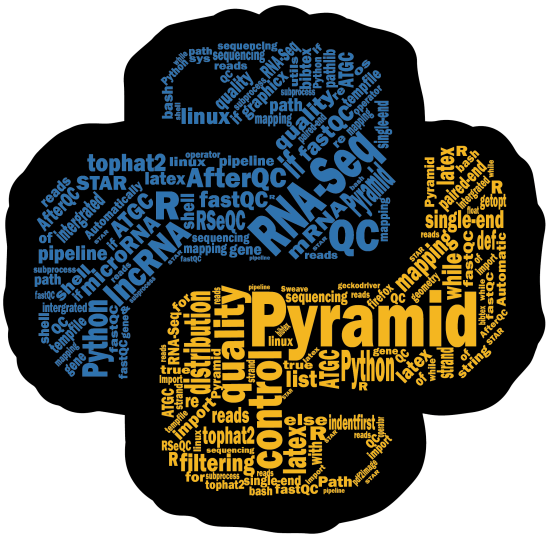
Python-based RNA-Seq pipeline for Automatically intergrated quality control



Sample-name : normal-1-A-1

Timestamp : Mon Feb 24 12:39:03 2020

Contents

Chapter 1

Introduction

1.1 Introduction

Here comes the introduction to a paper.

1.2 The result file of mapping.

Left reads:
Input : 8833067
Mapped : 3359138 (38.0% of input)
of these: 175874 (5.2%) have multiple alignments (3933 have >20)
Right reads:
Input : 8833067
Mapped : 6435414 (72.9% of input)
of these: 319524 (5.0%) have multiple alignments (3935 have >20)
55.4% overall read mapping rate.
Aligned pairs: 2784010
of these: 144059 (5.2%) have multiple alignments
124461 (4.5%) are discordant alignments
30.1% concordant pair alignment rate.

Chapter 2

FastQC Result

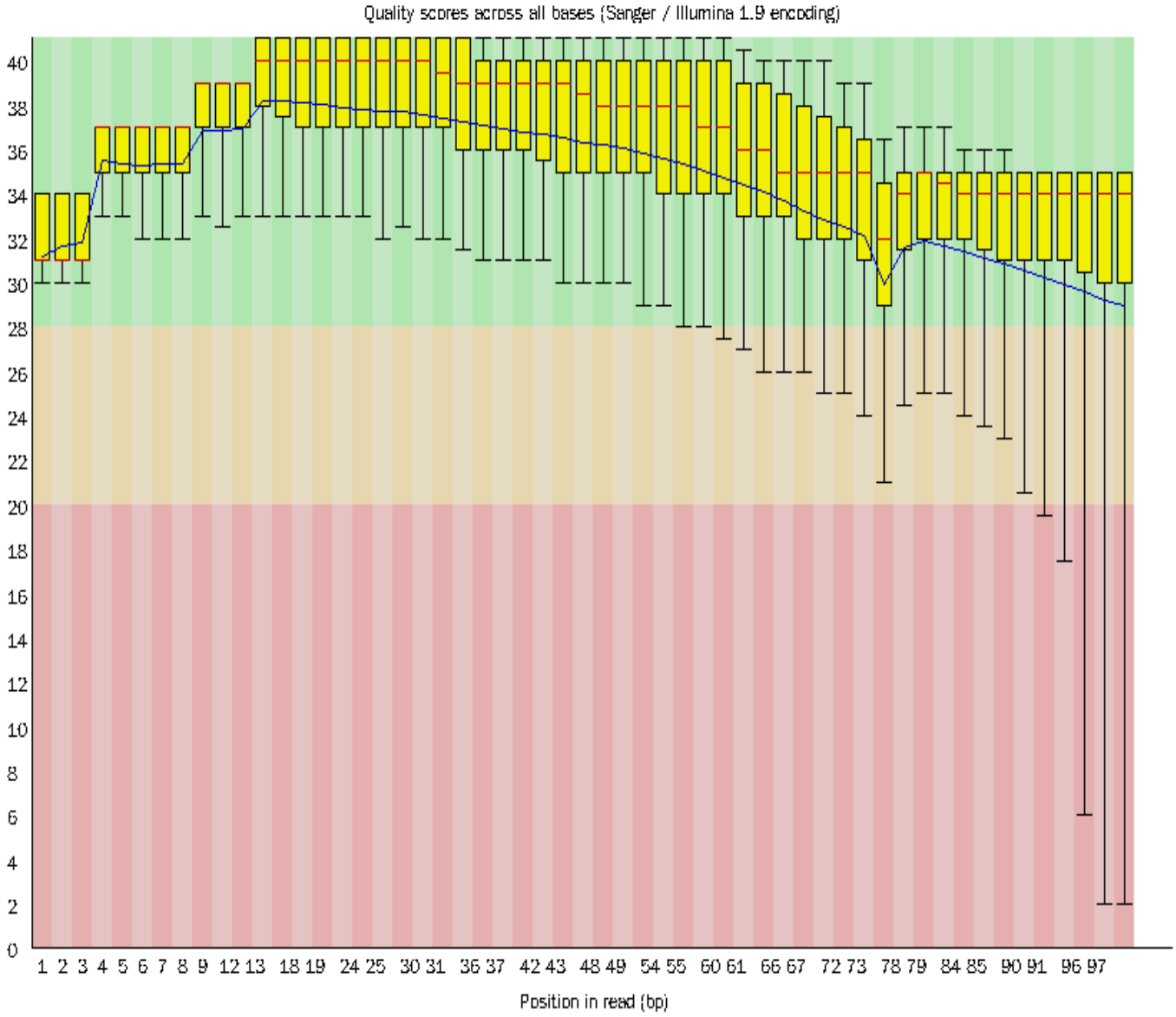
2.1 Per base sequence quality

This view shows an overview of all the base quality value ranges for each position in the FastQ file.The y-axis on the graph shows the quality scores. 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 title of the graph will describe the encoding FastQC thinks your file used.

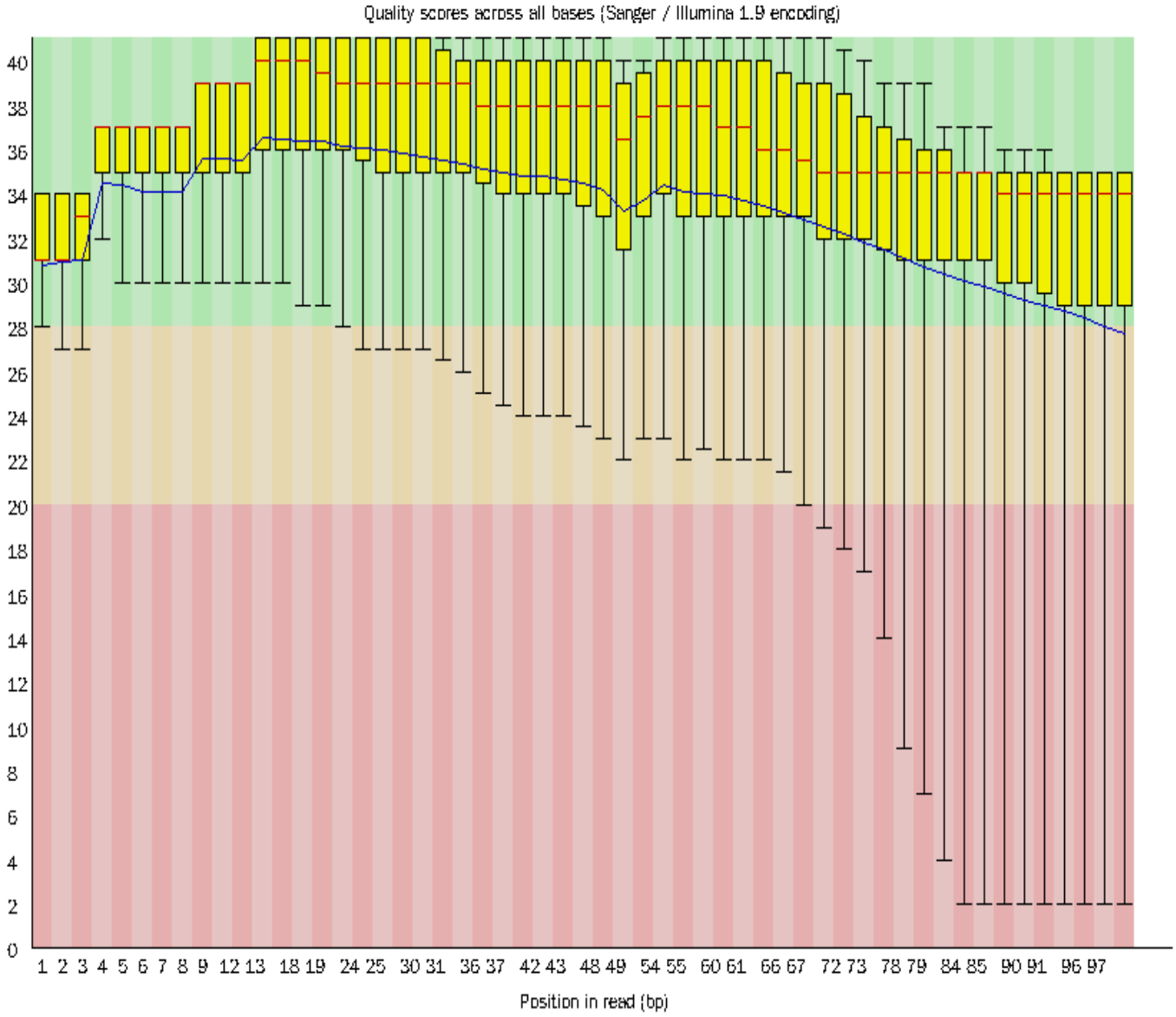
For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows:

- 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% and 90
- The blue line represents the mean quality.

This view shows an overview of all the base quality value ranges for each position in the FastQ file of R1.



This view shows an overview of all the base quality value ranges for each position in the FastQ file of R2.



Chapter 3

AfterQC results

3.1 AfterQC summary

This view is a summary of the digital results of the quality control analysis performed by the AfterQC software on the RAN-seq data input by the user. Content contains AfterQC Version, sequencing, total reads, filtered out reads, total bases, filtered out bases, estimated seq error, adapter trimmed reads, adapter trimmed bases, auto trimming, etc.

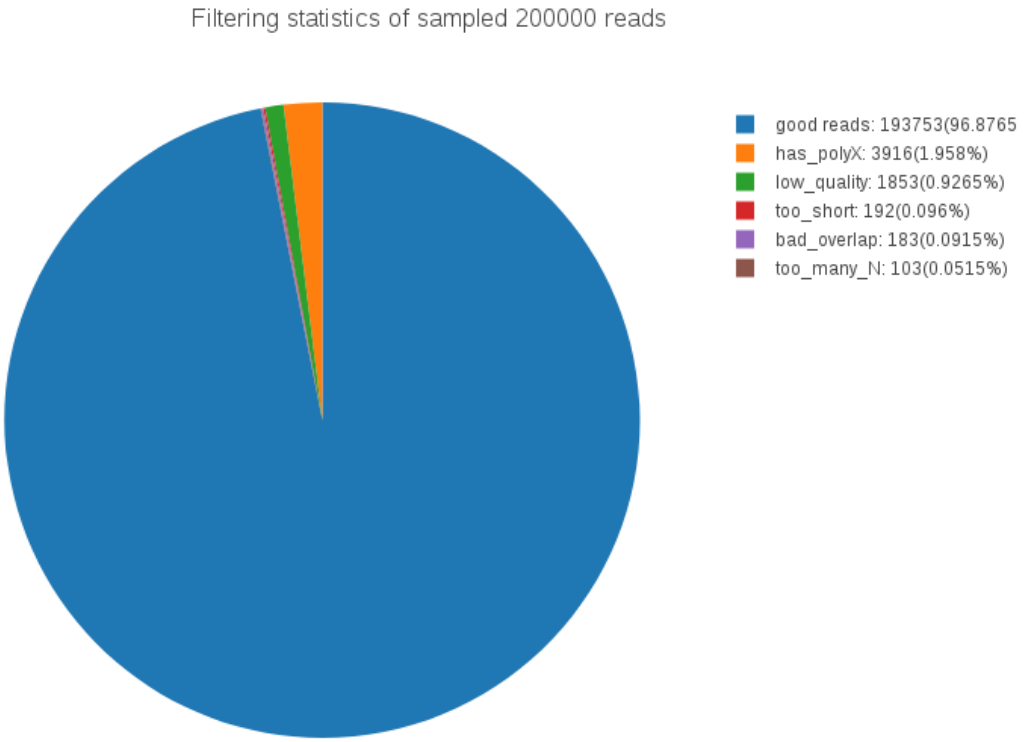
AfterQC Version:	0.9.6
sequencing:	2*100 pair end
total reads:	195.312 K
filtered out reads:	6247.000 (3.123%)
total bases:	19.073 M
filtered out bases:	1520004.000 (7.600%)
estimated seq error:	0.105%
adapter trimmed reads:	4.887 K
adapter trimmed bases:	261.443 K
auto trimming	front:4, tail:0 (use -f0 -t0 to disable)

3.2 Good reads and bad reads after filtering

This view is the result of AfterQC software filtering the user-input ran-seq data and summarizing the filtering results into a pie chart. The total number of reads is the statistical data of the 200,000 readings of the filtered samples. The content includes the circular pie graph and the legend corresponding to the figure, which are synthesized from good reads, has polyX reads, low quality reads, too short reads, bad overlap reads and too many N reads.

In the figure, blue indicates good reads, yellow indicates has ployX reads, green indicates low quality reads, red indicates too short reads, purple indicates bad overlap reads, and brown indicates too many N reads.

