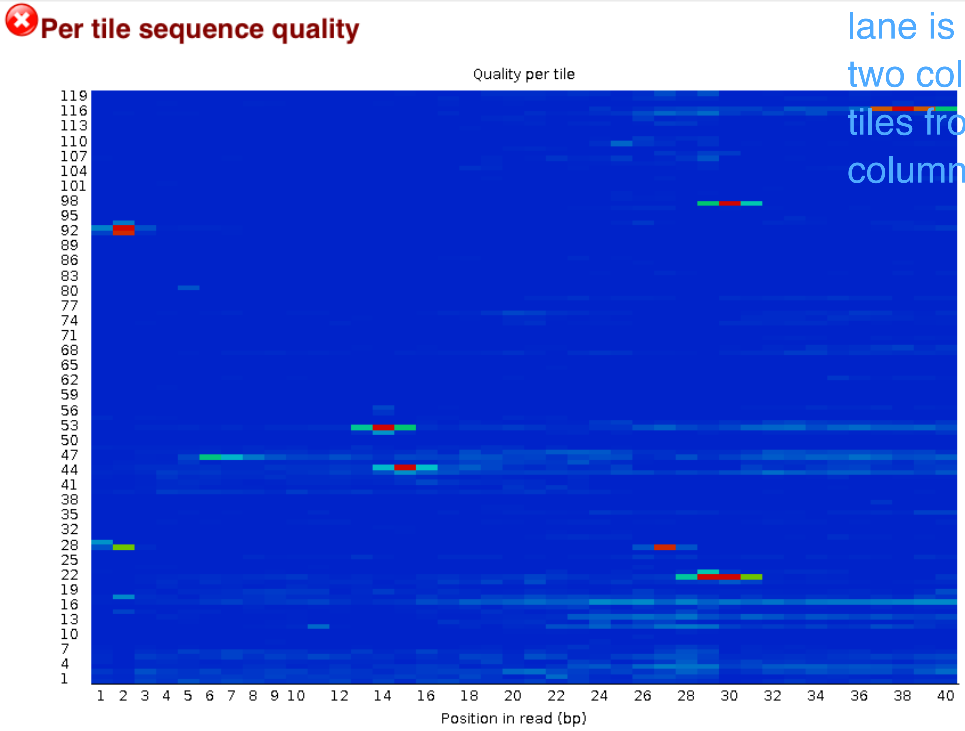
✓ Per base sequence quality

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

(3) FASTQC: Per tile sequence quality

This graph will only appear in your analysis results if you're using an Illumina library which retains its original sequence identifiers. Encoded in these is the flowcell tile from which each read came. The graph allows you to look at the quality scores from each tile across all of your bases to see if there was a loss in quality associated with only one part of the flowcell.





The plot shows the deviation from the average quality for each tile. The colours are on a cold to hot scale, with cold colours being positions where the quality was at or below the average for that base in the run, and hotter colours indicate that a tile had worse qualities than other tiles for that base. In the example below you can see that certain tiles show consistently poor quality. A good plot should be blue all over.

Reasons for seeing warnings or errors on this plot could be transient problems such as bubbles going through the flowcell, or they could be more permanent problems such as smudges on the flowcell or debris inside the flowcell lane.

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

What is a tile?

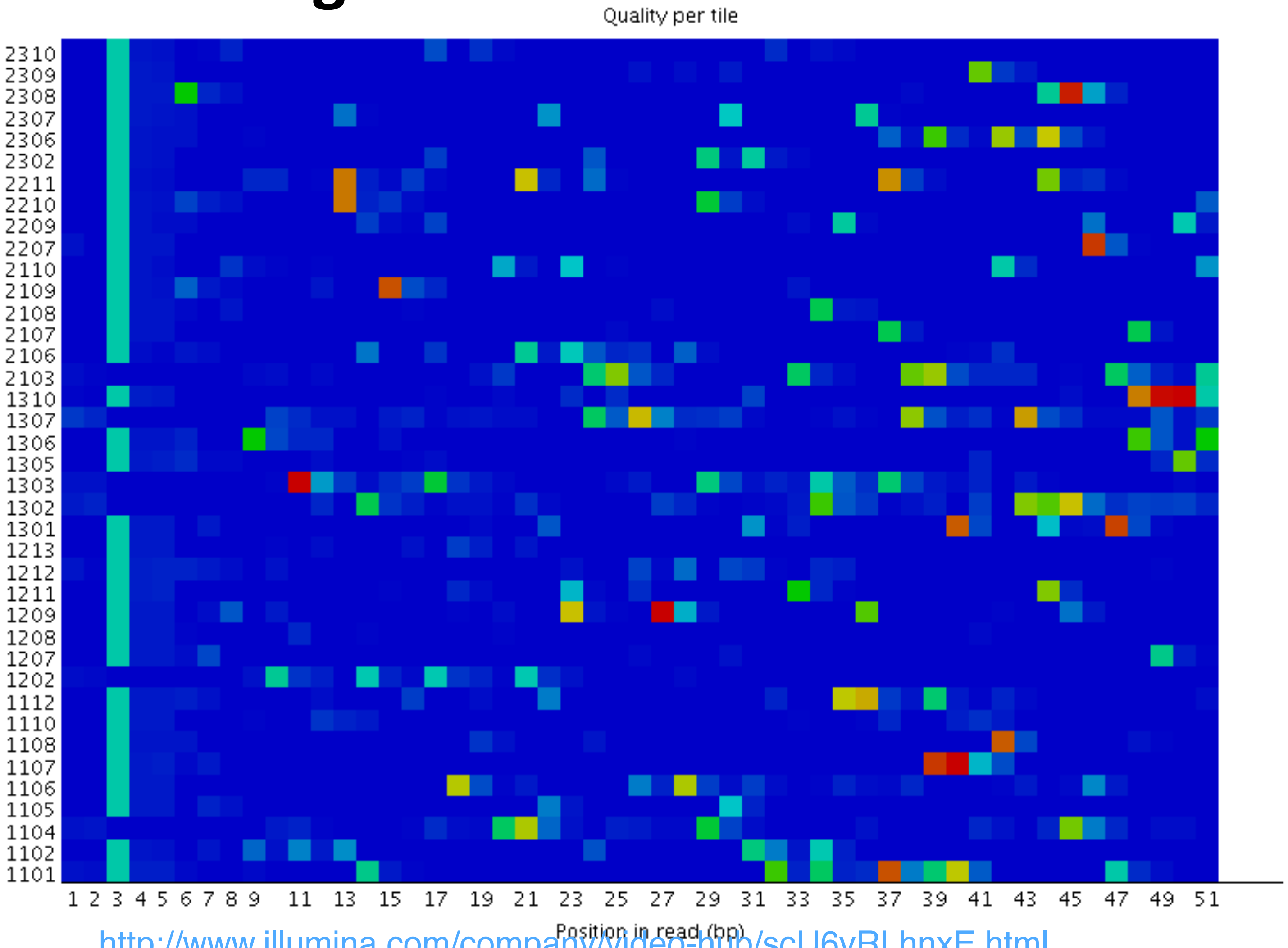
A tile is an image captured by the camera on the Genome Analyzer. flow cell contains eight lanes. Each lane is imaged in two columns with 6 tiles from each column.

We would generally ignore errors which mildly affected a small number of tiles for only 1 or 2 cycles, but would pursue larger effec

(3) FASTQC: Per tile sequence quality

Overclustering:

Overclustering (2 much DNA) creates image analysis problems, including loss of focus and poor template generation. The increased overall signal brightness of the flow cell makes it difficult for the MiSeq System to find the appropriate focal plane. Together these challenges impact sequencing data in the following ways:



http://www.illumina.com/company/video-hub/scU6vRLhnxE.html

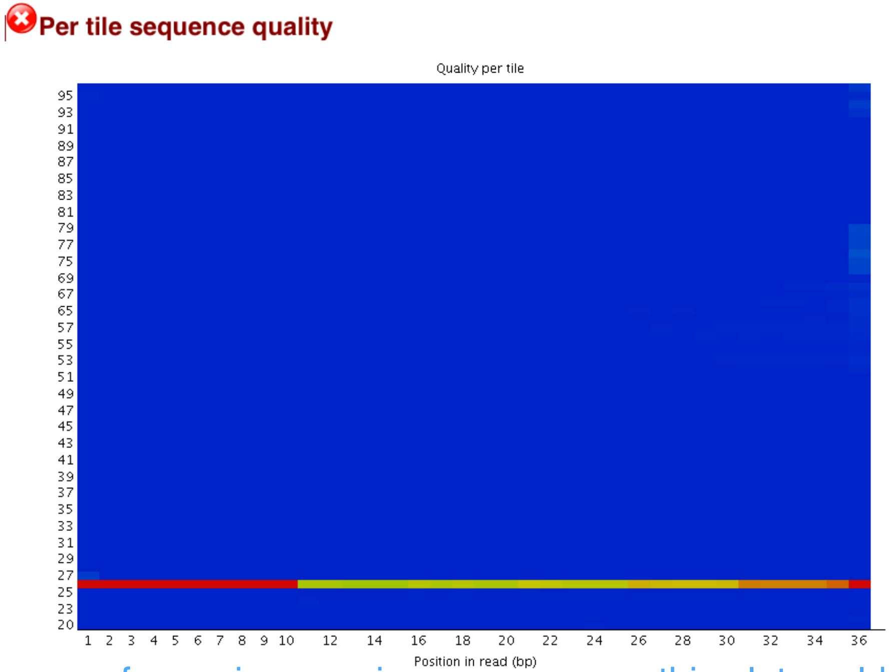
Simon Andrews

In this case events appear all over the flowcell rather than being confined to a specific area or range of cycles.

(3) FASTQC: Per tile sequence quality

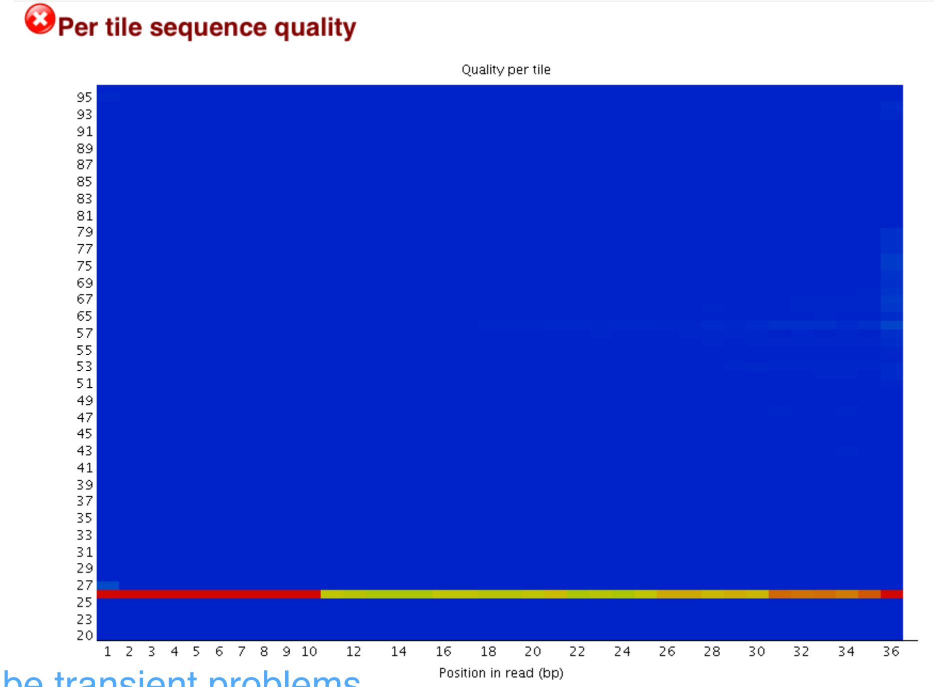
Tile fail:

SRR576938
anaerobic INPUT DNA



Per tile sequence quality

SRR576933
FNR IP ChIP-seq Anaerobic A



Per tile sequence quality

Reasons for seeing warnings or errors on this plot could be transient problems such as bubbles going through the flowcell, or they could be more permanent problems such as smudges on the flowcell or debris inside the flowcell lane.