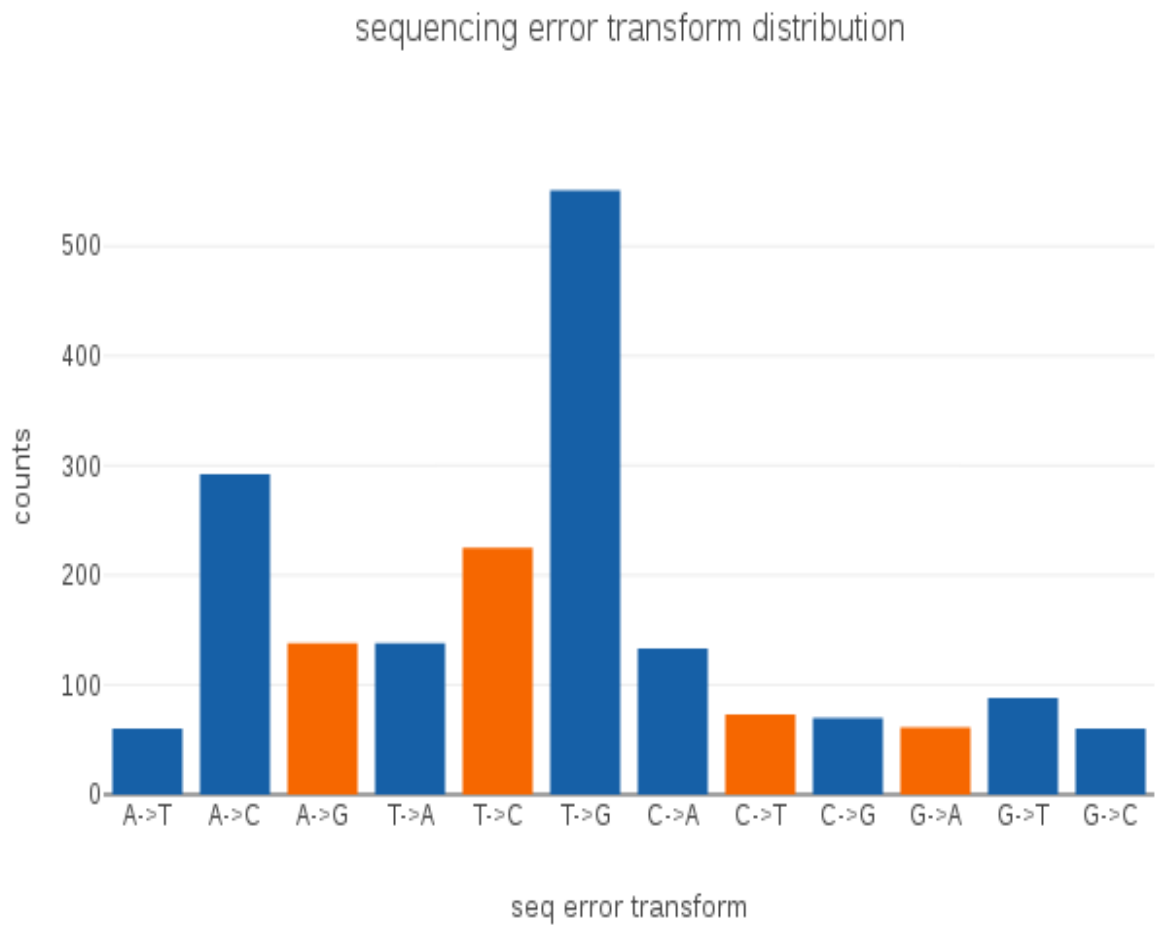
From the figure, it can be visually seen whether the various indicators of the sequencing result input by the user are up to standard.



3.3 Sequence error distribution

For each pair of pair-end sequenced FastQ files, AfterQC estimates such sequencing error rate and profiles the sequencing error transform distribution.error distribution is clearly sequencing platform dependent, different sequencing platforms have different error patterns, while the same sequencing platform's different sequencing runs share similar patterns.

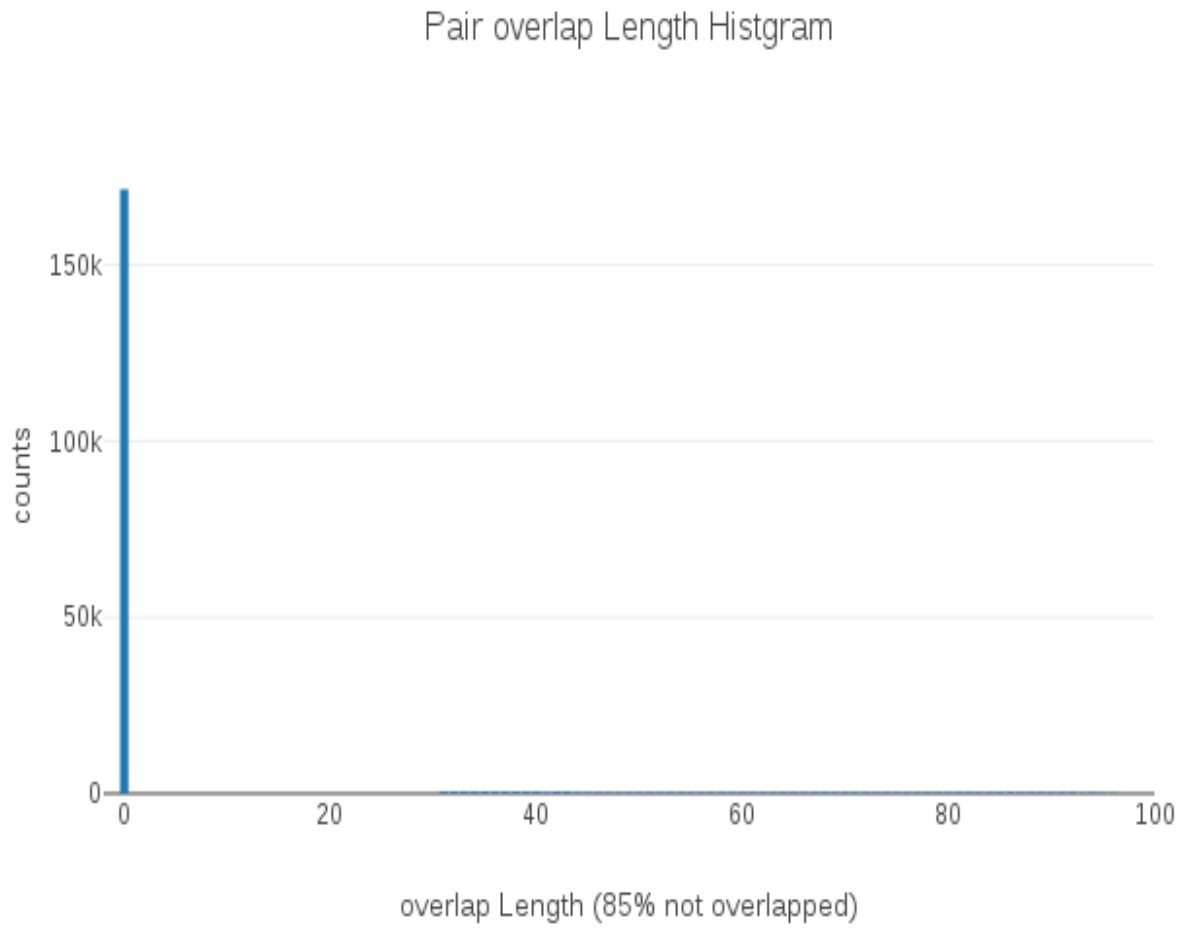
The view is sequence error distribution.Values in X-axis represent the sequencing error, and the values in Y-axis represent the counts calculated from a pair of FastQ files.



3.4 Overlap length distribution

Different from most tools, AfterQC analyses the overlapping of paired sequences for pair-end sequencing data. Each DNA template is sequenced twice in forward and reverse directions. When the DNA template length is less than twice of the sequencing length, the pair of reads will be overlapped. Note that each base in the overlapping region is actually sequenced twice, so the inconsistency of these pairs may reflect the sequencing errors.

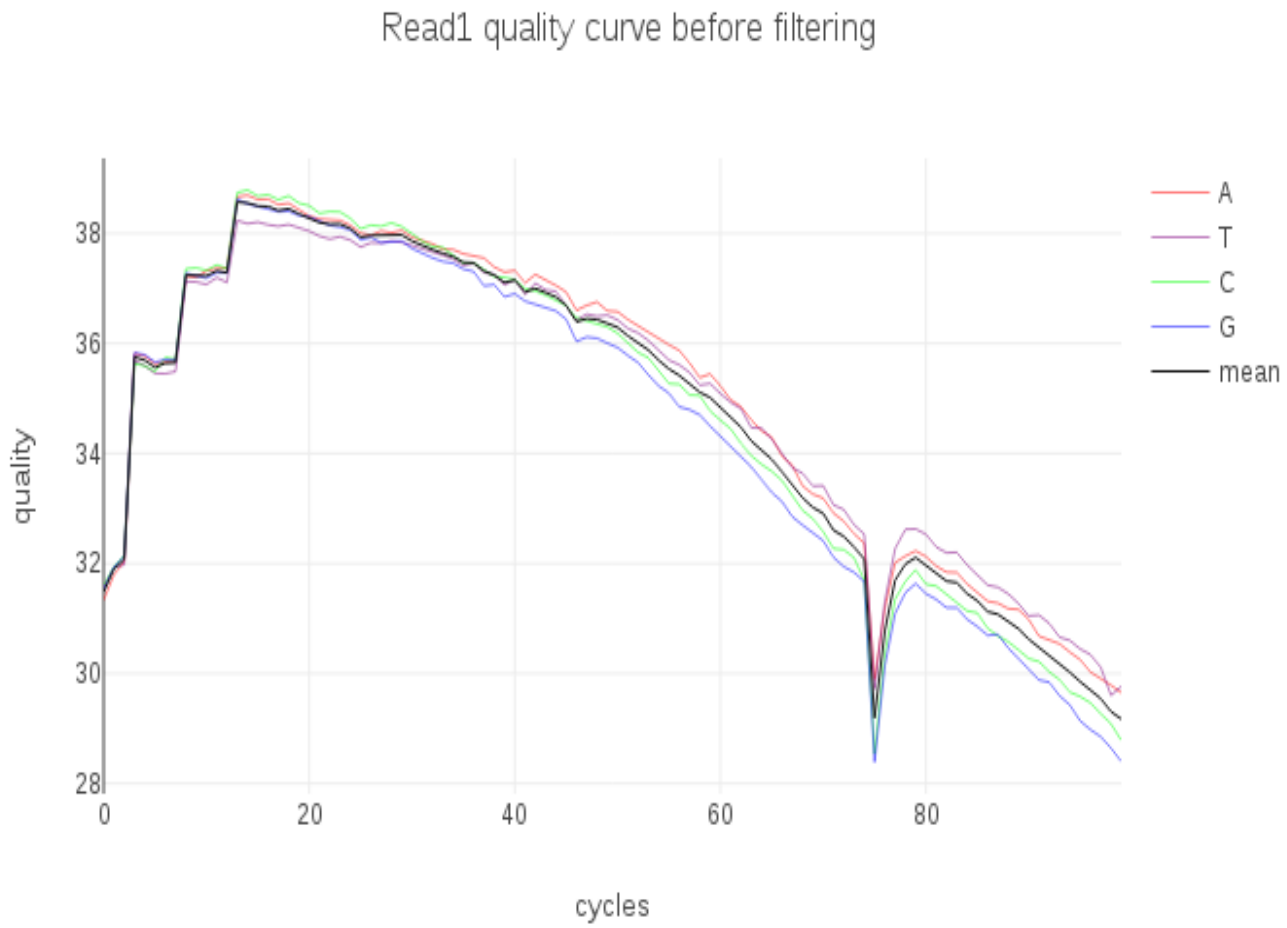
The view is a histogram drawn from the result of the overlap of the AfterQC analysis pairing sequence with the paired end sequencing data. The value in the X axis represents the overlap length, and the value in the Y axis represents the count of the overlap length. And the proportion of the length without overlap is included in the axis title of the X-axis.



3.5 Read1 quality curve before filtering

This view is the curve that quality of reads1 before filtering. It not only plots the quality of A, T, C, and G for each locus in the user-entered sequencing results file, but also plots the mass average of four bases per locus. The red curve represents the result of A base, the purple curve represents the result of T base , the green curve represents the result of C base , the blue curve represents the result of G base , and the black curve represents the result of the average.

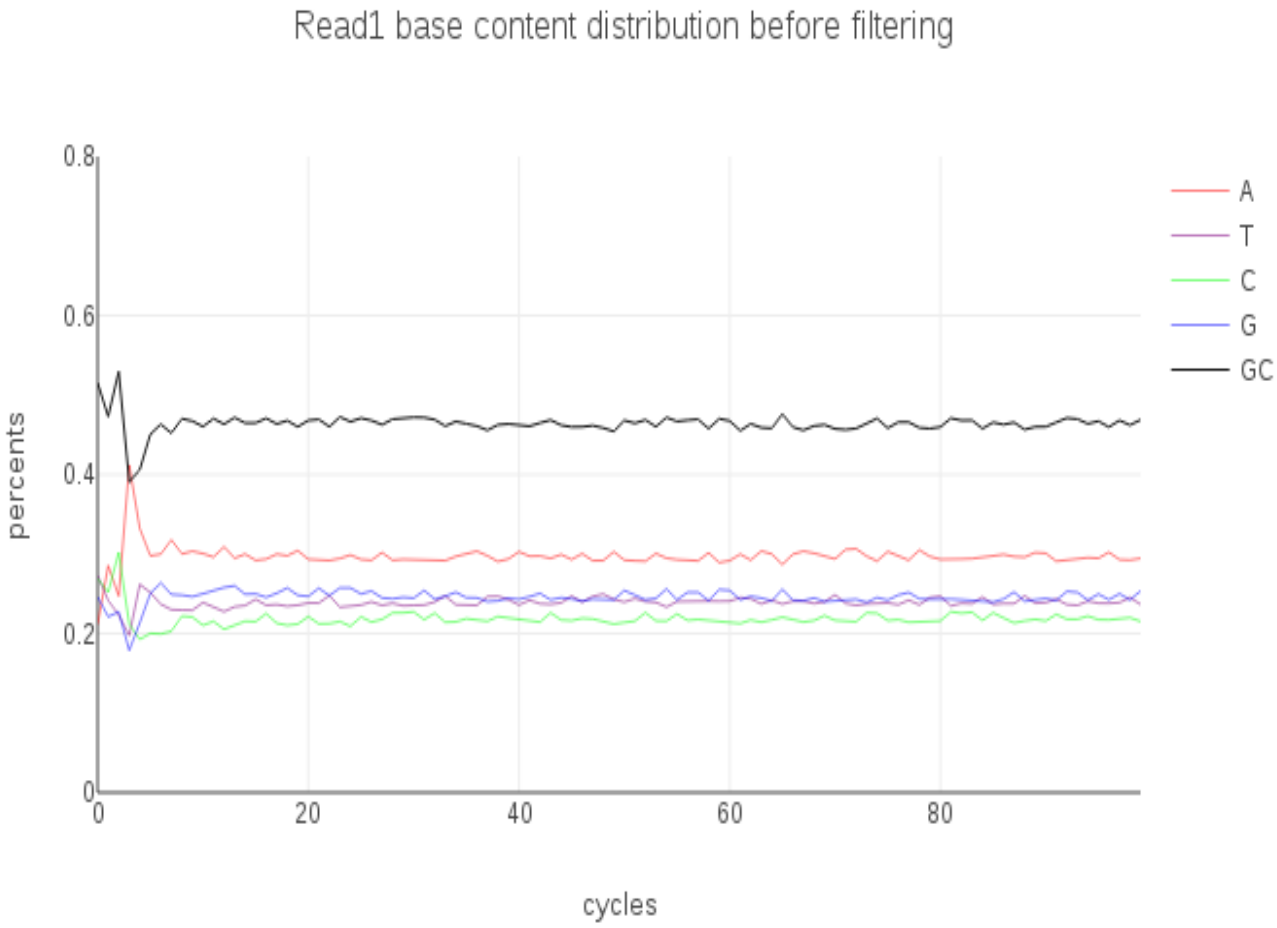
The abscissa represents the locus and the ordinate represents the mass fraction.



3.6 Read1 base content distribution before filtering

This view is the curve that the base content distribution of read1 before filtering. It plots the proportion of A, T, C, G, and GC at each locus in the sequencing results file entered by the user. The red curve indicates the result of the A base, the purple curve indicates the result of the T base, the green curve indicates the result of the C base, the blue curve indicates the result of the G base, and the black curve indicates the result of the sum of the GC.The abscissa indicates the locus and the ordinate indicates the proportion.

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.[1]

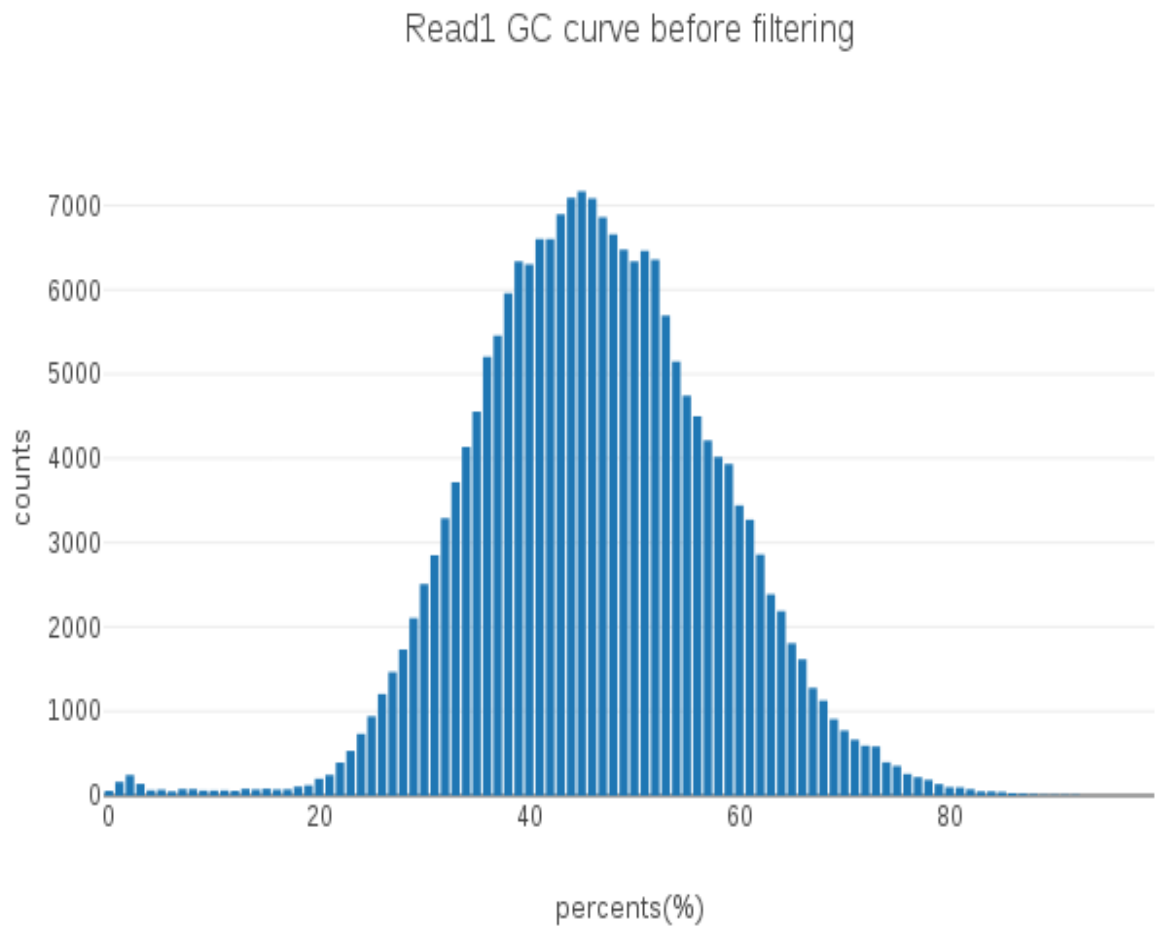


3.7 Read1 GC curve before filtering

This view measures the GC content across the whole length of each sequence in the fastq file of Read1 of before filtering.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.

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.

Values in X-axis represent the GC content(%), and the values in Y-axis represent the counts.

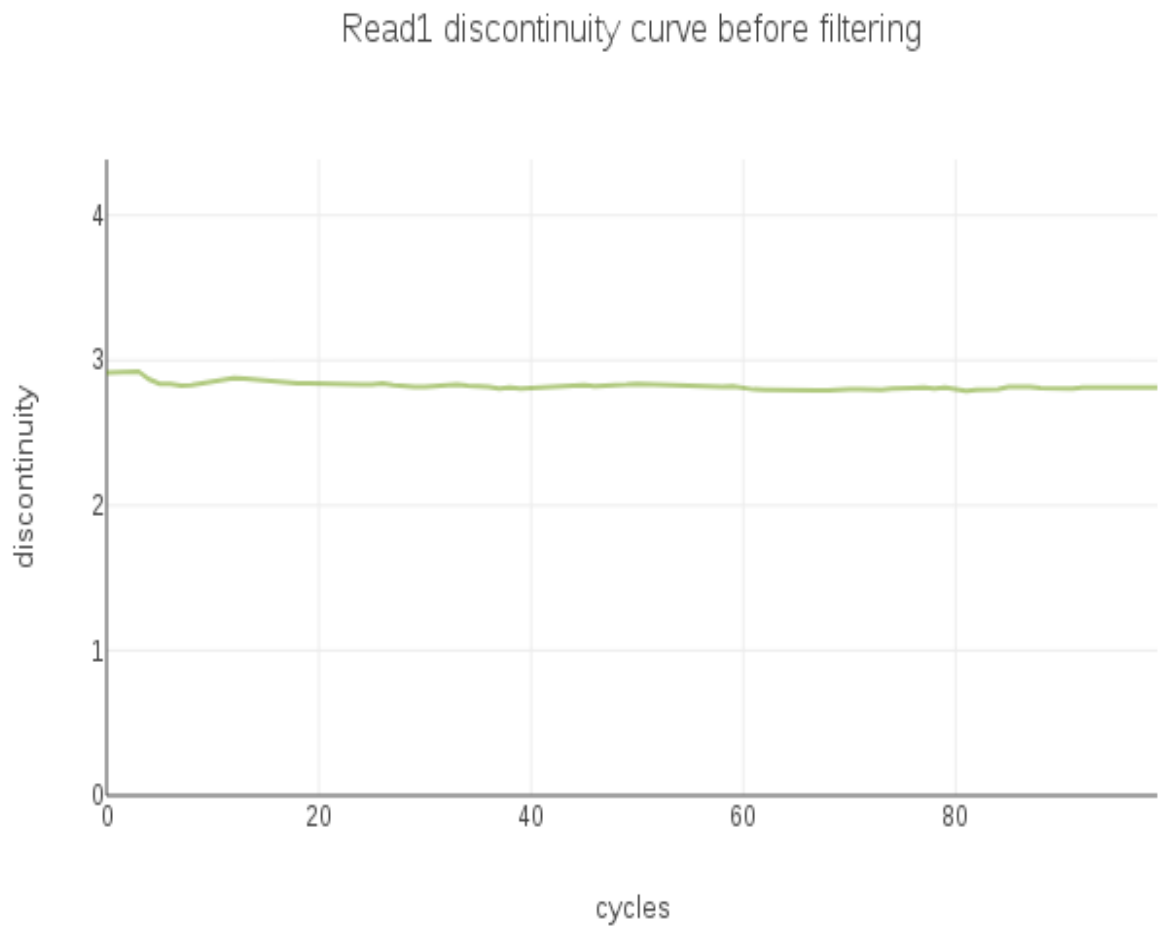


3.8 Read1 per base discontinuity before filtering

Besides normal per-cycle base content and quality profiling, AfterQC implements a methods to give more information about per-cycle discontinuity profiling to reflect sequencing quality instability.

The mean discontinuity should be more or less stable for all sequencing cycles. For a short window of sequencing cycles, it use the average discontinued base number in this window to calculate the discontinuity. If discontinuity drops down significantly cycle by cycle, it usually reflects a sequencing issue, which may be caused by the per-cycle washing process not working well.

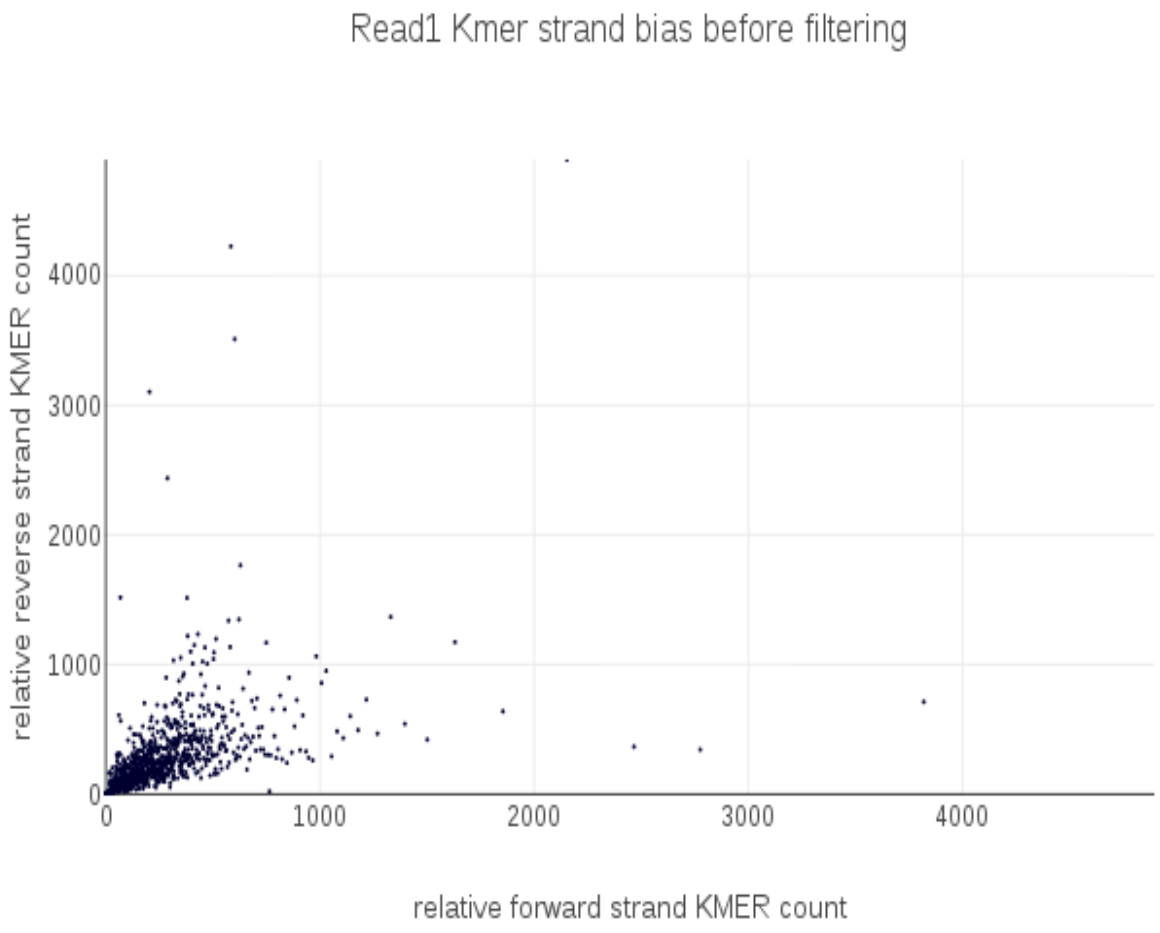
The view is a curve that the per base discontinuity before filtering of reads1. The abscissa indicates cycles and the ordinate indicates discontinuity.



3.9 Read1 kmer strand bias before filtering

Besides normal per-cycle base content and quality profiling, AfterQC implements novel method to give more information about sequencing quality: strand bias profiling to reflect amplification bias. The method is based on a hypothesis: if the DNA amplification process and sequencing process have only little non-uniformity, the repeat count of a short K-MER should be close to the repeat count of its reverse complement. So we plot each K-MER and its reverse complement's counts, and check whether most points are near the line y=x.