GSE41187: Genome-wide analysis of FNR and s70 in E. coli under aerobic and anaerobic growth conditions: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41187

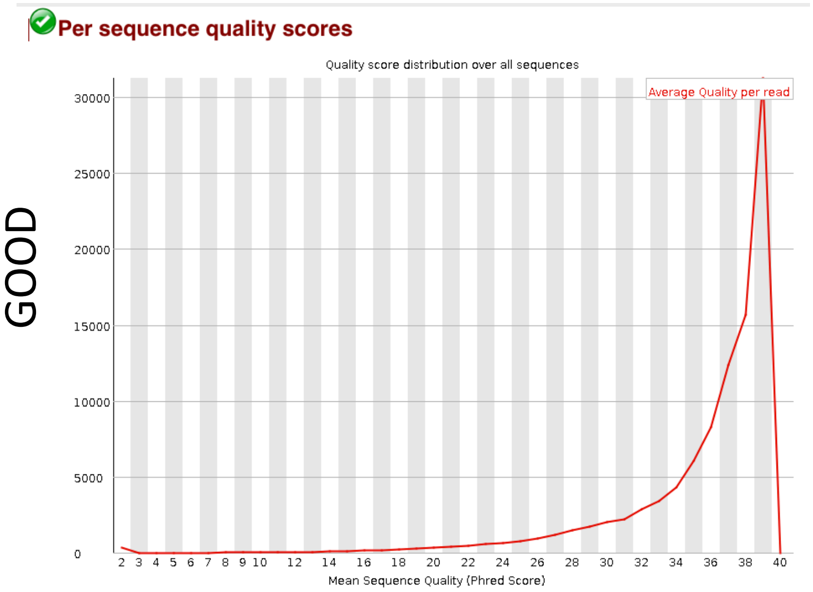
(4) FASTQC: Per sequence quality scores

The per sequence quality score report allows you to see if a subset of your sequences have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge of the field of view etc), however these should represent only a small percentage of the total sequences.

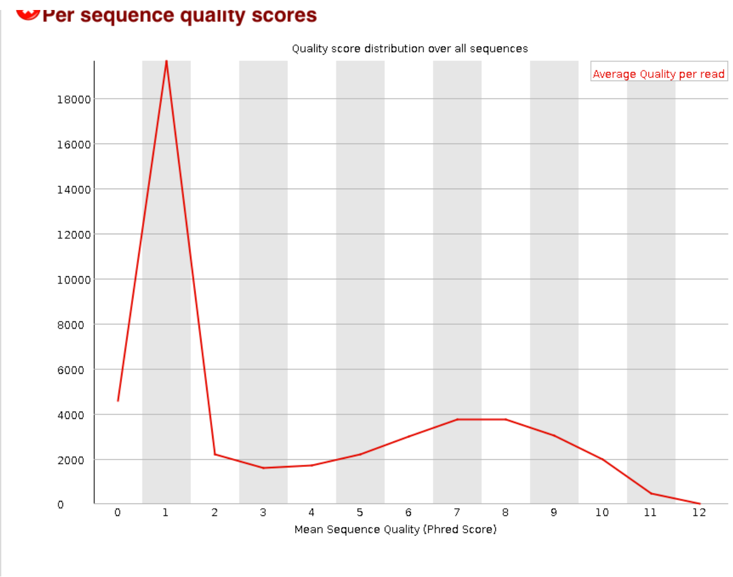
If a significant proportion of the sequences in a run have overall low quality then this could indicate some kind of systematic problem - possibly with just part of the run (for example one end of a flowcell).

Results from this module will not be displayed if your input is a BAM/SAM file in which quality scores have not been recorded.

GOOD



BAD



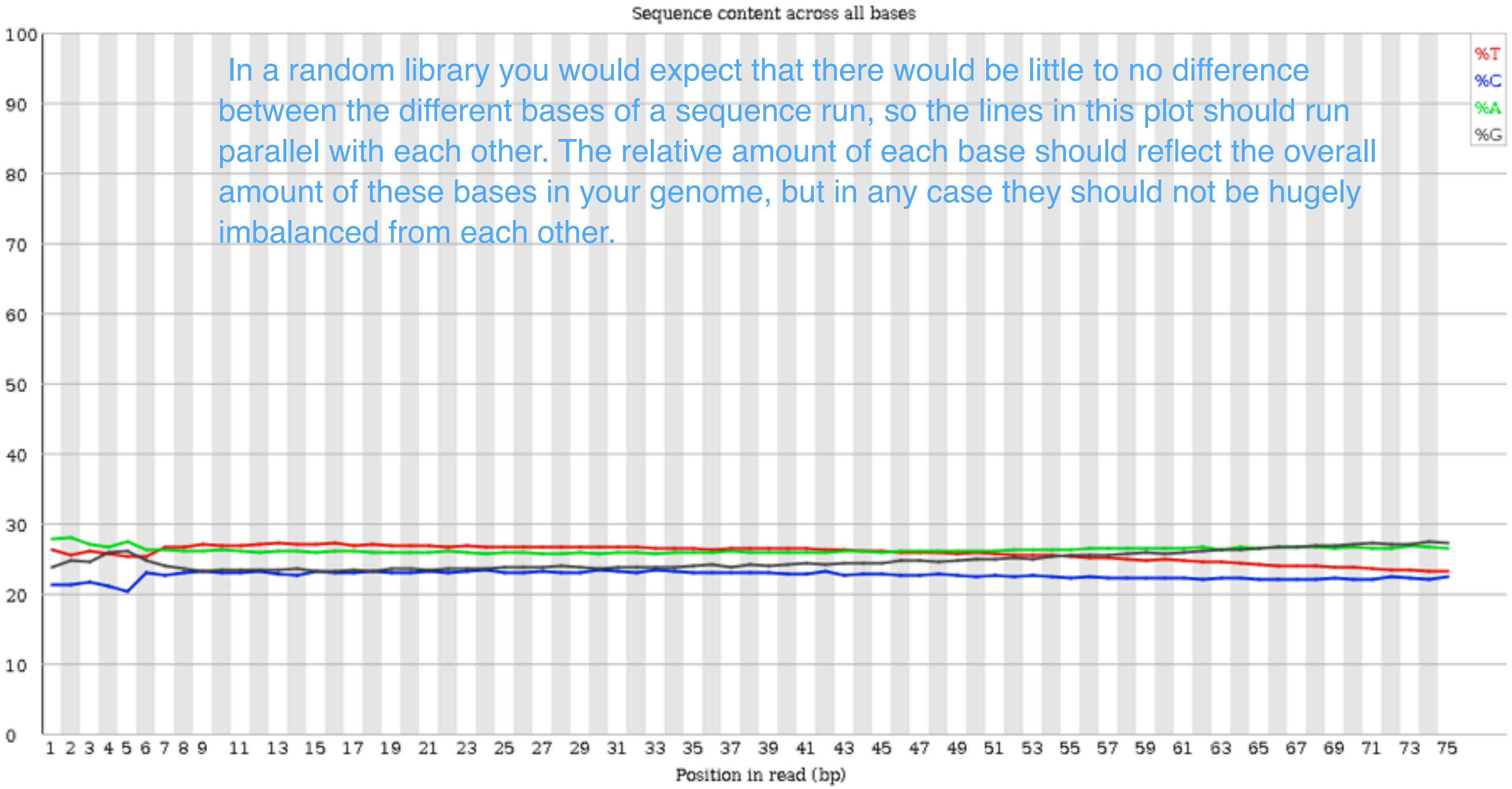
An error is raised if the most frequently observed mean quality is below 20 - this equates to a 1% error rate.

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

(5) FASTQC: Per base sequence content

✓ Per base sequence content

Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.



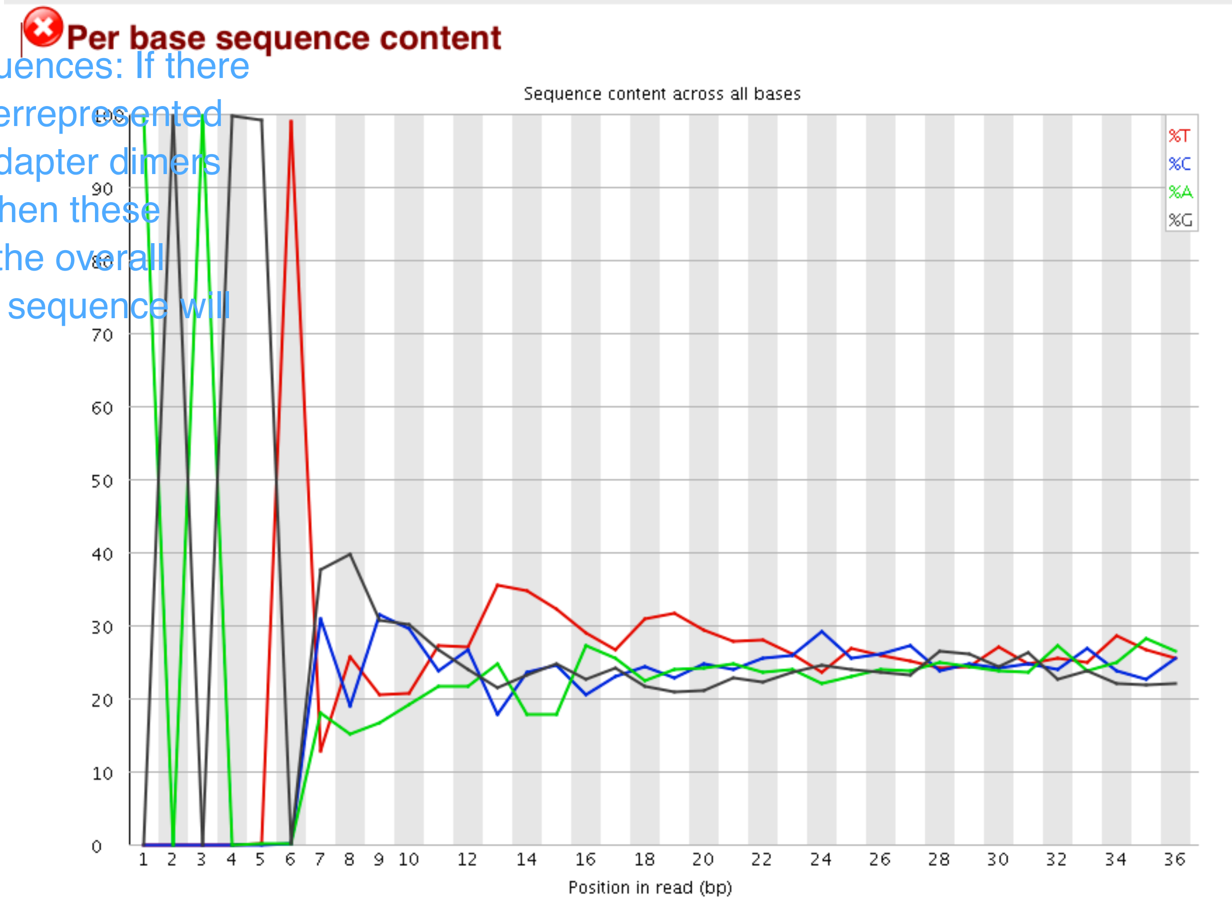
In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. The relative amount of each base should reflect the overall amount of these bases in your genome, but in any case they should not be hugely imbalanced from each other.

http://bio-hpc.kisti.re.kr/MDS_03_normal_chr21.1.fq_fastqc/fastqc_report.html#M3

(5) FASTQC: Per base sequence content

Biased sequence composition (adapters?)

Overrepresented sequences: If there is any evidence of overrepresented sequences such as adapter dimers or rRNA in a sample then these sequences may bias the overall composition and their sequence will emerge from this plot.