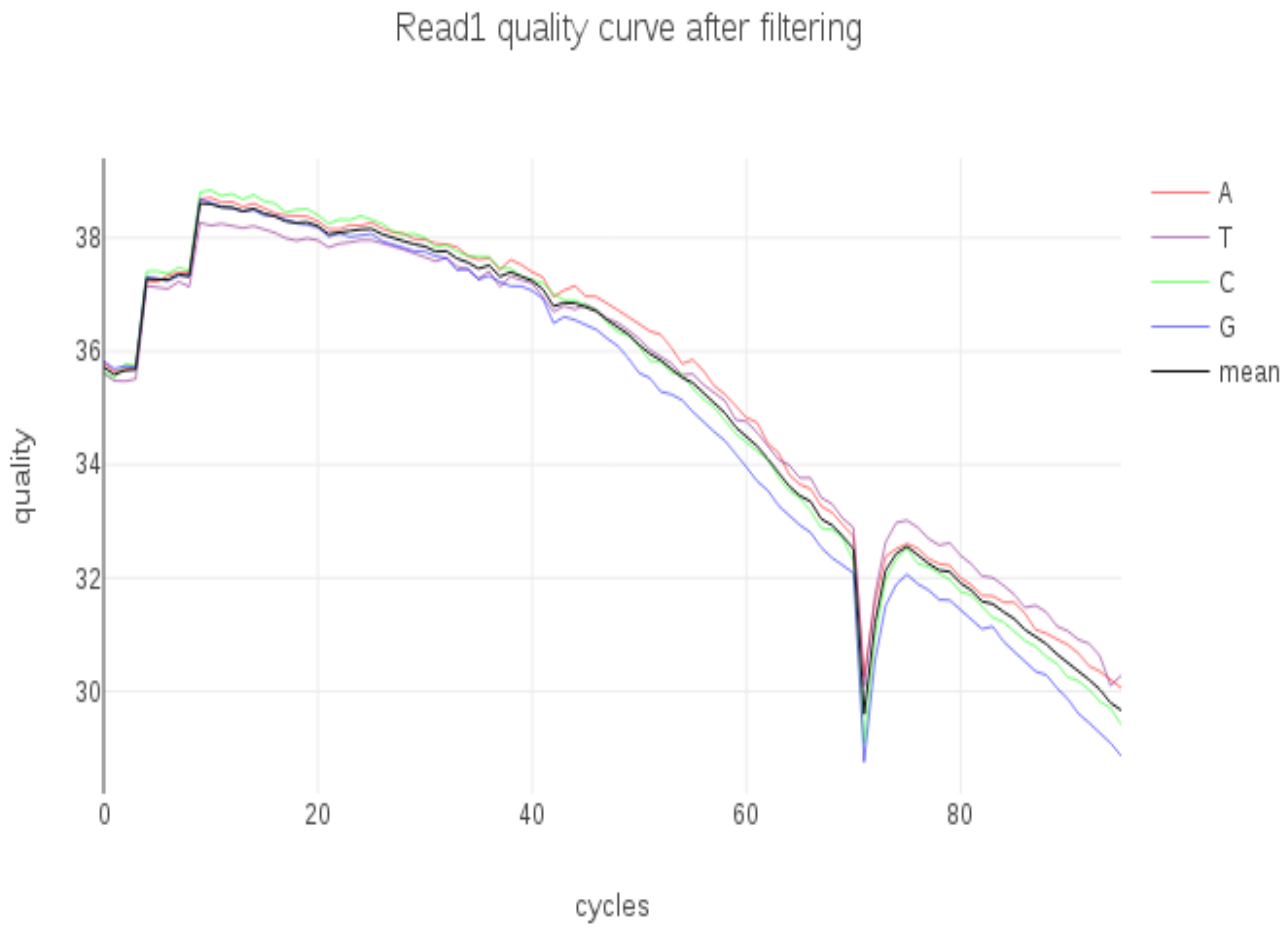
The view is kmer strand bias before filtering of R1.X-axis is about the counts of relative forward strand K-MERs, while the Y-axis is about relative reverse ones.



3.10 Read1 quality curve after filtering

This view is the curve that quality of reads1 after filtering. It not only plots the quality of A, T, C, and G for each locus in the user-entered sequencing results file, but also plots the mass average of four bases per locus. The red curve represents the result of A base, the purple curve represents the result of T base , the green curve represents the result of C base , the blue curve represents the result of G base , and the black curve represents the result of the average.

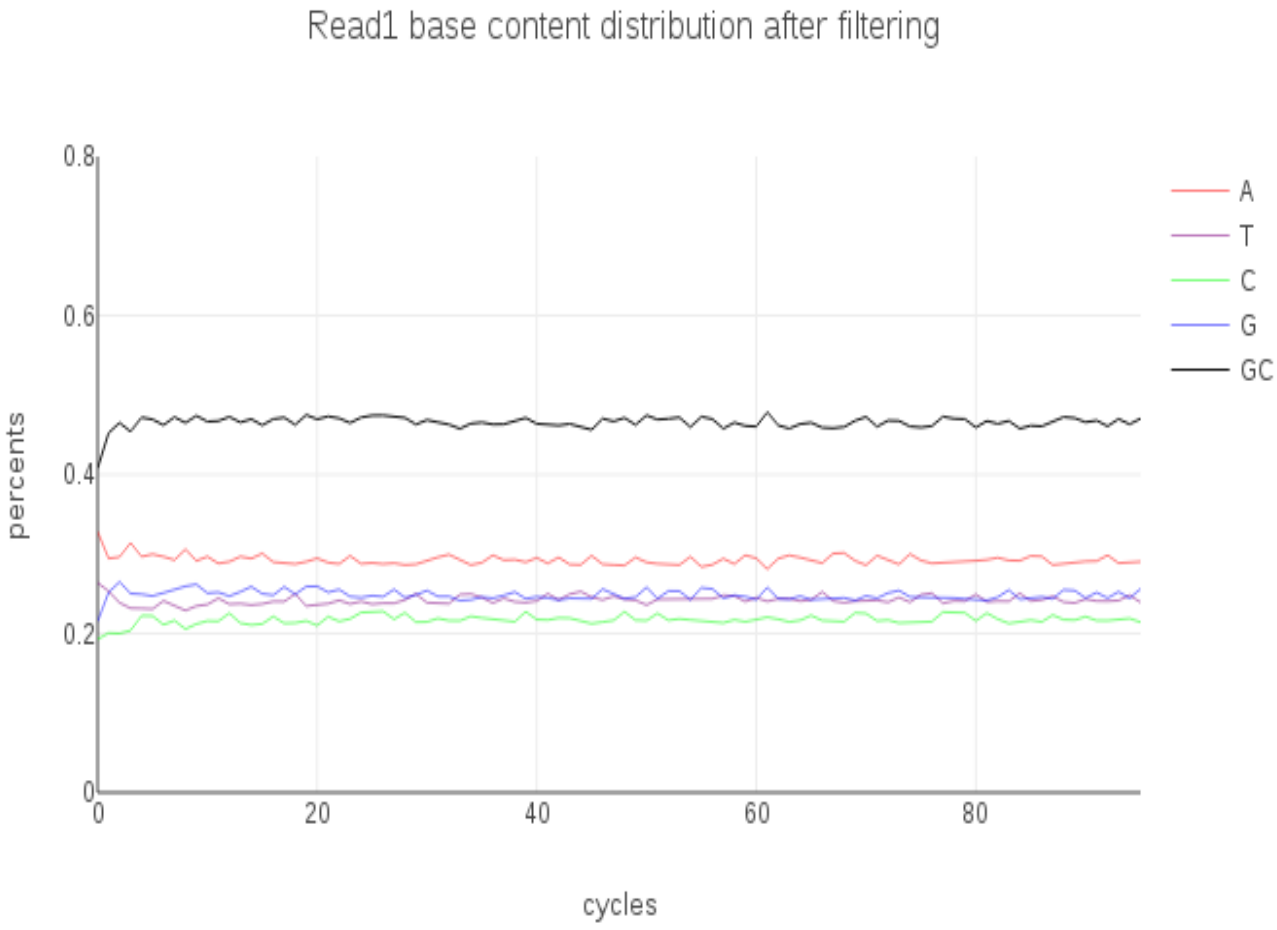
The abscissa represents the locus and the ordinate represents the mass fraction.



3.11 Read1 base content distribution after filtering

This view is the curve that the base content distribution of read1 after filtering. It plots the proportion of A, T, C, G, and GC at each locus in the sequencing results file entered by the user. The red curve indicates the result of the A base, the purple curve indicates the result of the T base, the green curve indicates the result of the C base, the blue curve indicates the result of the G base, and the black curve indicates the result of the sum of the GC.The abscissa indicates the locus and the ordinate indicates the proportion.

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.

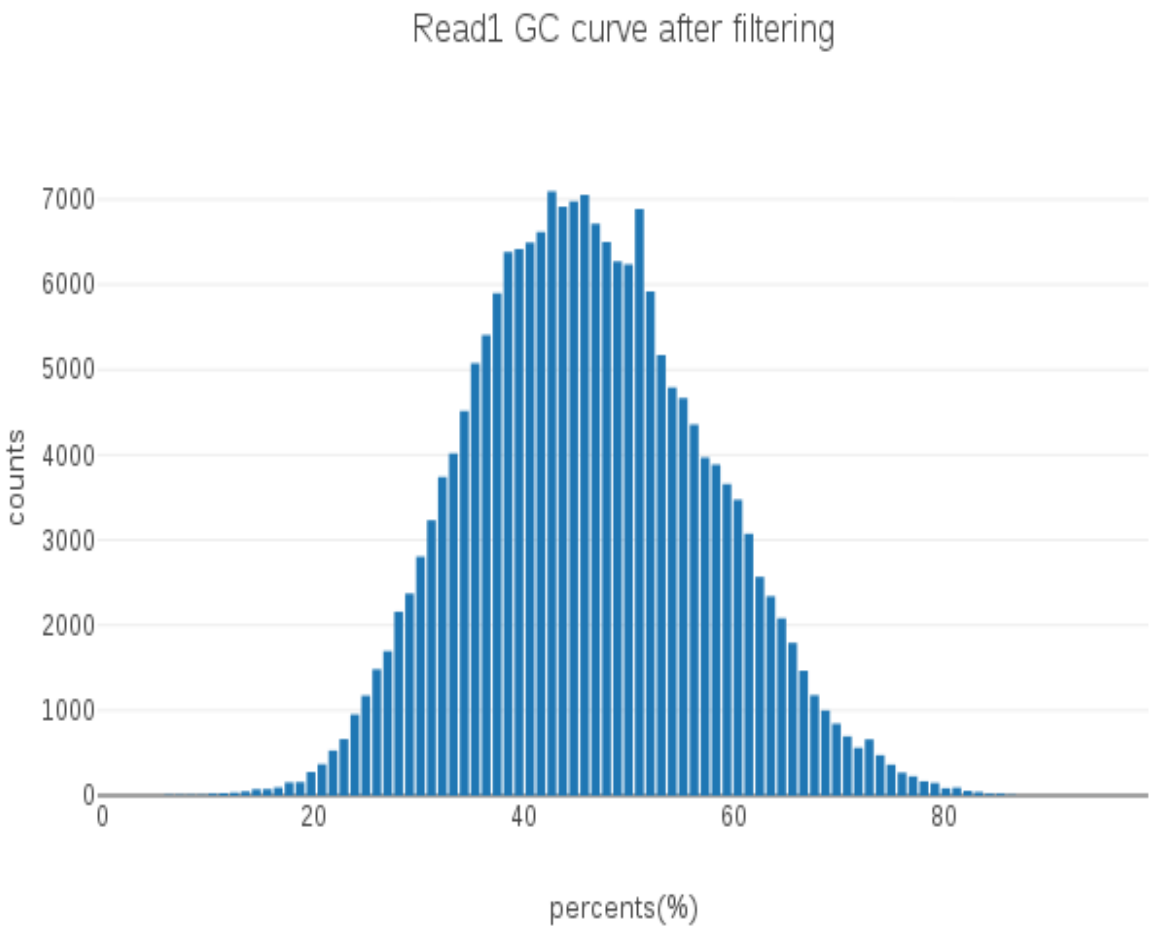


3.12 Read1 GC curve after filtering

This view measures the GC content across the whole length of each sequence in the fastq file of Read1 of after filtering.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.

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.

Values in X-axis represent the GC content(%), and the values in Y-axis represent the counts.

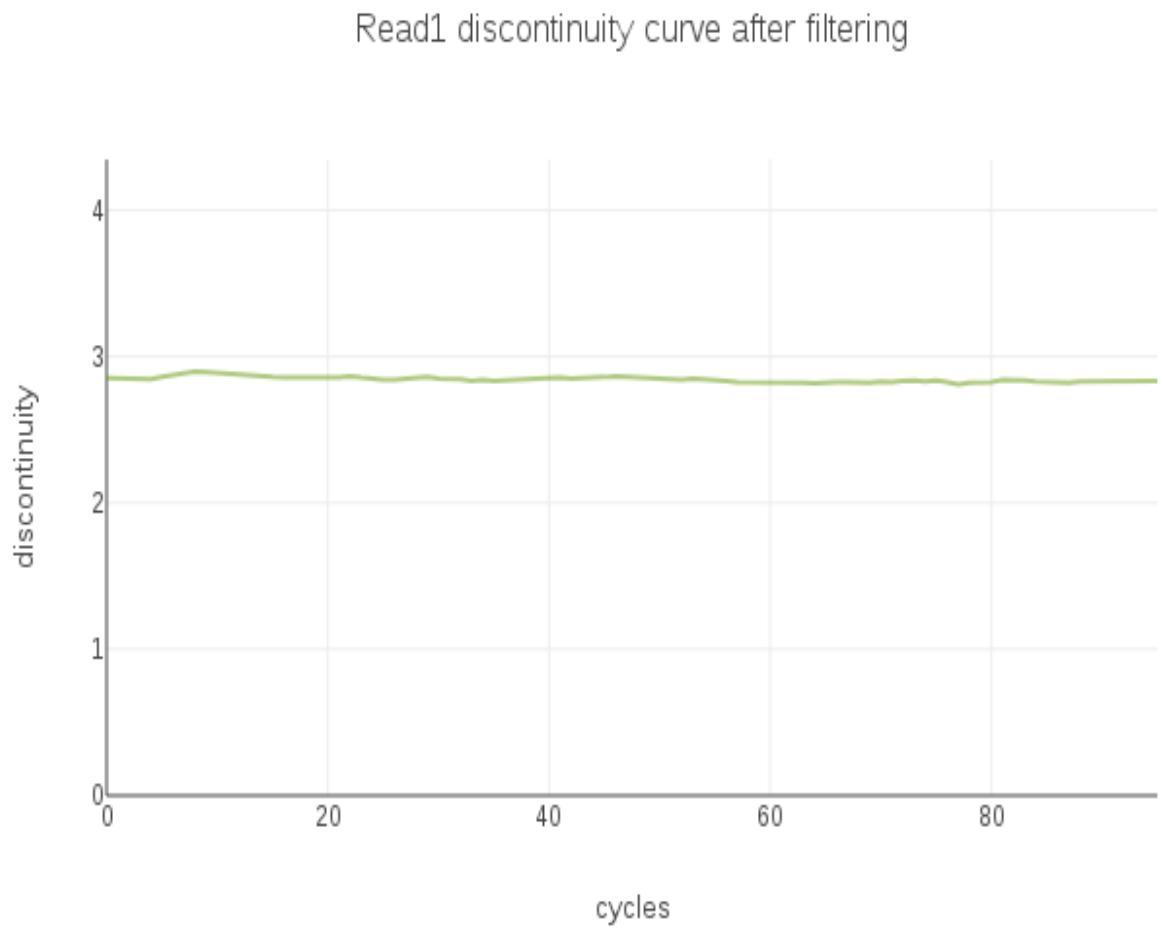


3.13 Read1 per base discontinuity after filtering

Besides normal per-cycle base content and quality profiling, AfterQC implements a methods to give more information about per-cycle discontinuity profiling to reflect sequencing quality instability.

The mean discontinuity should be more or less stable for all sequencing cycles. For a short window of sequencing cycles, it use the average discontinued base number in this window to calculate the discontinuity. If discontinuity drops down significantly cycle by cycle, it usually reflects a sequencing issue, which may be caused by the per-cycle washing process not working well.