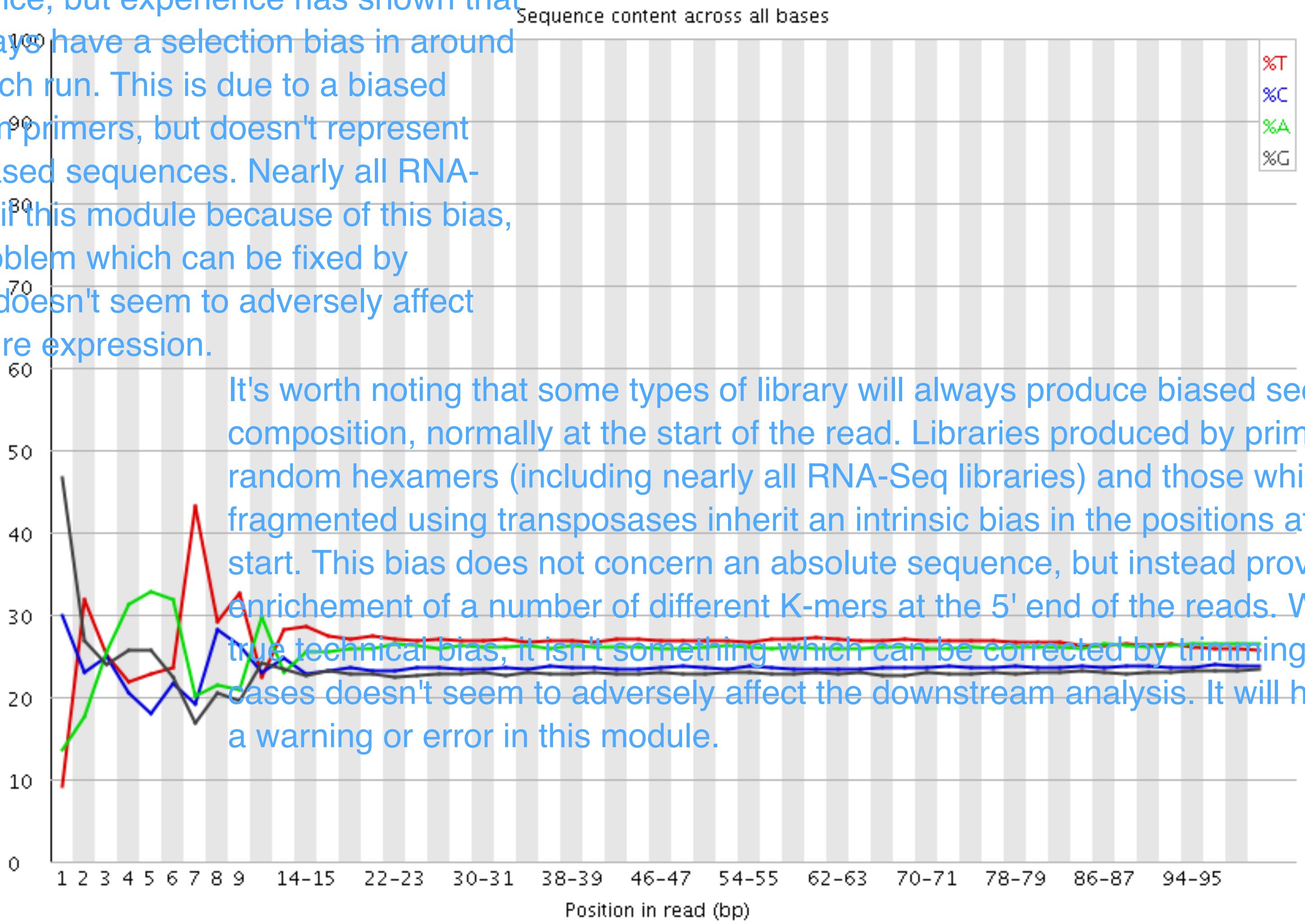
Per base sequence content

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

(5) FASTQC: Per base sequence content

Biased fragmentation: Any library which is generated based on the ligation of random hexamers or through tagmentation should theoretically have good diversity through the sequence, but experience has shown that these libraries always have a selection bias in around the first 12bp of each run. This is due to a biased selection of random primers, but doesn't represent any individually biased sequences. Nearly all RNA-Seq libraries will fail this module because of this bias, but this is not a problem which can be fixed by processing, and it doesn't seem to adversely affect the ablity to measure expression.

Unavoidable – RNA-Seq



It's worth noting that some types of library will always produce biased sequence composition, normally at the start of the read. Libraries produced by priming using random hexamers (including nearly all RNA-Seq libraries) and those which were fragmented using transposases inherit an intrinsic bias in the positions at which reads start. This bias does not concern an absolute sequence, but instead provides enrichement of a number of different K-mers at the 5' end of the reads. Whilst this is a true technical bias, it isn't something which can be corrected by trimming and in most cases doesn't seem to adversely affect the downstream analysis. It will however produce a warning or error in this module.

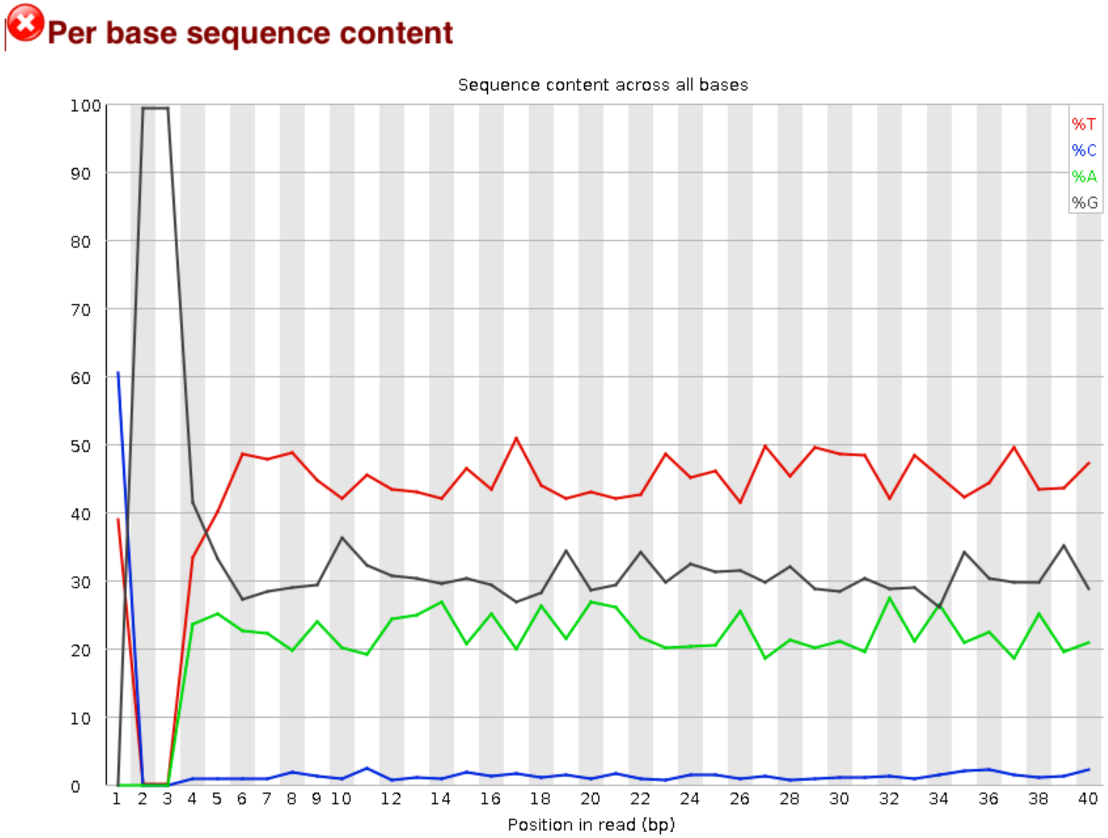
Simon Andrews

(5) FASTQC: Per base sequence content

Unavoidable – RRBS

Devoided of cytosines because the library was treated with sodium bisulphite
(which will have converted most of the C to T)

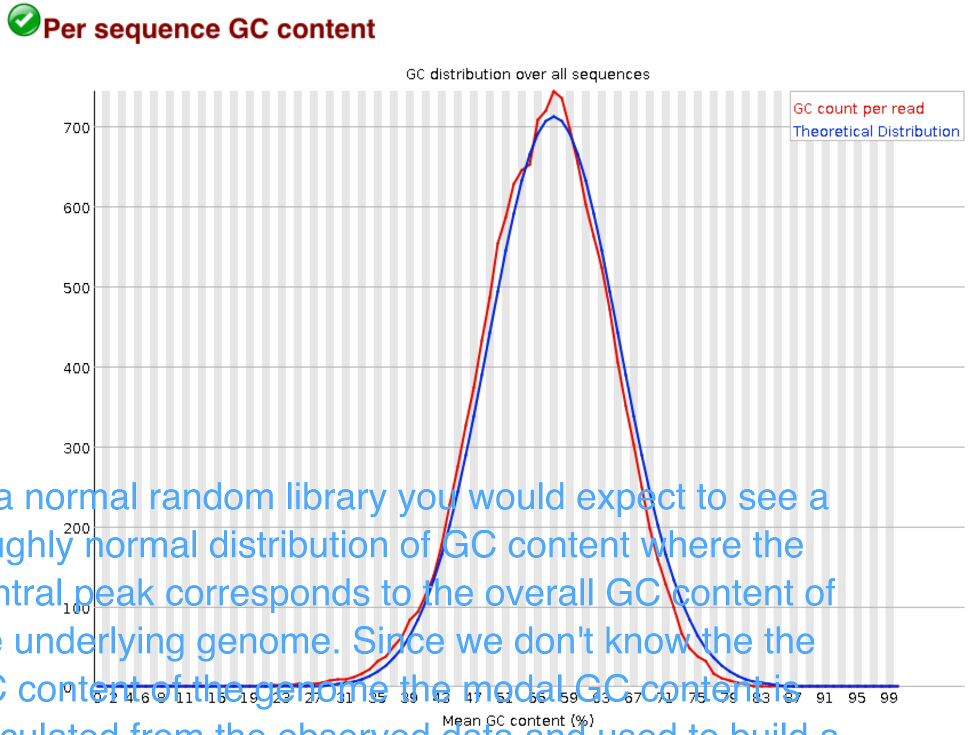
a library which has been treated with sodium bisulphite which will then have converted most of the cytosines to thymines, meaning that the base composition will be almost devoid of cytosines and will thus trigger an error, despite this being entirely normal for that type of library



Per base sequence content

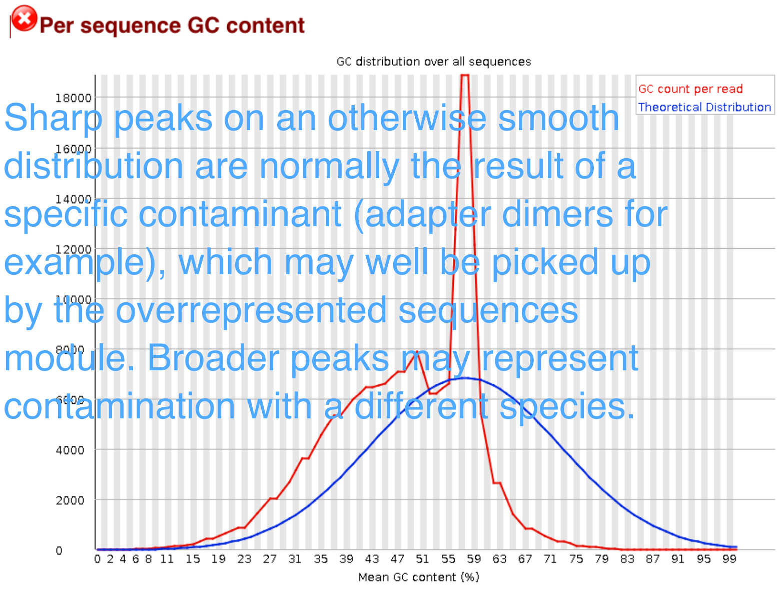
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/RRBS_fastqc.html#M4

(6) FASTQC: Per sequence GC content

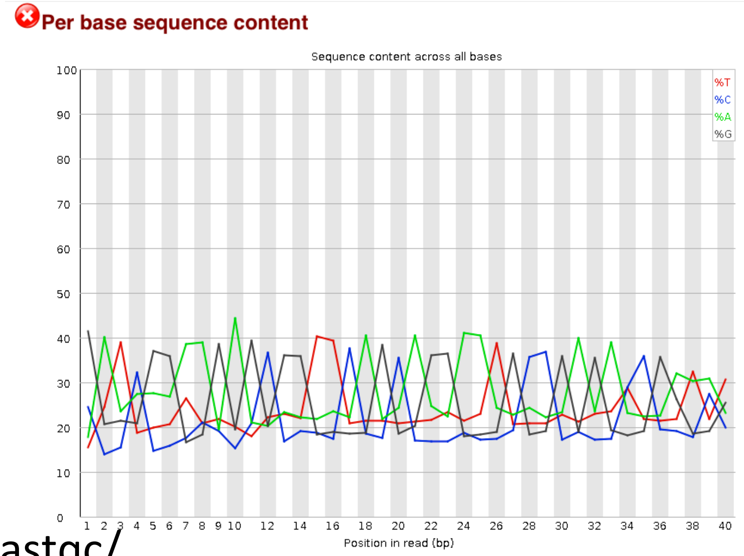


In a normal random library you would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome. Since we don't know the the GC content of the genome the modal GC content is calculated from the observed data and used to build a reference distribution.

An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position. If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an error by the module since it doesn't know what your genome's GC content should be.

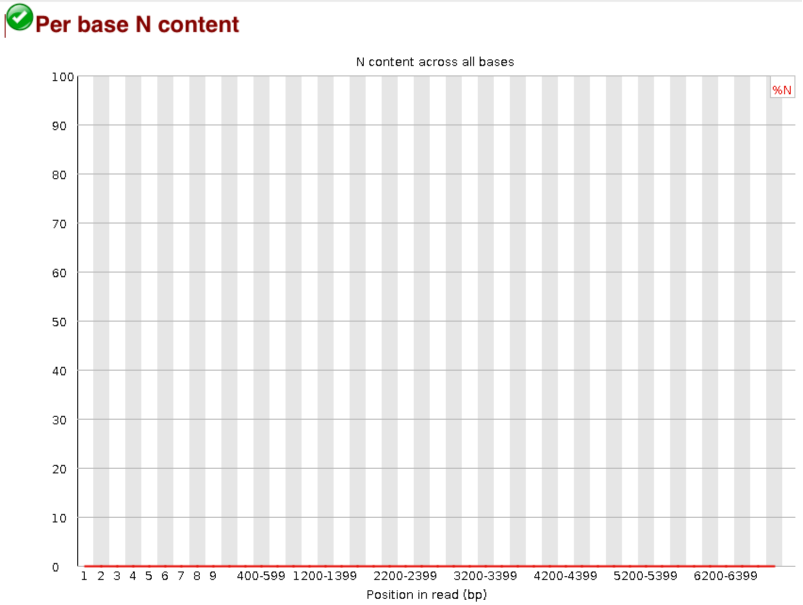


Sharp peaks on an otherwise smooth distribution are normally the result of a specific contaminant (adapter dimers for example), which may well be picked up by the overrepresented sequences module. Broader peaks may represent contamination with a different species.



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

(7) FASTQC: Per base N content

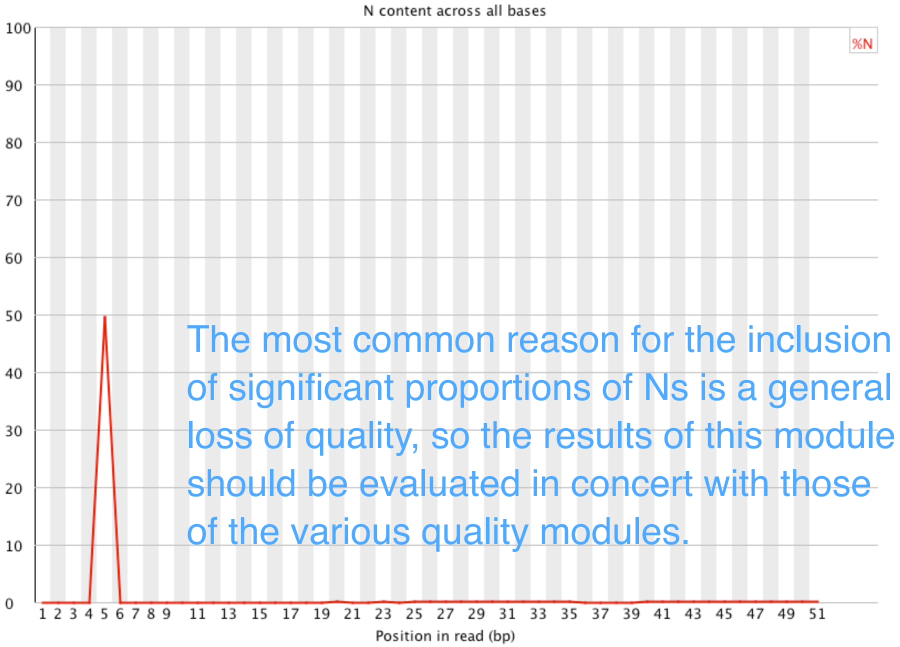


If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base] call

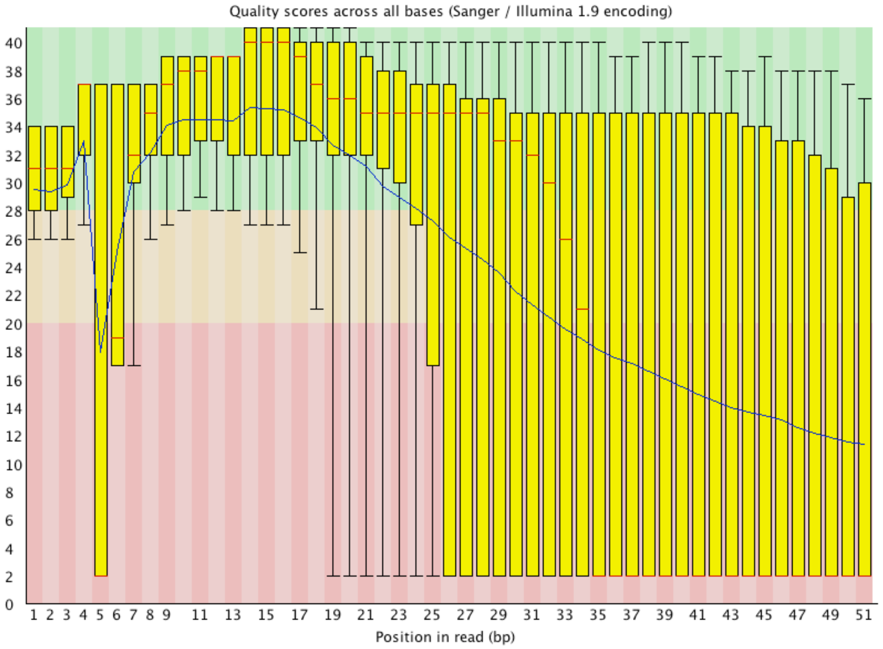
This module plots out the percentage of base calls at each position for which an N was called.

It's not unusual to see a very low proportion of Ns appearing in a sequence, especially nearer the end of a sequence. However, if this proportion rises above a few percent it suggests that the analysis pipeline was unable to interpret the data well enough to make valid base calls.

http://cbio.mskcc.org/~lianos/files/scott/2011-11-21/qc/

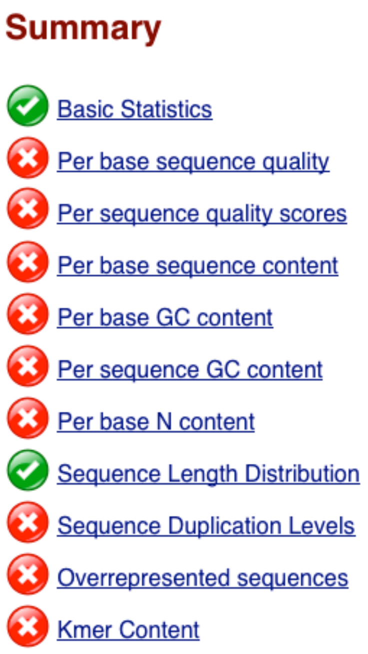


The most common reason for the inclusion of significant proportions of Ns is a general loss of quality, so the results of this module should be evaluated in concert with those of the various quality modules.



(8) FASTQC: Sequence Length Distribution

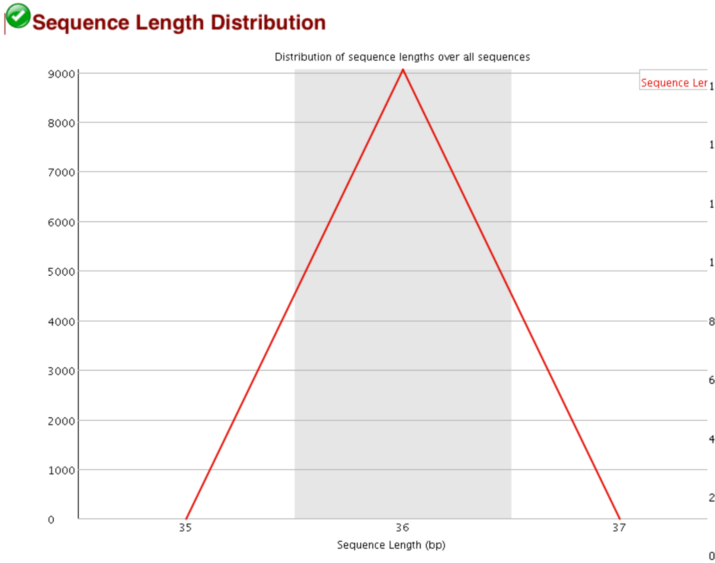
For Illumina it would be typical to obtain the same sequence length for all reads.



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 Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✗ Kmer Content

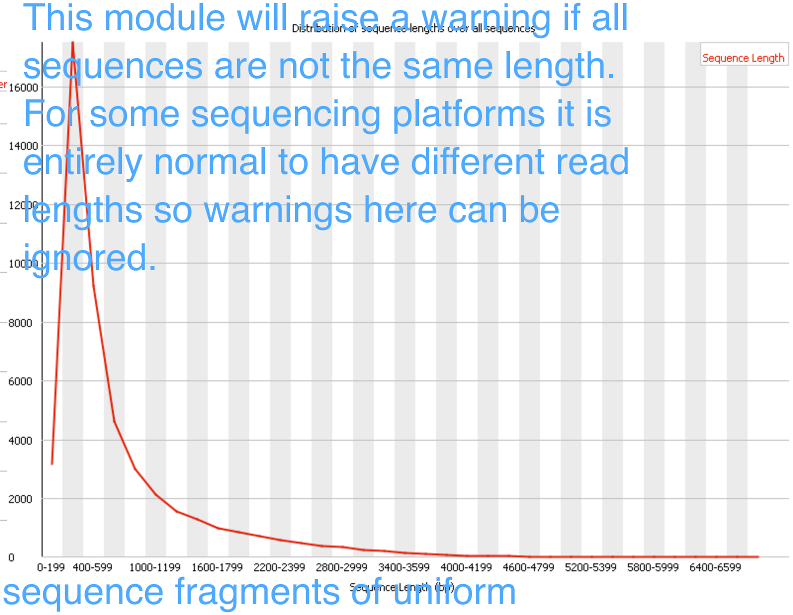
Sequence fragments of uniform length (36bp)



✓ Sequence Length Distribution

Reads of variable length:

This module will raise a warning if all sequences are not the same length. For some sequencing platforms it is entirely normal to have different read lengths so warnings here can be ignored.



Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end.

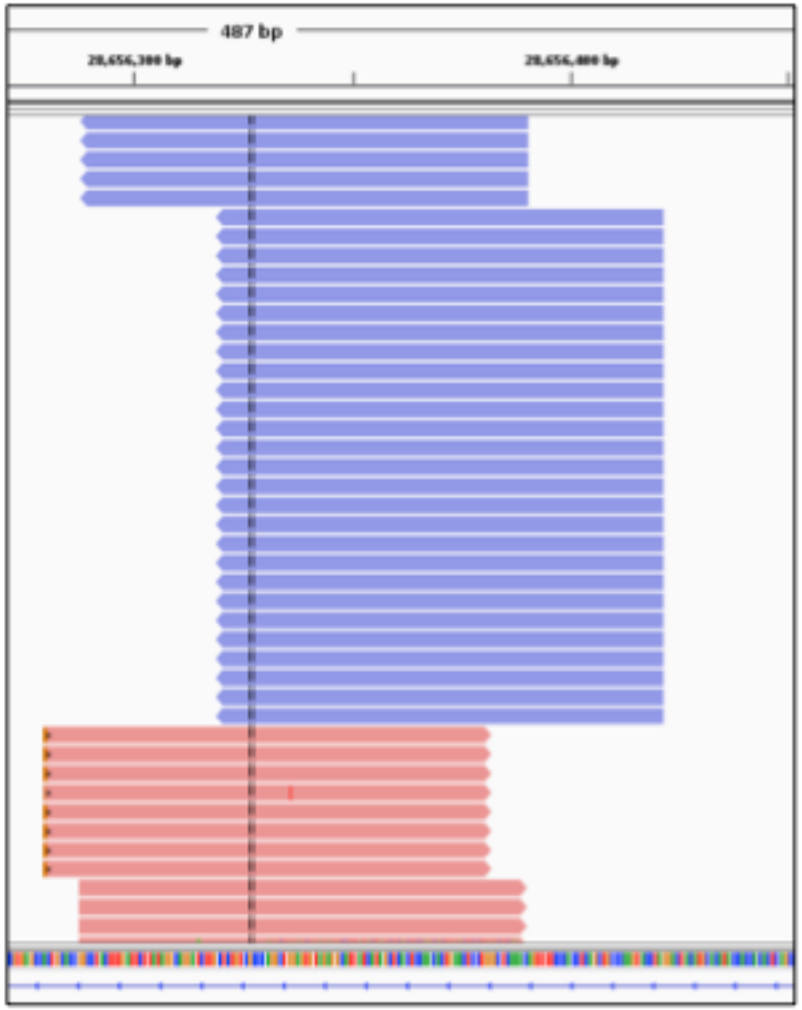
This module generates a graph showing the distribution of fragment sizes in the file which was analysed.