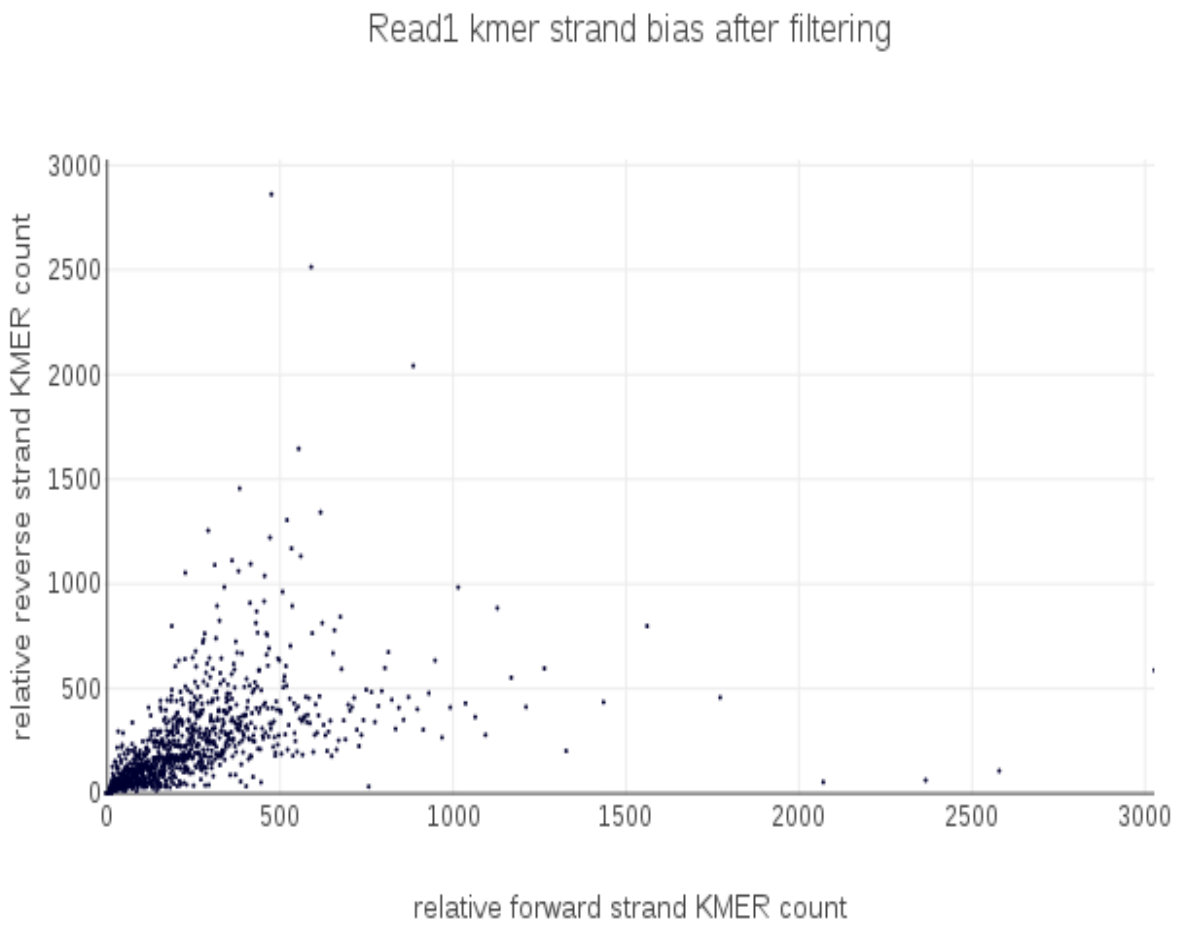
The view is a curve that the per base discontinuity after filtering of reads1. The abscissa indicates cycles and the ordinate indicates discontinuity.



3.14 Read1 kmer strand bias after filtering

Besides normal per-cycle base content and quality profiling, AfterQC implements novel method to give more information about sequencing quality: strand bias profiling to reflect amplification bias. The method is based on a hypothesis: if the DNA amplification process and sequencing process have only little non-uniformity, the repeat count of a short K-MER should be close to the repeat count of its reverse complement. So we plot each K-MER and its reverse complement's counts, and check whether most points are near the line y=x.

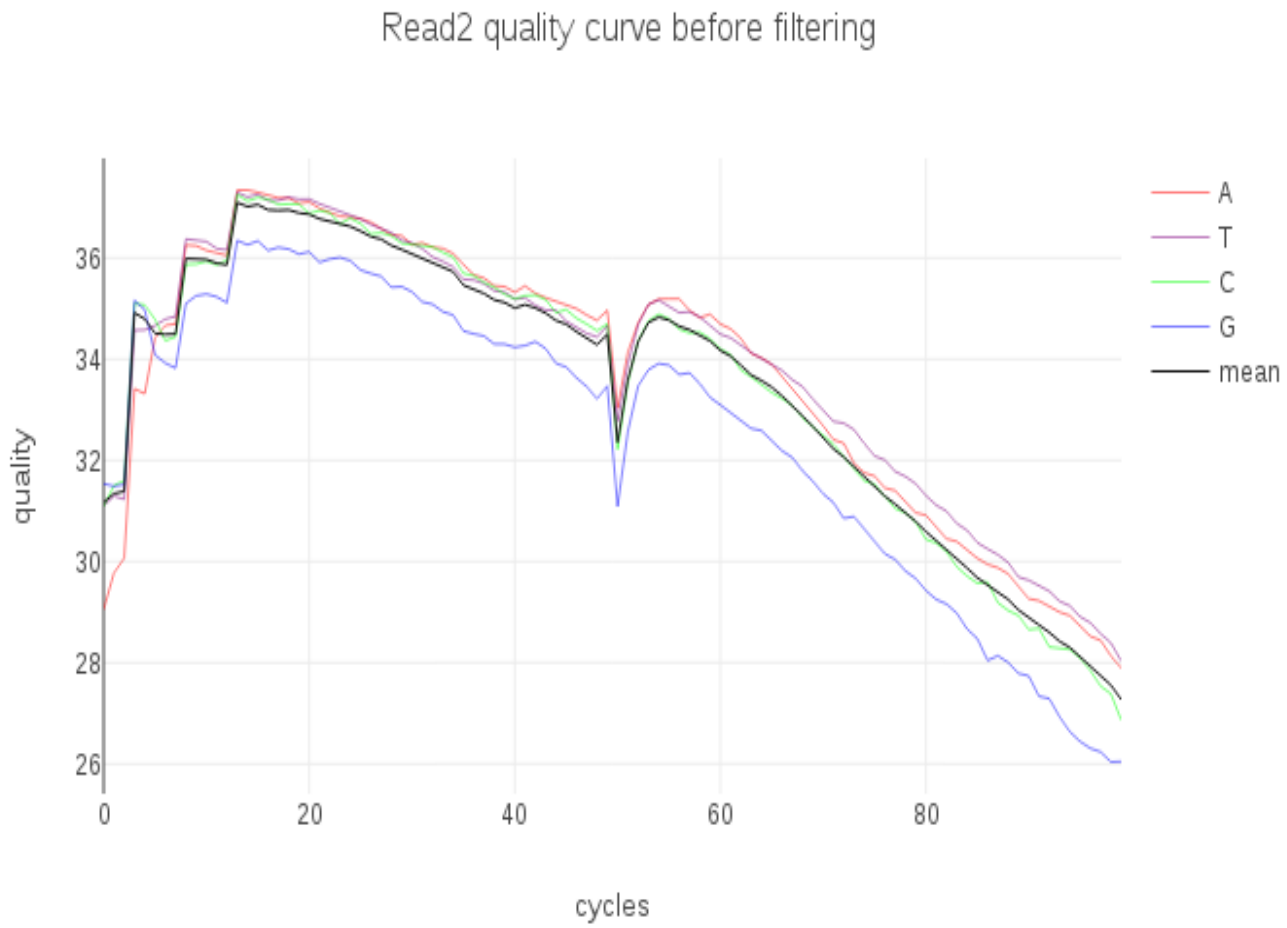
The view is kmer strand bias after filtering of R1.X-axis is about the counts of relative forward strand K-MERs, while the Y-axis is about relative reverse ones.



3.15 Read2 quality curve before filtering

This view is the curve that quality of reads2 before filtering. It not only plots the quality of A, T, C, and G for each locus in the user-entered sequencing results file, but also plots the mass average of four bases per locus. The red curve represents the result of A base, the purple curve represents the result of T base , the green curve represents the result of C base , the blue curve represents the result of G base , and the black curve represents the result of the average.

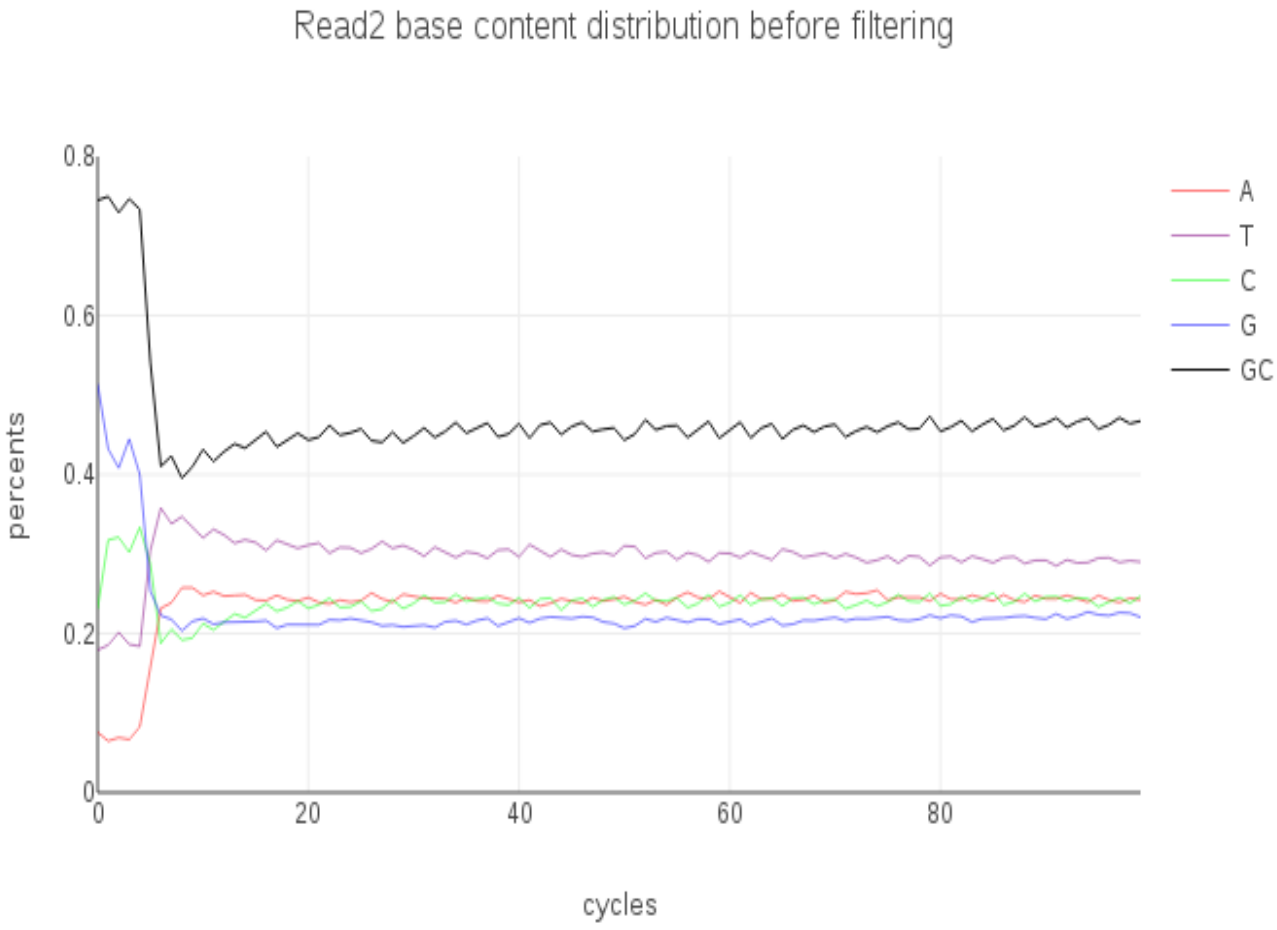
The abscissa represents the locus and the ordinate represents the mass fraction.



3.16 Read2 base content distribution before filtering

This view is the curve that the base content distribution of read2 before filtering. It plots the proportion of A, T, C, G, and GC at each locus in the sequencing results file entered by the user. The red curve indicates the result of the A base, the purple curve indicates the result of the T base, the green curve indicates the result of the C base, the blue curve indicates the result of the G base, and the black curve indicates the result of the sum of the GC.The abscissa indicates the locus and the ordinate indicates the proportion.

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.

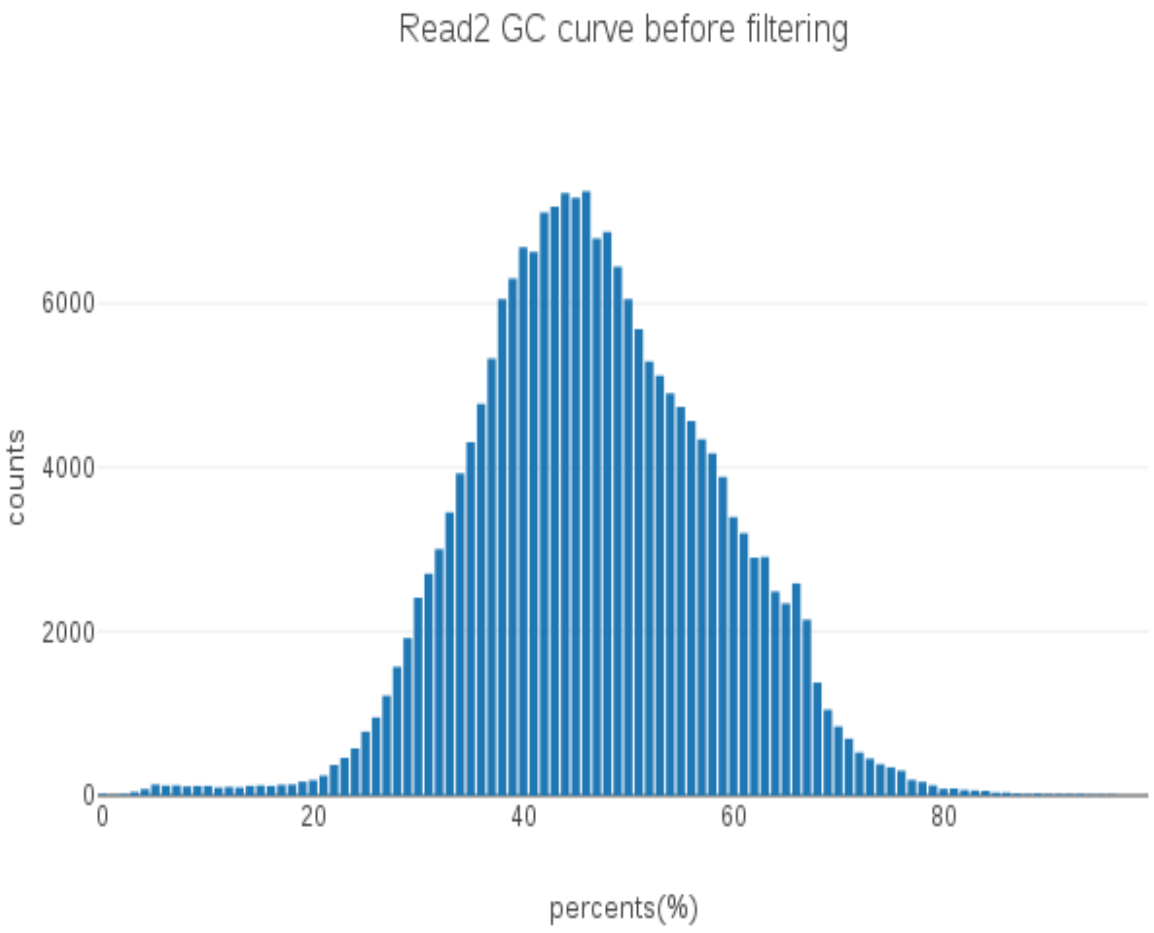


3.17 Read2 GC curve before filtering

This view measures the GC content across the whole length of each sequence in the fastq file of Read2 of before filtering.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.