In many cases this will produce a simple graph showing a peak only at one size, but for variable length FastQ files this will show the relative amounts of each different size of sequence fragment.

http://cbio.mskcc.org/~lianos/files/scott/2011-11-21/qc/Bcnr2_ATCACG_L001_R1_001_fastqc/fastqc_report.html#M2

(9) FASTQC: Sequence duplication levels

- PCR duplicates during sample preparation
- Optical duplicates: read the same cluster twice in the sequencer
- High duplication can lead to problems in downstream analysis (e.g. skew allele frequencies)

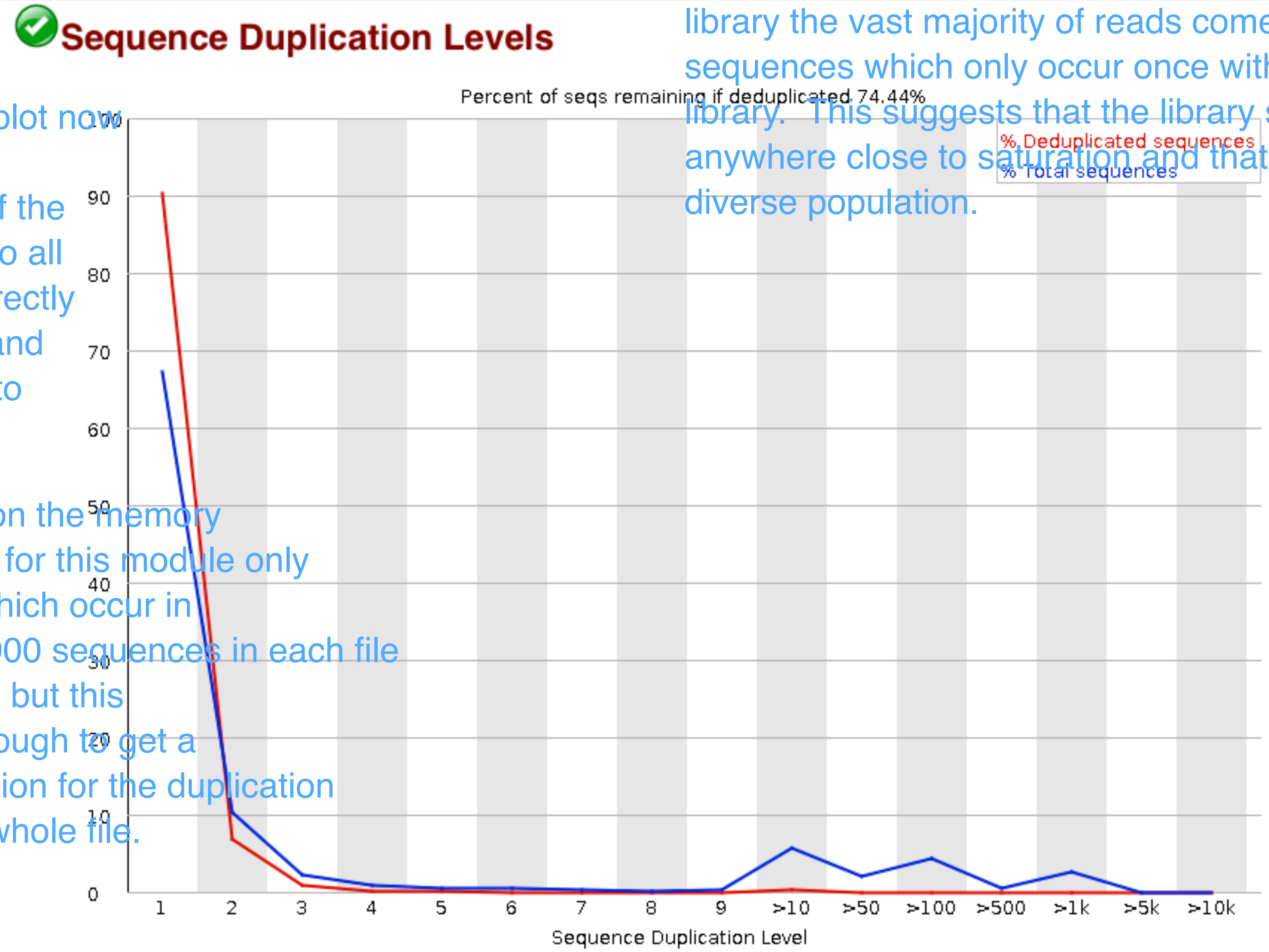


http://bioinformatics.org.au/ws14/wp-content/uploads/ws14/sites/5/2014/07/Felicity-Newell_presentation.pdf

(9) FASTQC: Sequence duplication levels

Very diverse library

Here we have a library with low duplication. In both the raw and deduplicated versions of the library the vast majority of reads come from sequences which only occur once within the library. This suggests that the library sampling isn't anywhere close to saturation and that you have a diverse population.



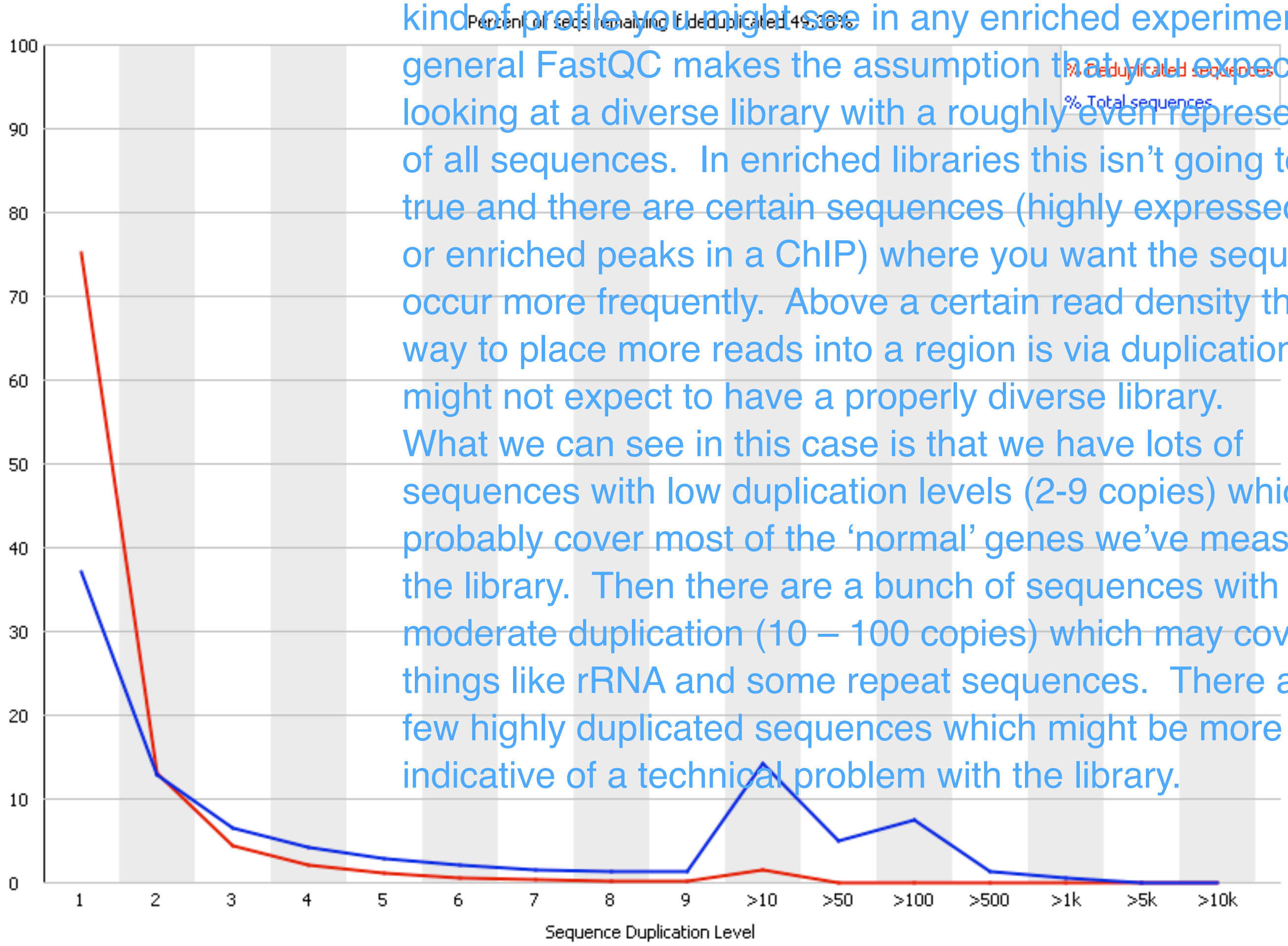
y-axis of the plot now represents a percentage of the total library, so all values are directly comparable and much easier to understand.

To cut down on the memory requirements for this module only sequences which occur in the first 200,000 sequences in each file are analysed, but this should be enough to get a good impression for the duplication levels in the whole file.

http://proteo.me.uk/2013/09/a-new-way-to-look-at-duplication-in-fastqc-v0-11/

(9) FASTQC: Sequence duplication levels

A good RNA-Seq library (although dup levels > 50%)



This is a good RNA-Seq library, but actually represents the kind of profile you might see in any enriched experiment. In general FastQC makes the assumption that you expect to be looking at a diverse library with a roughly even representation of all sequences. In enriched libraries this isn’t going to be true and there are certain sequences (highly expressed genes or enriched peaks in a ChIP) where you want the sequence to occur more frequently. Above a certain read density the only way to place more reads into a region is via duplication so we might not expect to have a properly diverse library.

What we can see in this case is that we have lots of sequences with low duplication levels (2-9 copies) which probably cover most of the ‘normal’ genes we’ve measured in the library. Then there are a bunch of sequences with moderate duplication (10 – 100 copies) which may cover things like rRNA and some repeat sequences. There are very few highly duplicated sequences which might be more indicative of a technical problem with the library.

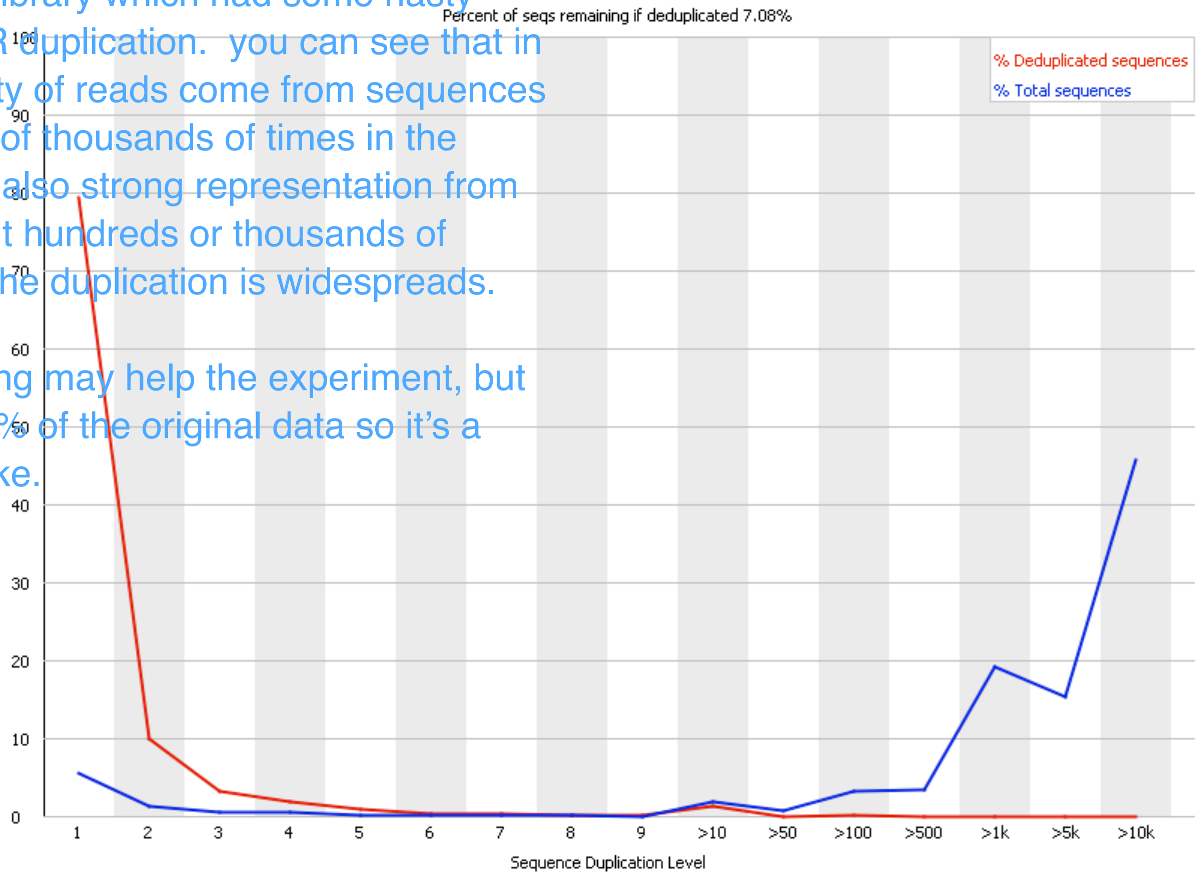
http://proteo.me.uk/2013/09/a-new-way-to-look-at-duplication-in-fastqc-v0-11/

(9) FASTQC: Sequence duplication levels

PCR duplication

This was an RNA-Seq library which had some nasty contamination and PCR duplication. you can see that in the raw data the majority of reads come from sequences which are present tens of thousands of times in the library, but that there is also strong representation from reads which are present hundreds or thousands of times, suggesting that the duplication is widespreads.

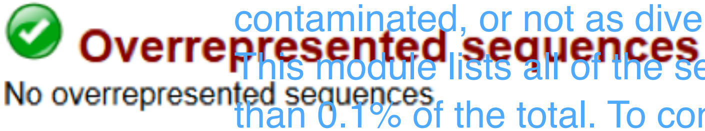
In this case deduplicating may help the experiment, but it causes the loss of 93% of the original data so it's a pretty drastic step to take.



http://proteo.me.uk/2013/09/a-new-way-to-look-at-duplication-in-fastqc-v0-11/

(10) FASTQC: Over-represented sequences

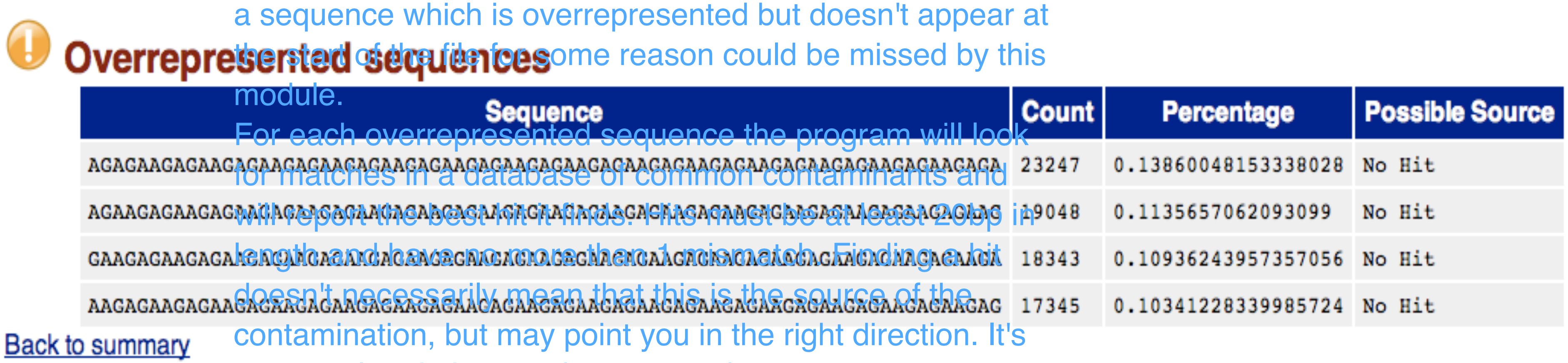
Good dataset



✓ Overrepresented sequences

No overrepresented sequences

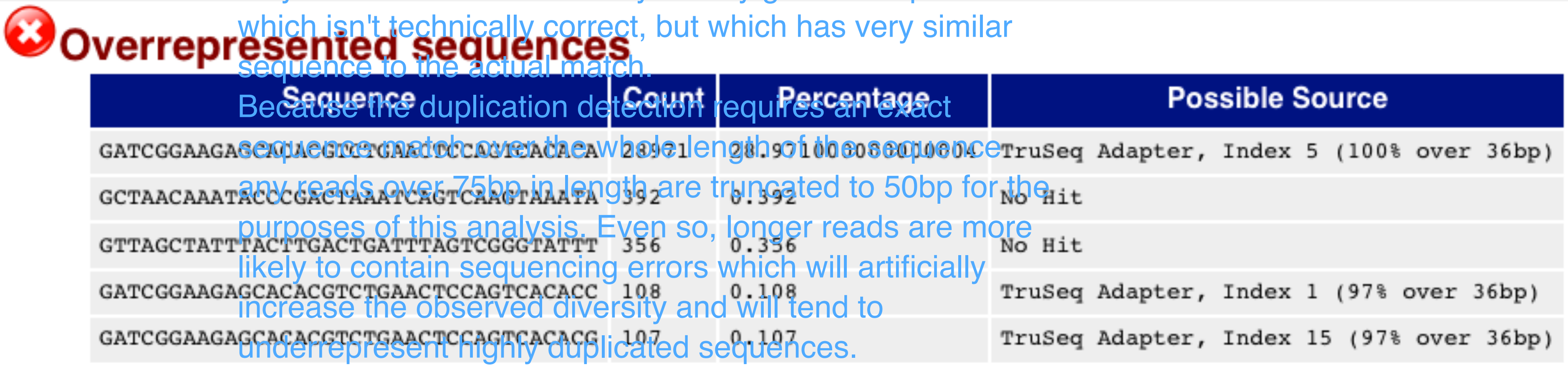
Bad datasets:



! Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGAAGAGAAG[illegible]AAGAGAAGAGAAGAGA	23247	0.13860048153338028	No Hit
AGAAGAGAAGAG[illegible]AG	19048	0.1135657062093099	No Hit
GAAGAGAAGAGA[illegible]AGA	18343	0.10936243957357056	No Hit
AAGAGAAGAGAA[illegible]GAAGAGAAGAG	17345	0.10341228339985724	No Hit

Back to summary



✕ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGA[illegible]CACA	28971	28.971000300010004	TruSeq Adapter, Index 5 (100% over 36bp)
GCTAACAAATACCCGACTAAATCAGTCAAGTAAATA	392	0.392	No Hit
GTTAGCTATTTACTTGACTGATTTAGTCGGGTATTT	356	0.356	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACC	108	0.108	TruSeq Adapter, Index 1 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACG	107	0.107	TruSeq Adapter, Index 15 (97% over 36bp)

A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected.

This module lists all of the sequence which make up more than 0.1% of the total. To conserve memory only sequences which appear in the first 100,000 sequences are tracked to the end of the file. It is therefore possible that a sequence which is overrepresented but doesn't appear at the start of the file for some reason could be missed by this module.

For each overrepresented sequence the program will look for matches in a database of common contaminants and will report the best hit it finds. Hits must be at least 20bp in length and have no more than 1 mismatch. Finding a hit doesn't necessarily mean that this is the source of the contamination, but may point you in the right direction. It's also worth pointing out that many adapter sequences are very similar to each other so you may get a hit reported which isn't technically correct, but which has very similar sequence to the actual match.

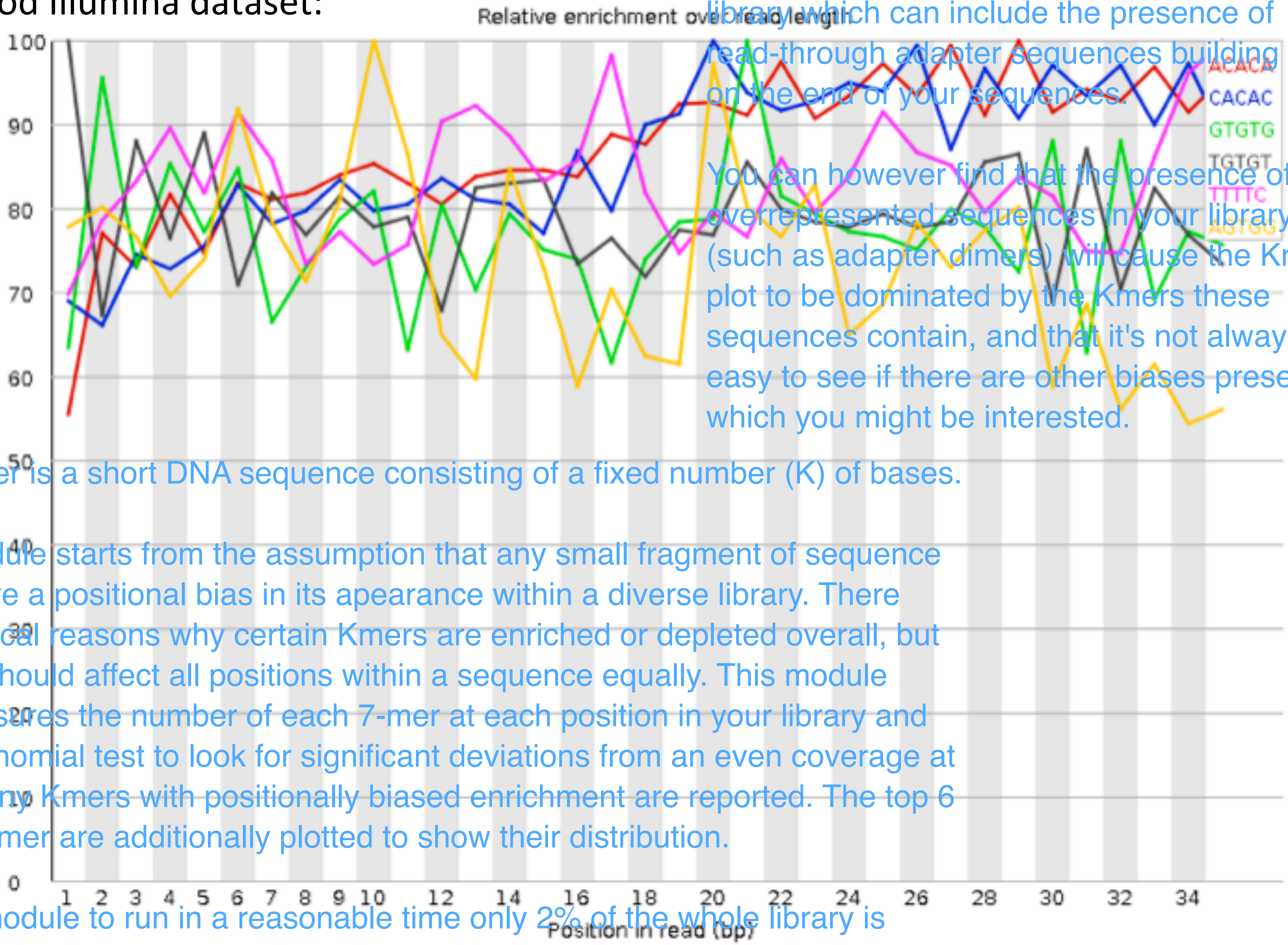
Because the duplication detection requires an exact sequence match over the whole length of the sequence any reads over 75bp in length are truncated to 50bp for the purposes of this analysis. Even so, longer reads are more likely to contain sequencing errors which will artificially increase the observed diversity and will tend to underrepresent highly duplicated sequences.