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.

Values in X-axis represent the GC content(%), and the values in Y-axis represent the counts.

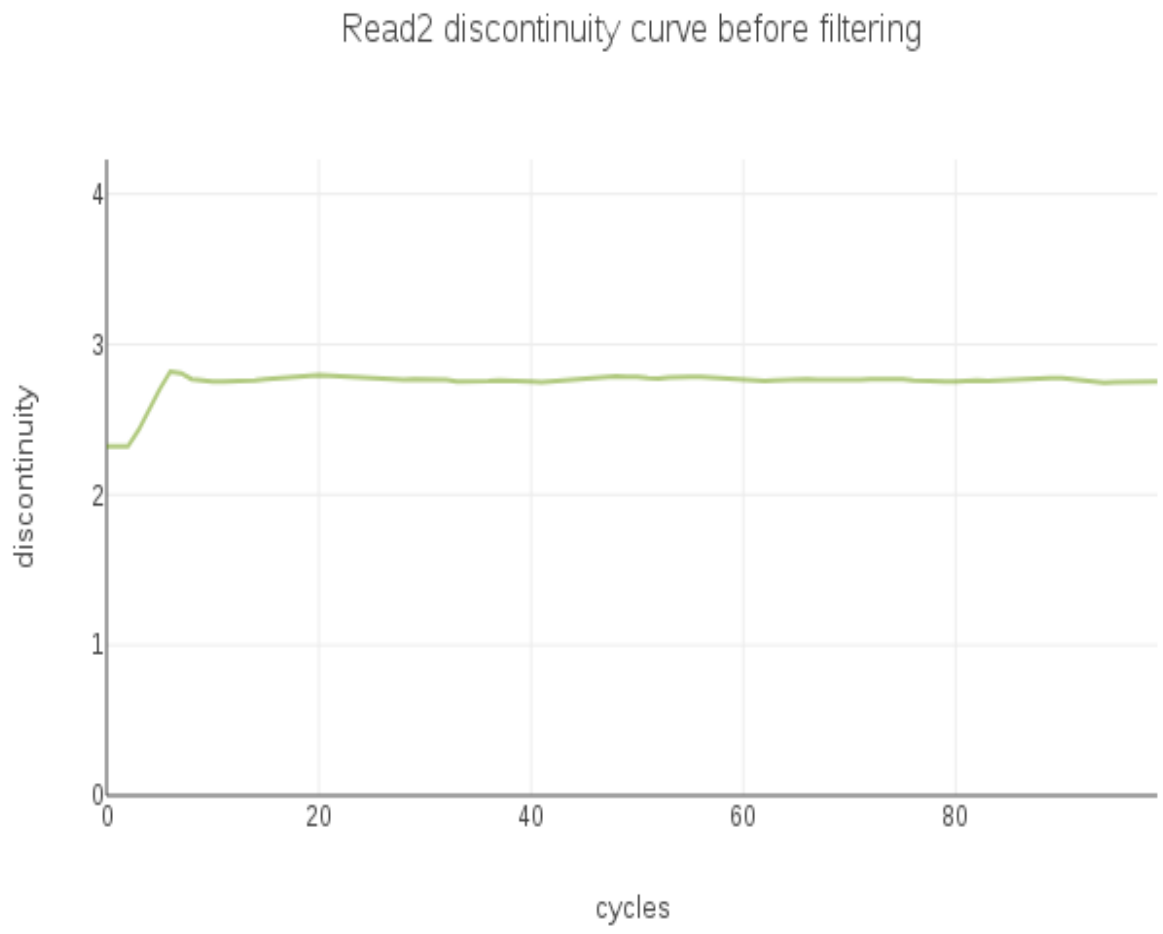


3.18 Read2 per base discontinuity before filtering

Besides normal per-cycle base content and quality profiling, AfterQC implements a methods to give more information about per-cycle discontinuity profiling to reflect sequencing quality instability.

The mean discontinuity should be more or less stable for all sequencing cycles. For a short window of sequencing cycles, it use the average discontinued base number in this window to calculate the discontinuity. If discontinuity drops down significantly cycle by cycle, it usually reflects a sequencing issue, which may be caused by the per-cycle washing process not working well.

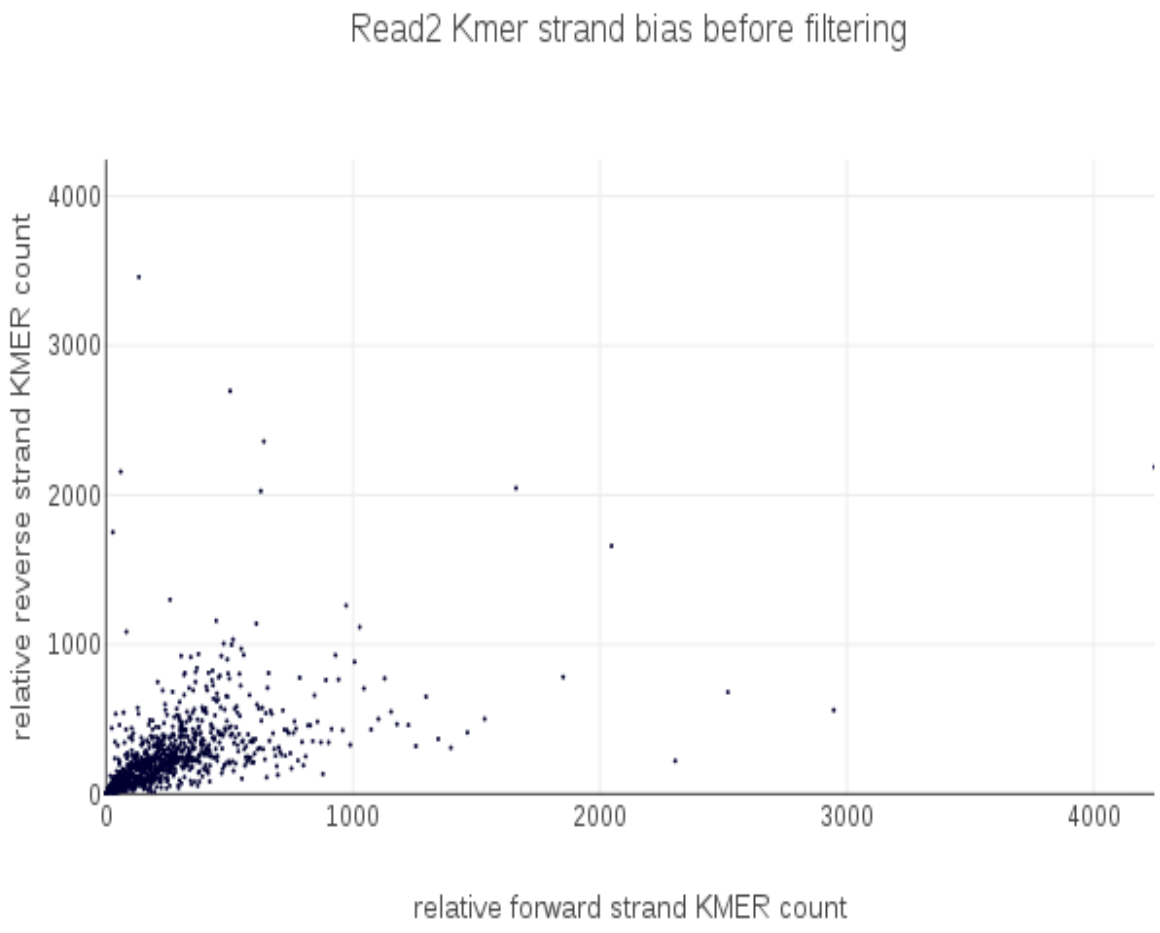
The view is a curve that the per base discontinuity before filtering of reads2. The abscissa indicates cycles and the ordinate indicates discontinuity.



3.19 Read2 kmer strand bias before filtering

Besides normal per-cycle base content and quality profiling, AfterQC implements novel method to give more information about sequencing quality: strand bias profiling to reflect amplification bias. The method is based on a hypothesis: if the DNA amplification process and sequencing process have only little non-uniformity, the repeat count of a short K-MER should be close to the repeat count of its reverse complement. So we plot each K-MER and its reverse complement's counts, and check whether most points are near the line y=x.

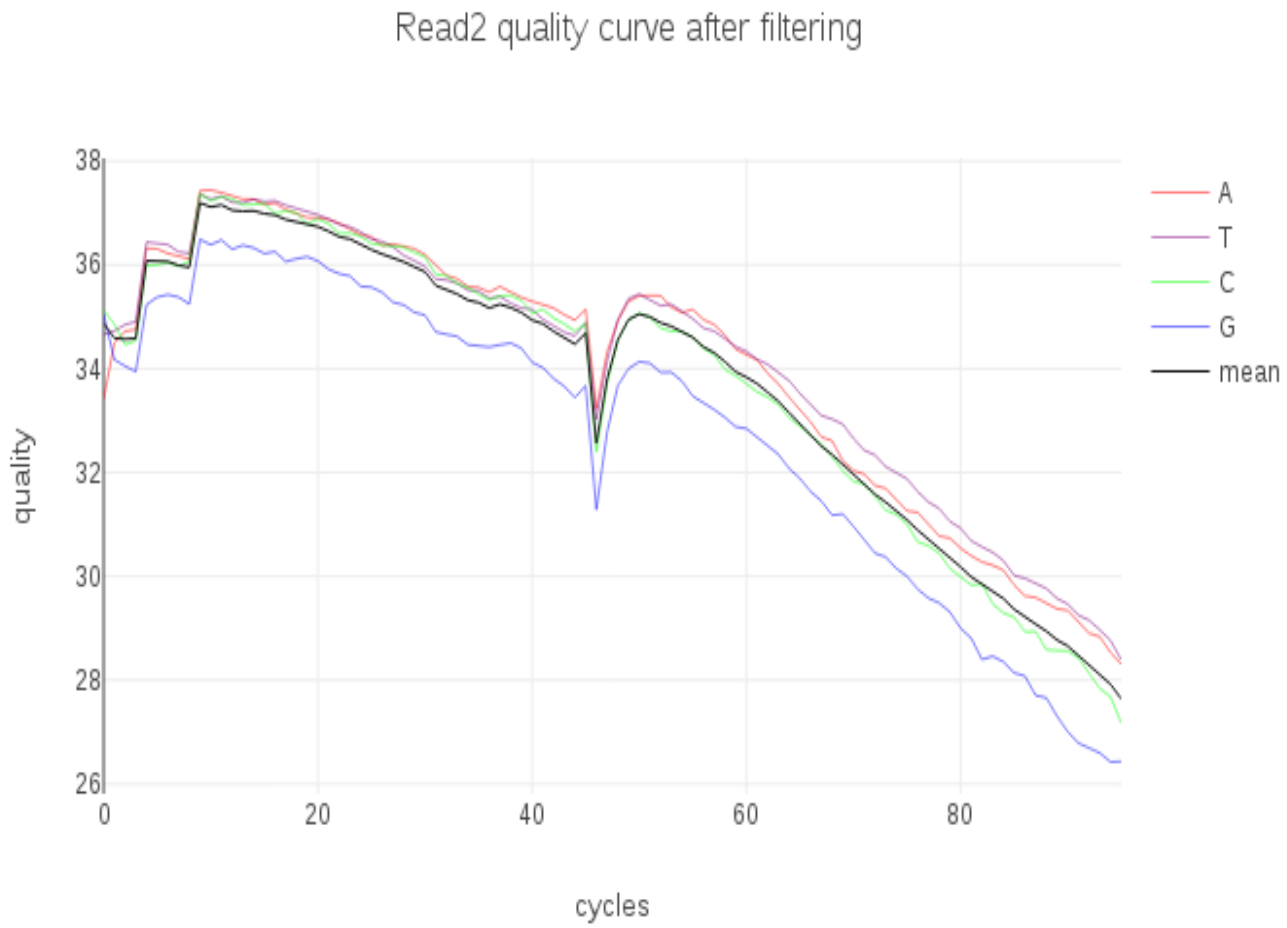
The view is kmer strand bias before filtering of R2.X-axis is about the counts of relative forward strand K-MERs, while the Y-axis is about relative reverse ones.



3.20 Read2 quality curve after filtering

This view is the curve that quality of reads2 after filtering. It not only plots the quality of A, T, C, and G for each locus in the user-entered sequencing results file, but also plots the mass average of four bases per locus. The red curve represents the result of A base, the purple curve represents the result of T base , the green curve represents the result of C base , the blue curve represents the result of G base , and the black curve represents the result of the average.