(11) FASTQC: Kmer content

The Kmer Content module will do a generic analysis of all of the Kmers in your library to find those which do not have even coverage through the length of your reads. This can find a number of different sources of bias in the library which can include the presence of read-through adapter sequences building up on the end of your sequences.

You can however find that the presence of any overrepresented sequences in your library (such as adapter dimers) will cause the Kmer plot to be dominated by the Kmers these sequences contain, and that it's not always easy to see if there are other biases present in which you might be interested.

Good Illumina dataset:



A kmer or k-mer is a short DNA sequence consisting of a fixed number (K) of bases.

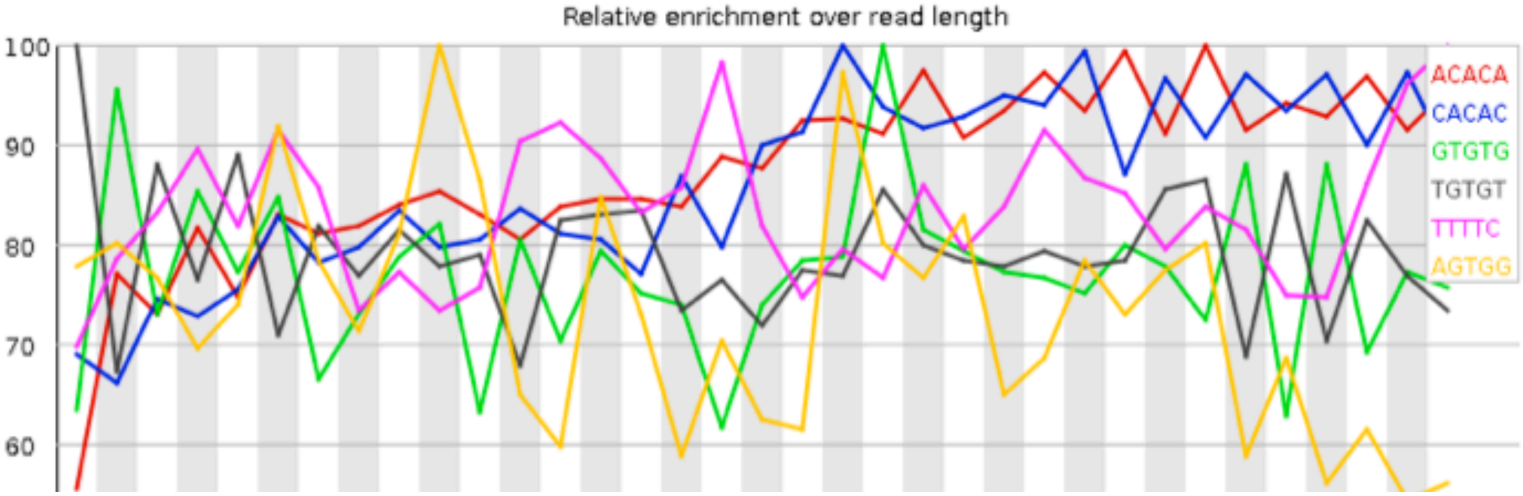
The Kmer module starts from the assumption that any small fragment of sequence should not have a positional bias in its apearance within a diverse library. There may be biological reasons why certain Kmers are enriched or depleted overall, but these biases should affect all positions within a sequence equally. This module therefore measures the number of each 7-mer at each position in your library and then uses a binomial test to look for significant deviations from an even coverage at all positions. Any Kmers with positionally biased enrichment are reported. The top 6 most biased Kmer are additionally plotted to show their distribution.

To allow this module to run in a reasonable time only 2% of the whole library is analysed and the results are extrapolated to the rest of the library. Sequences longer than 500bp are truncated to 500bp for this analysis.

http://www.slideshare.net/suryasaha/sequencing-quality-filtering?related=1

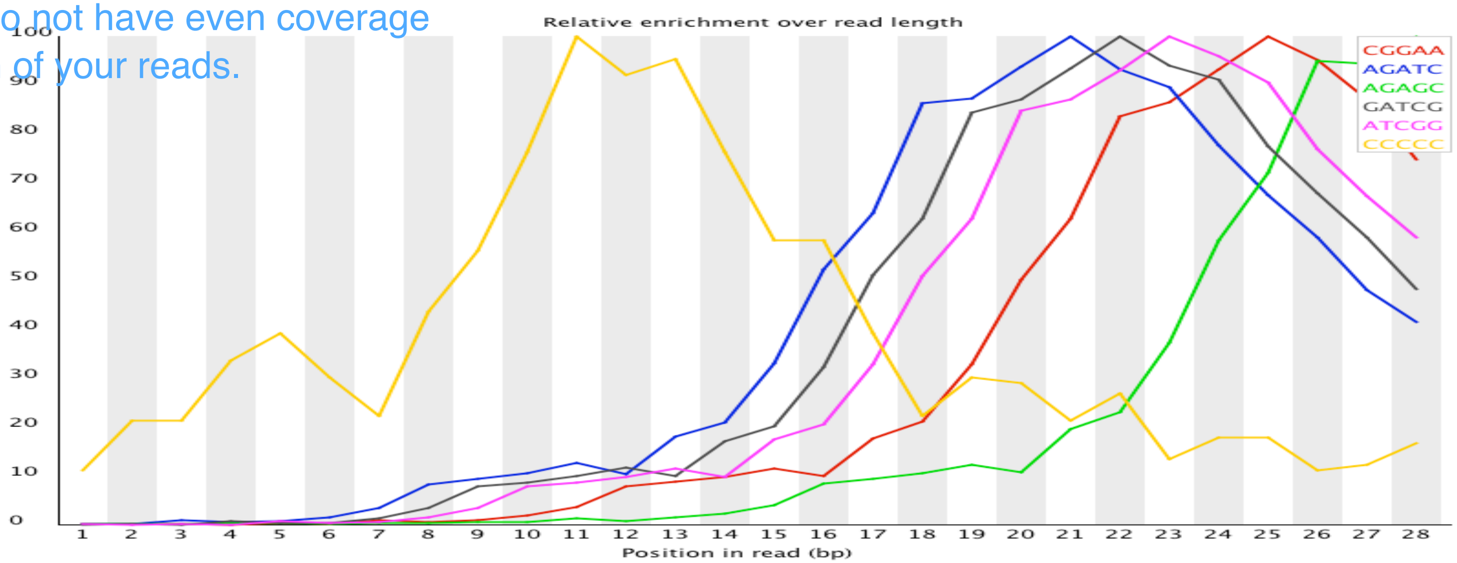
(11) FASTQC: Kmer content

GOOD



The Kmer Content module will do a generic analysis of all of the Kmers in your library to find those which do not have even coverage through the length of your reads.

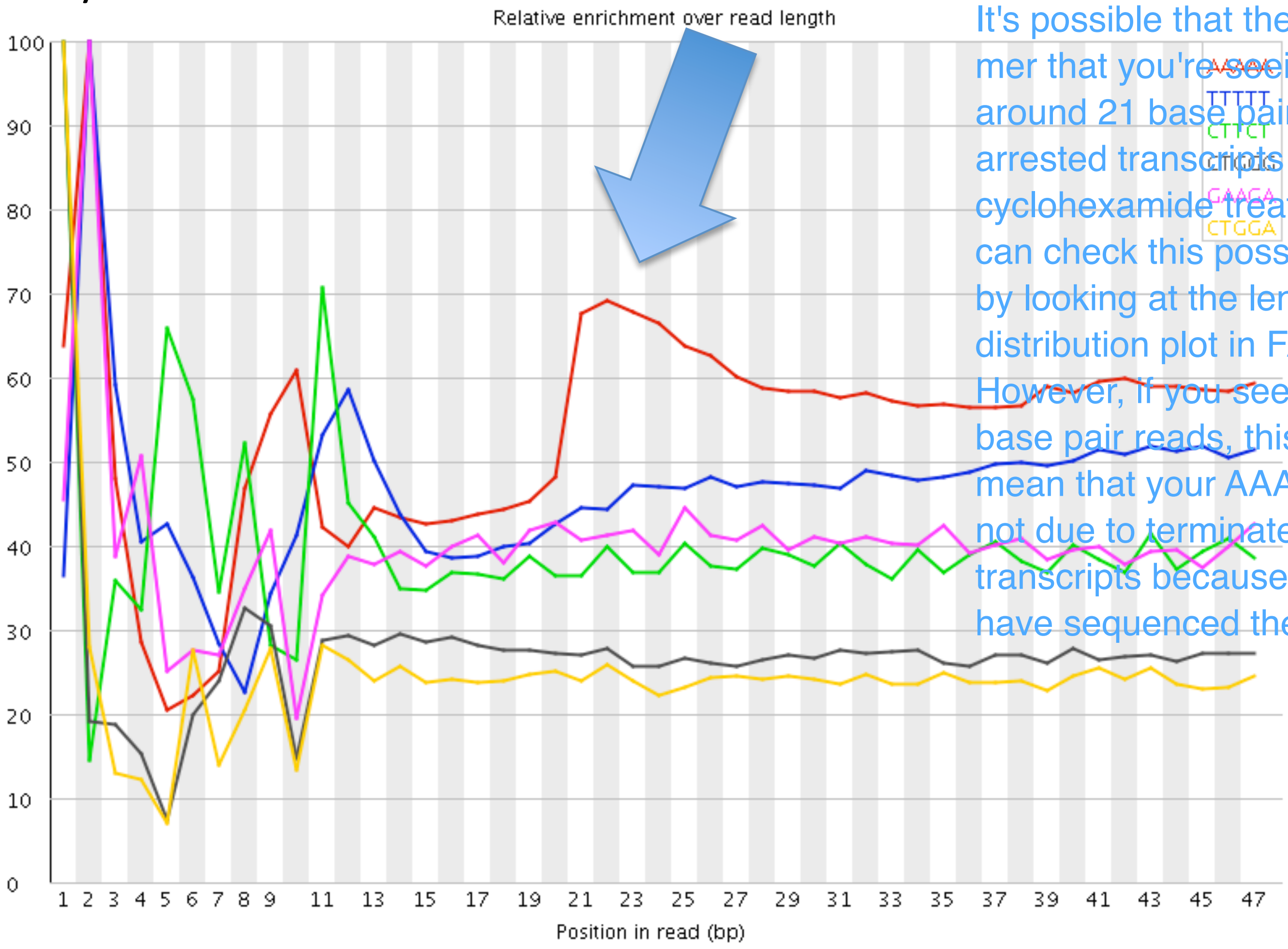
BAD



http://www.slideshare.net/suryasaha/sequencing-quality-filtering?related=1

(11) FASTQC: Kmer content

AAAA k-mer that you're seeing at around 21 base pairs are arrested transcripts caused by cyclohexamide treatment.