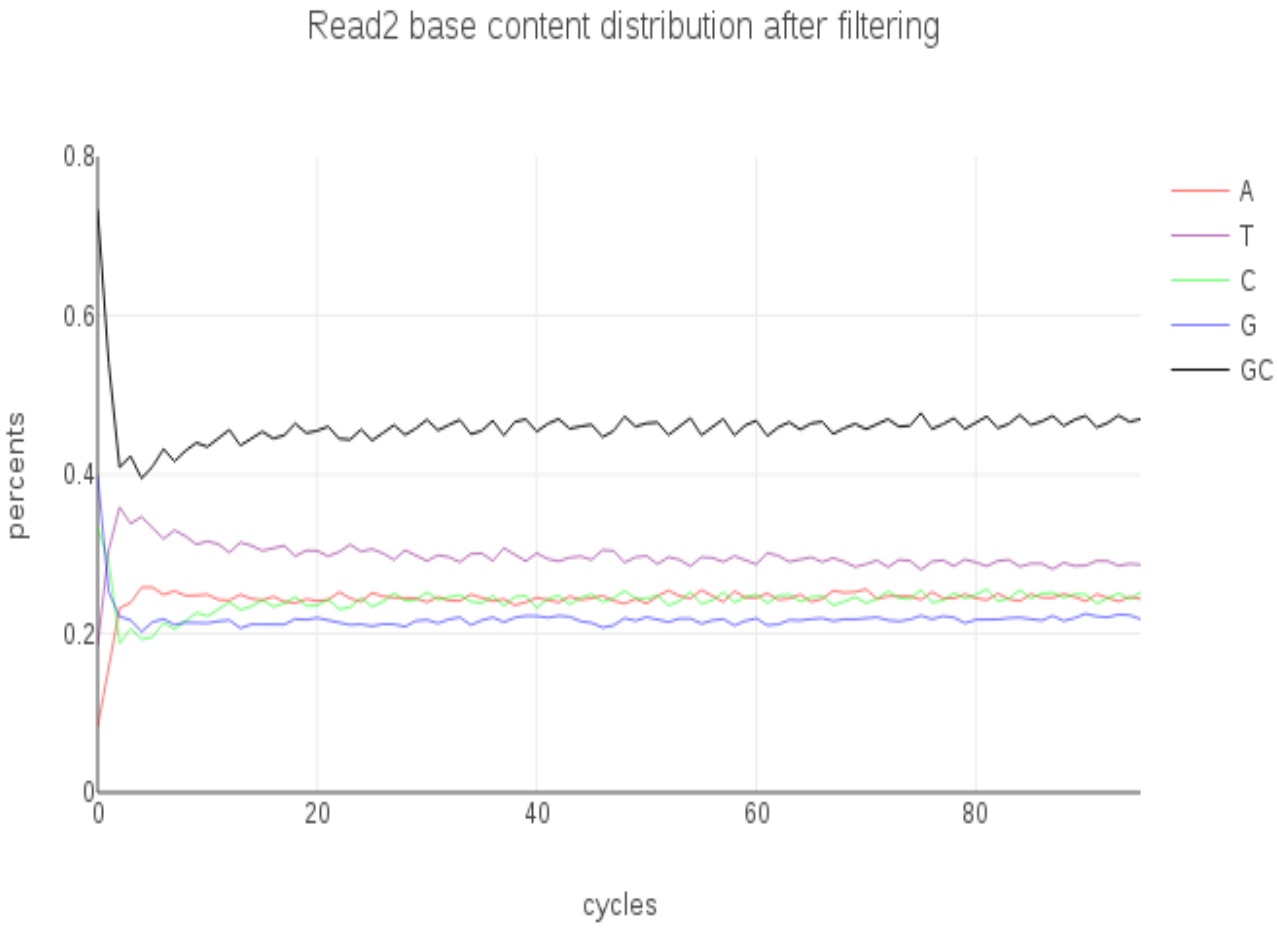
The abscissa represents the locus and the ordinate represents the mass fraction.



3.21 Read2 base content distribution after filtering

This view is the curve that the base content distribution of read2 after filtering. It plots the proportion of A, T, C, G, and GC at each locus in the sequencing results file entered by the user. The red curve indicates the result of the A base, the purple curve indicates the result of the T base, the green curve indicates the result of the C base, the blue curve indicates the result of the G base, and the black curve indicates the result of the sum of the GC.The abscissa indicates the locus and the ordinate indicates the proportion.

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.

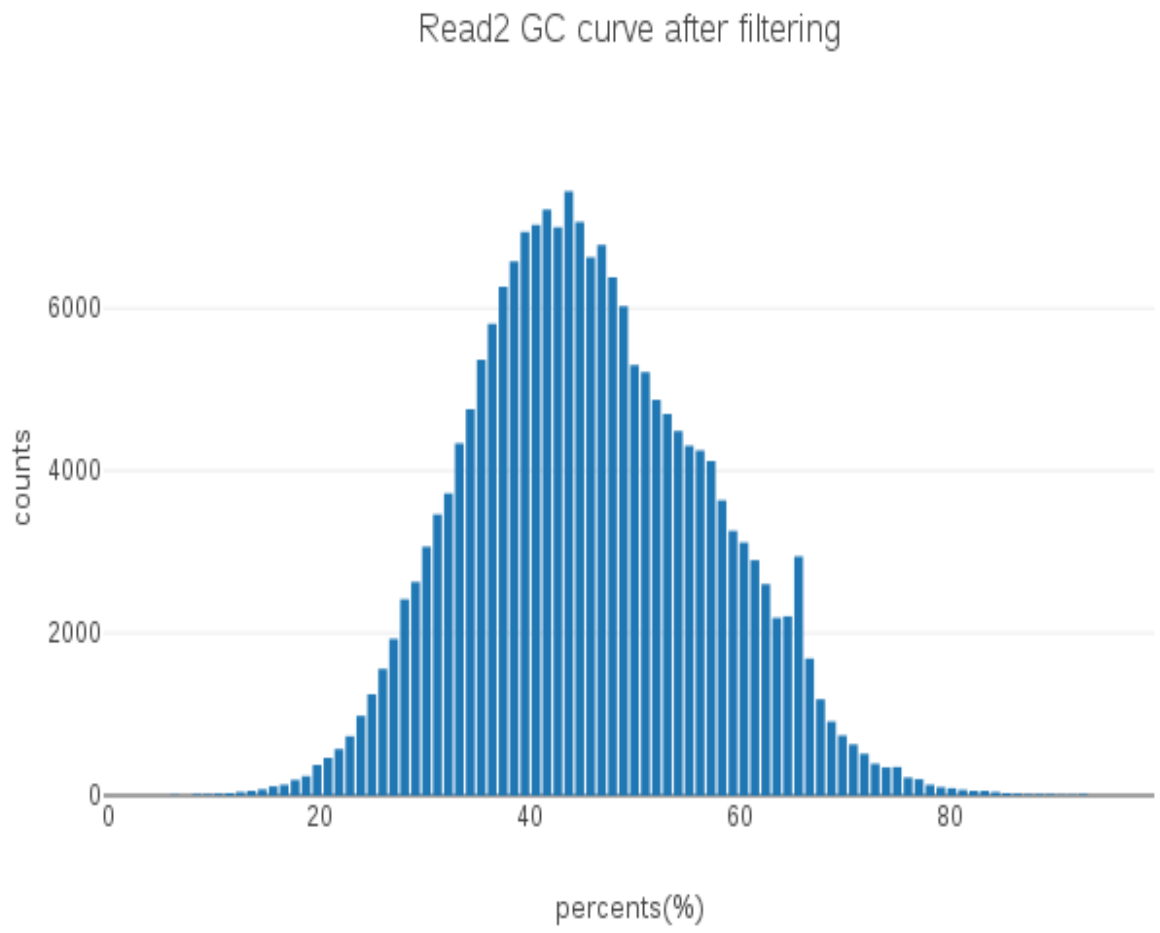


3.22 Read2 GC curve after filtering

This view measures the GC content across the whole length of each sequence in the fastq file of Read2 of after filtering.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.

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.

Values in X-axis represent the GC content(%), and the values in Y-axis represent the counts.

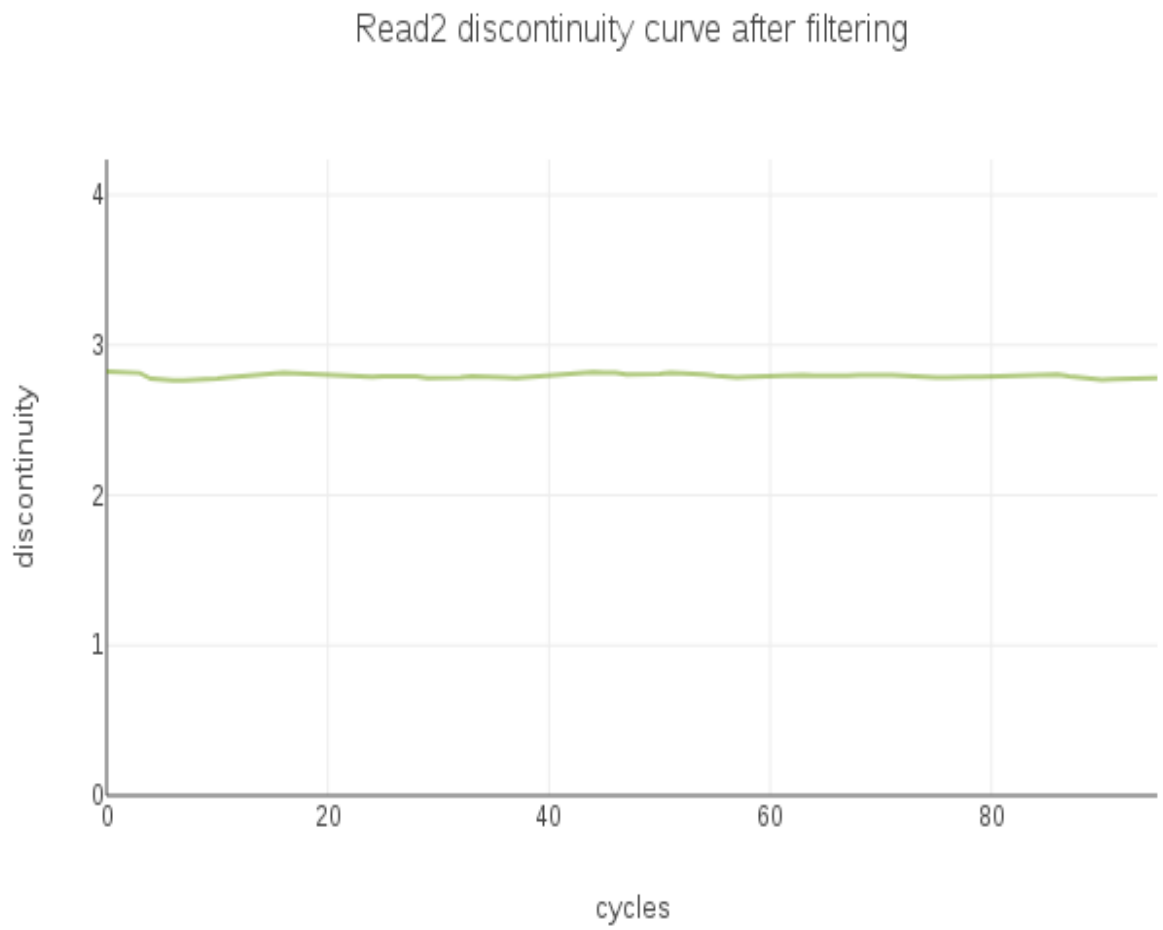


3.23 Read2 per base discontinuity after filtering

Besides normal per-cycle base content and quality profiling, AfterQC implements a methods to give more information about per-cycle discontinuity profiling to reflect sequencing quality instability.

The mean discontinuity should be more or less stable for all sequencing cycles. For a short window of sequencing cycles, it use the average discontinued base number in this window to calculate the discontinuity. If discontinuity drops down significantly cycle by cycle, it usually reflects a sequencing issue, which may be caused by the per-cycle washing process not working well.

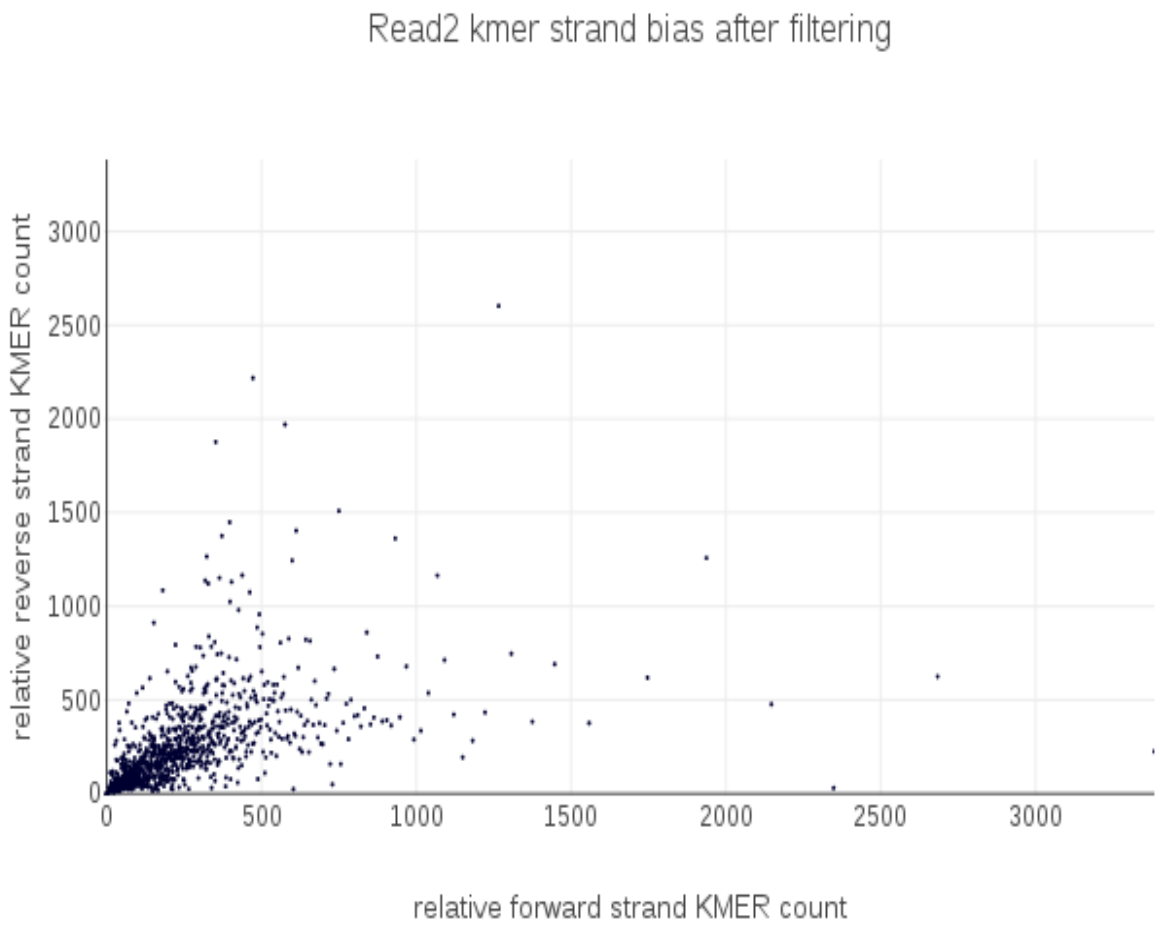
The view is a curve that the per base discontinuity after filtering of reads2. The abscissa indicates cycles and the ordinate indicates discontinuity.