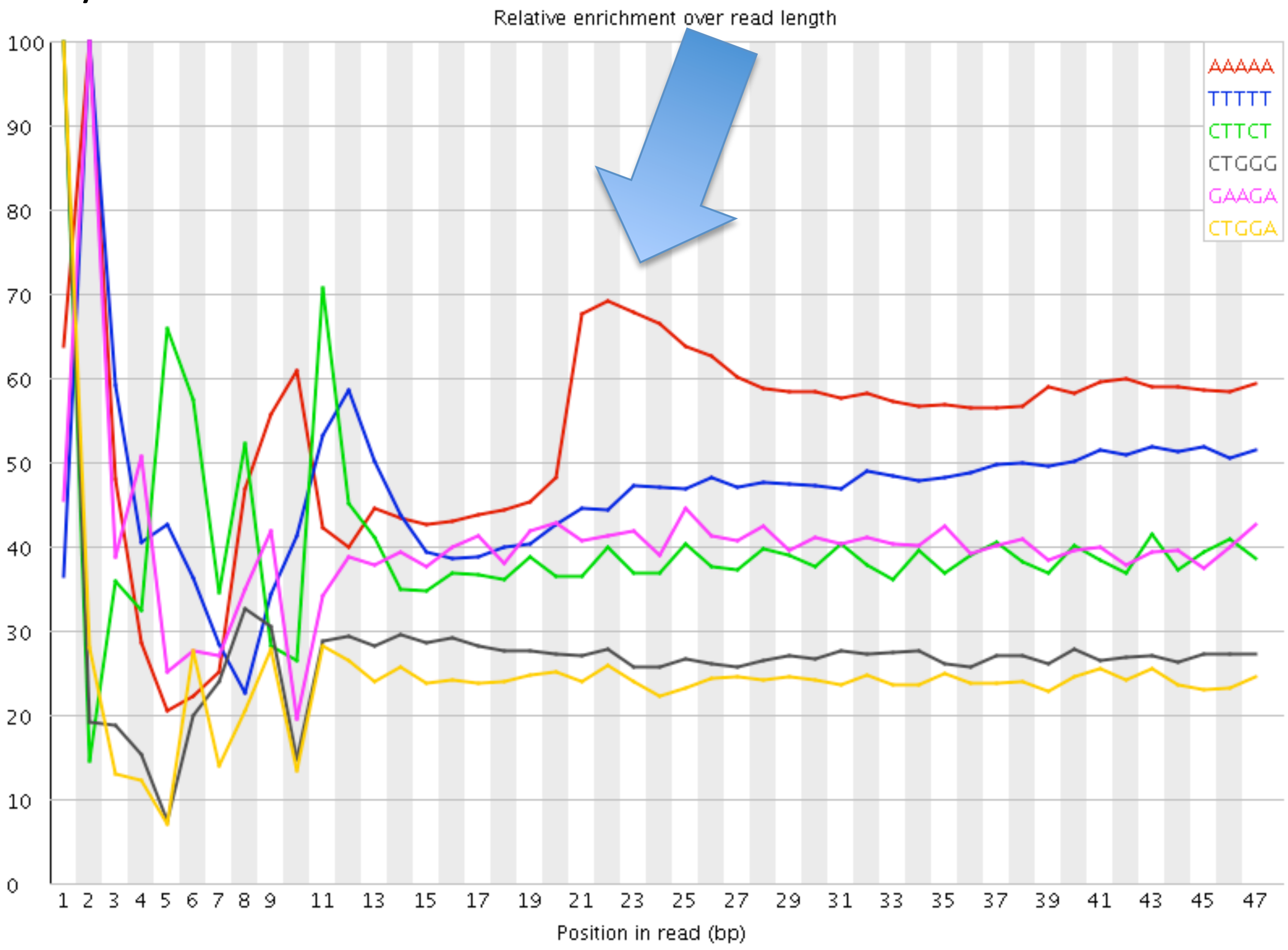


It's possible that the AAAA k-mer that you're seeing at around 21 base pairs are arrested transcripts caused by cyclohexamide treatment. You can check this possibility out by looking at the length distribution plot in FASTQC. However, if you see uniform 51 base pair reads, this doesn't mean that your AAAA k-mer is not due to terminated transcripts because you may have sequenced the 3' UTR.

http://seqanswers.com/forums/showthread.php?t=18447

(11) FASTQC: Kmer content

AAAA k-mer that you're seeing at around 21 base pairs are arrested transcripts caused by cyclohexamide treatment.



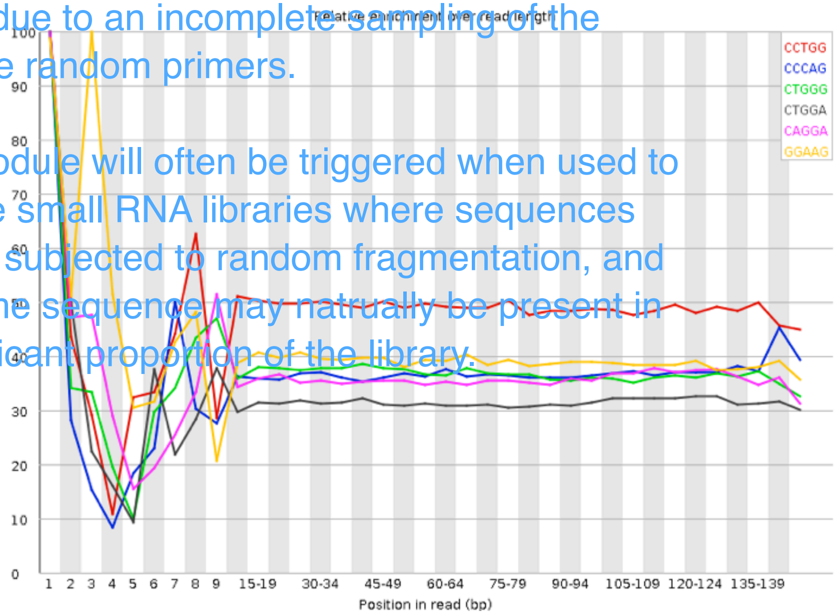
http://seqanswers.com/forums/showthread.php?t=18447

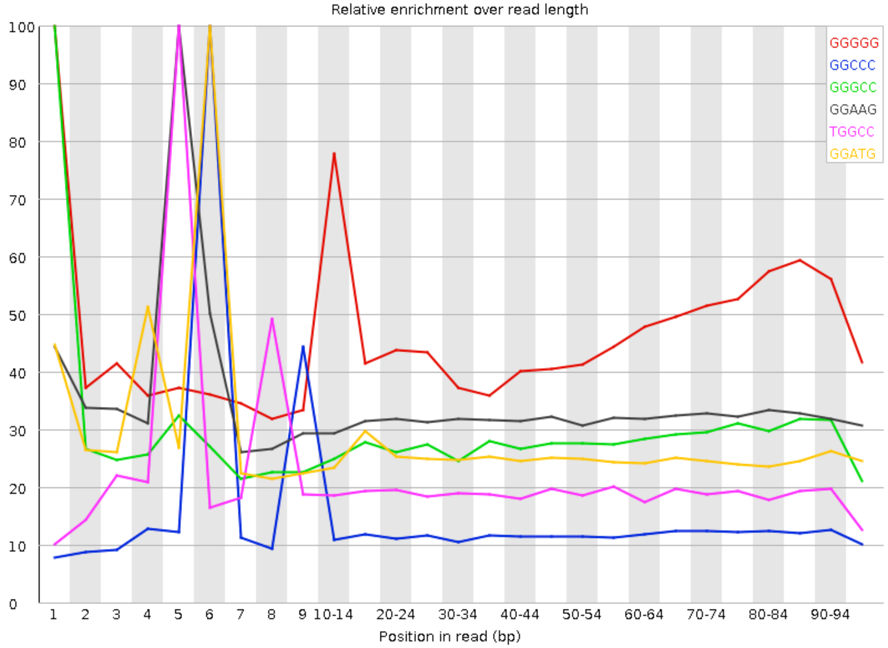
(11) FASTQC: Kmer content

"Random" hexamer primer in RNA-seq libraries
(not that random after all)

Libraries which derive from random priming will nearly always show Kmer bias at the start of the library due to an incomplete sampling of the possible random primers.

This module will often be triggered when used to analyse small RNA libraries where sequences are not subjected to random fragmentation, and the same sequence may natrually be present in a significant proportion of the library.

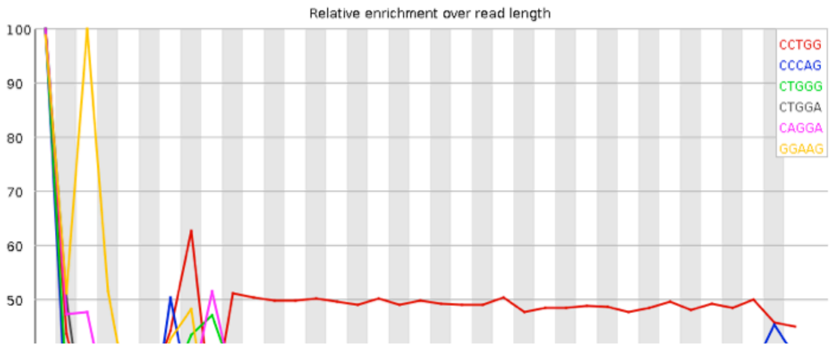


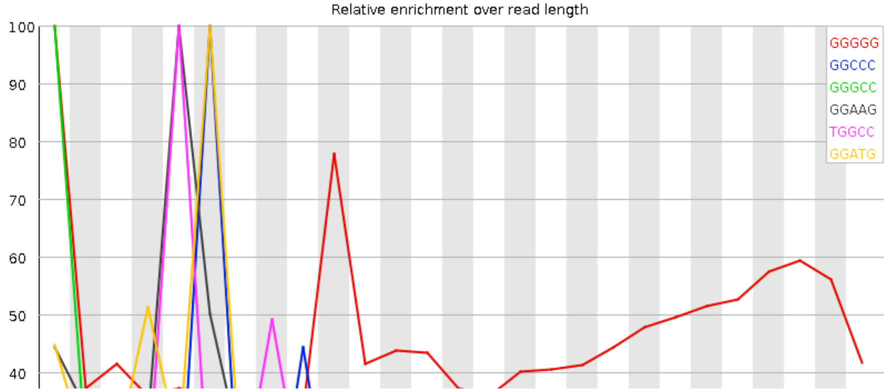


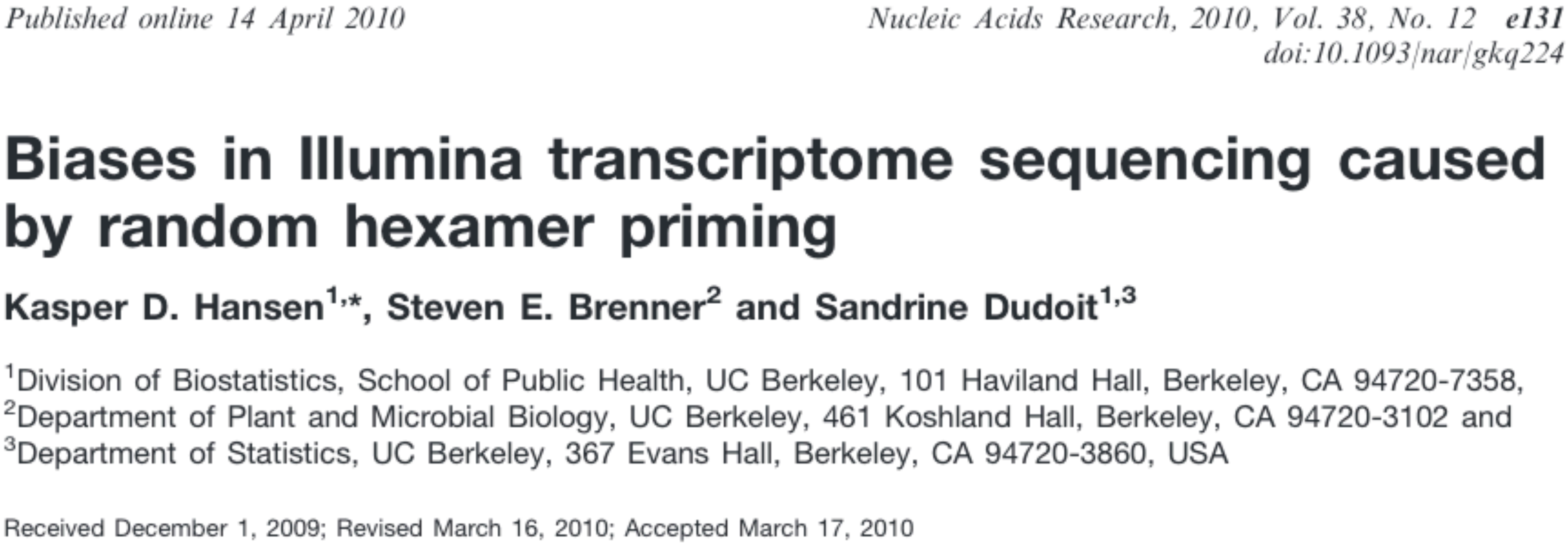
http://seqanswers.com/forums/showthread.php?t=44770&highlight=kmer+fastq
http://seqanswers.com/forums/showthread.php?t=16669

(11) FASTQC: Kmer content

“Random” hexamer primer in RNA-seq libraries
(not that random afterall)







Published online 14 April 2010

Nucleic Acids Research, 2010, Vol. 38, No. 12 ***e131***
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen[1],*, Steven E. Brenner[2] and Sandrine Dudoit[1,3]

[1]Division of Biostatistics, School of Public Health, UC Berkeley, 101 Haviland Hall, Berkeley, CA 94720-7358, [2]Department of Plant and Microbial Biology, UC Berkeley, 461 Koshland Hall, Berkeley, CA 94720-3102 and [3]Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA

Received December 1, 2009; Revised March 16, 2010; Accepted March 17, 2010

Hands on exercise:

Fastqc_sweave.pdf

Examples of FASTQC runs and preprocessing