3.24 Read2 kmer strand bias after filtering

Besides normal per-cycle base content and quality profiling, AfterQC implements novel method to give more information about sequencing quality: strand bias profiling to reflect amplification bias. The method is based on a hypothesis: if the DNA amplification process and sequencing process have only little non-uniformity, the repeat count of a short K-MER should be close to the repeat count of its reverse complement. So we plot each K-MER and its reverse complement's counts, and check whether most points are near the line y=x.

The view is kmer strand bias after filtering of R2.X-axis is about the counts of relative forward strand K-MERs, while the Y-axis is about relative reverse ones.



Chapter 4

RSeQC Result

Currently, a few tools are available for the QC of high-throughput sequencing data, but most of them (FastQC , htSeqTools, FASTX-ToolKit and SAM-Stat) only focus on raw sequence-related metrics . RNA-SeQC is the only tool designed for RNA-seq QC, but it still lacks many important functions, such as saturation checking .

RSeQC can solve such problems. RSeQC package is used to comprehensively assess the quality of RNA-seq experiments performed on clinical samples or other well-annotated model organisms, such as mouse, fly, Caenorhabditis elegans and yeast.

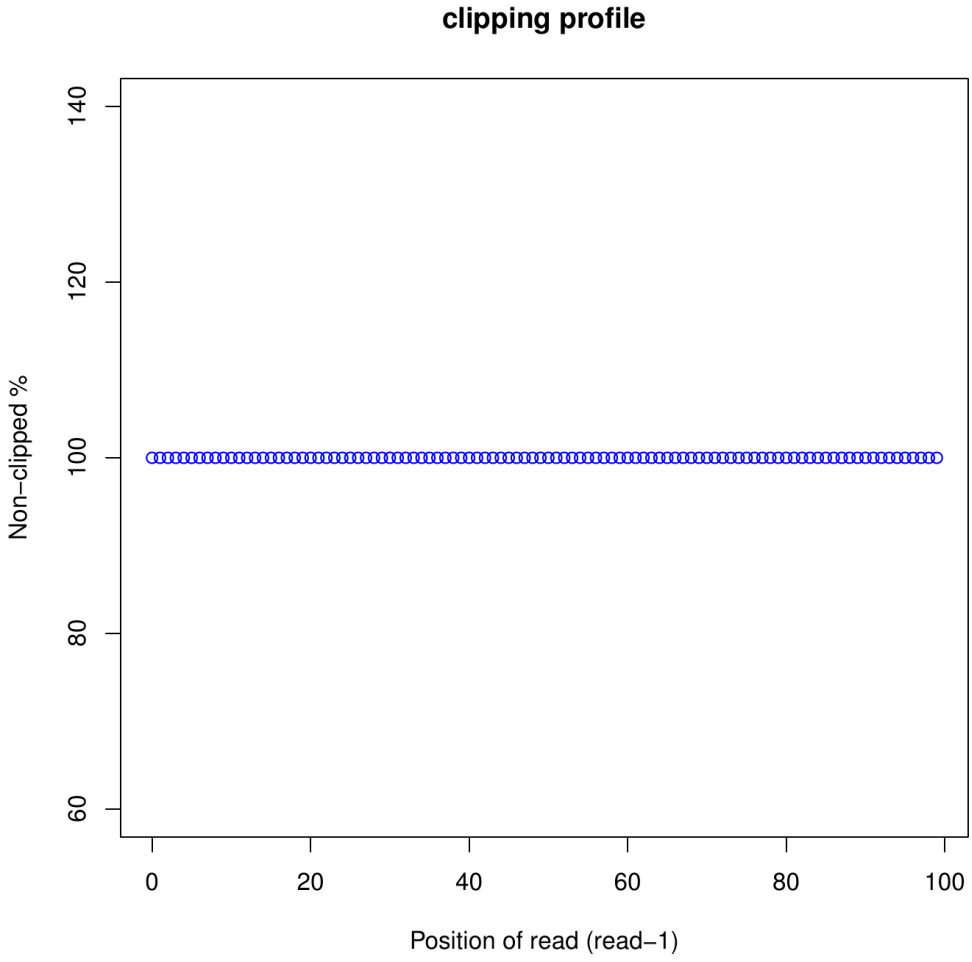
It provides a number of useful modules that can comprehensively evaluate high throughput sequence data especially RNA-seq data.

”Basic modules” quickly inspect sequence quality, nucleotide composition bias, PCR bias and GC bias, while ”RNA-seq specific modules” investigate sequencing saturation status of both splicing junction detection and expression estimation, mapped reads clipping profile, mapped reads distribution, coverage uniformity over gene body, reproducibility, strand specificity and splice junction annotation.[2]

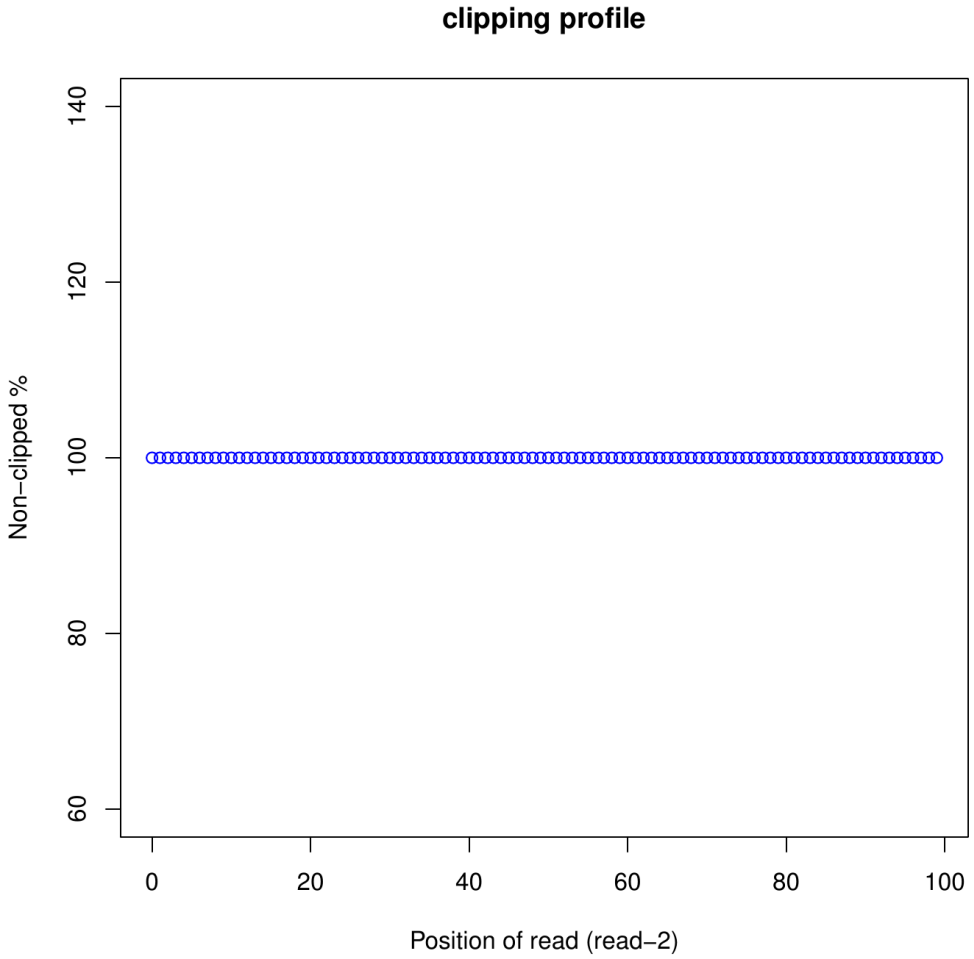
4.1 Calculate the distributions of clipped nucleotides across reads

This module is used to estimate clipping profile of RNA-seq reads from BAM file. Double-end sequencing has two result plots, one for read1 and one for read2. There is only one result plot in single-ended sequencing.

The view is the estimated clipping profile of reads1 in the bam file by user inputed.The abscissa indicates position of reads and the ordinate indicates the ratio of non-clipped.



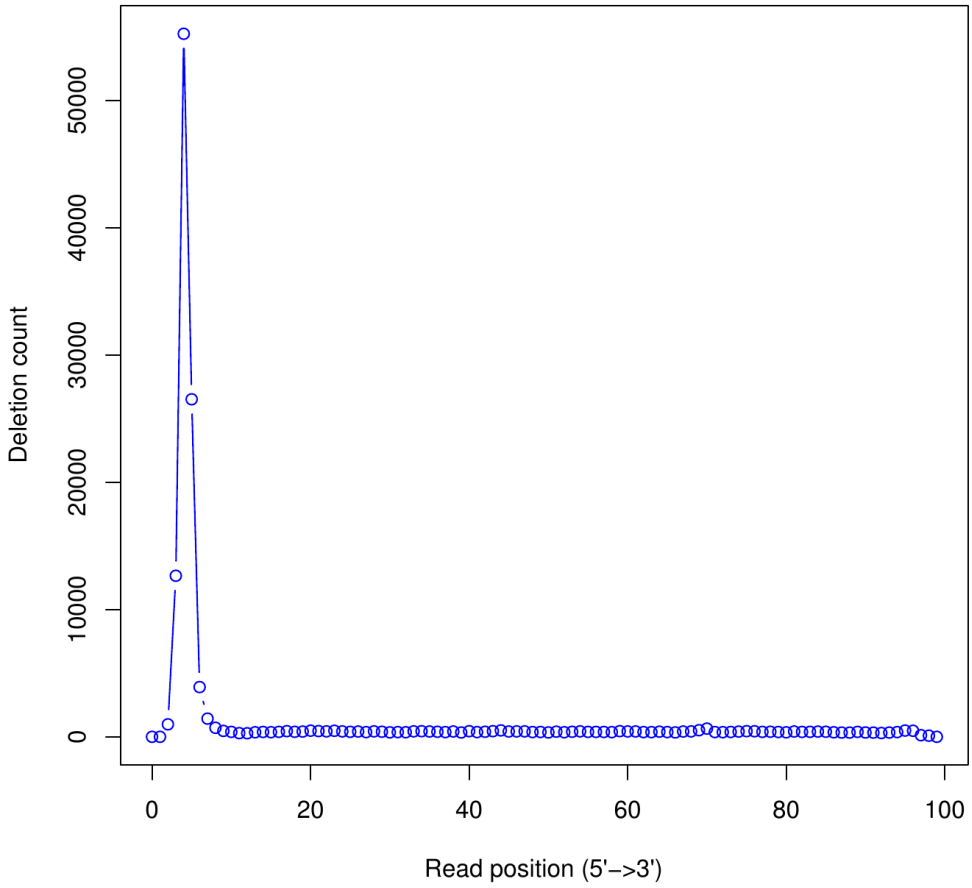
The view is the estimated clipping profile of reads2 in the bam file by user inputed.The abscissa indicates position of reads and the ordinate indicates the ratio of non-clipped.



4.2 Calculate the distributions of deletions across reads

The view is used to calculating the distributions of deletions across reads.

The abscissa indicates position of reads (from 5 to 3 end) and the ordinate indicates deletion count.

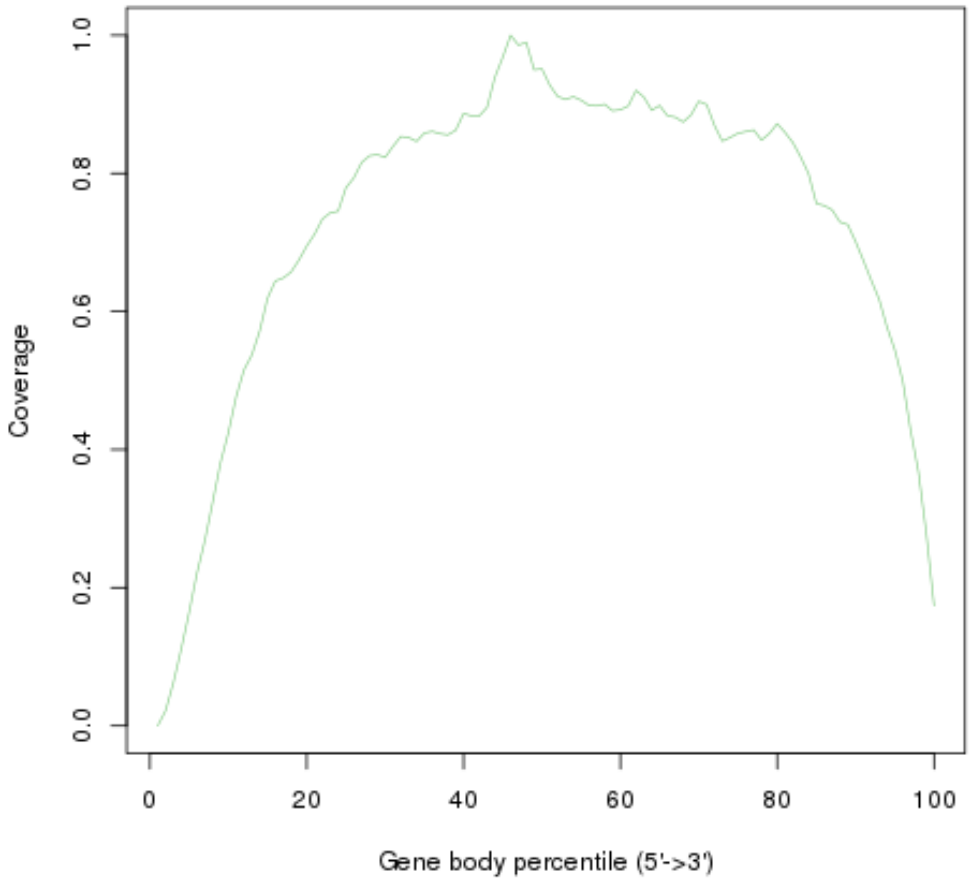


4.3 Calculate the RNA-seq reads coverage over gene body

The module scales all transcripts to 100 nt and calculates the number of reads covering each nucleotide position. Finally, it generates a plot illustrating the coverage profile along the gene body.

This view is the result curve of calculating the RNA-seq reads coverage over gene body.

The abscissa indicates gene body percentile(From 5 to 3 end) and the ordinate indicates coverage.

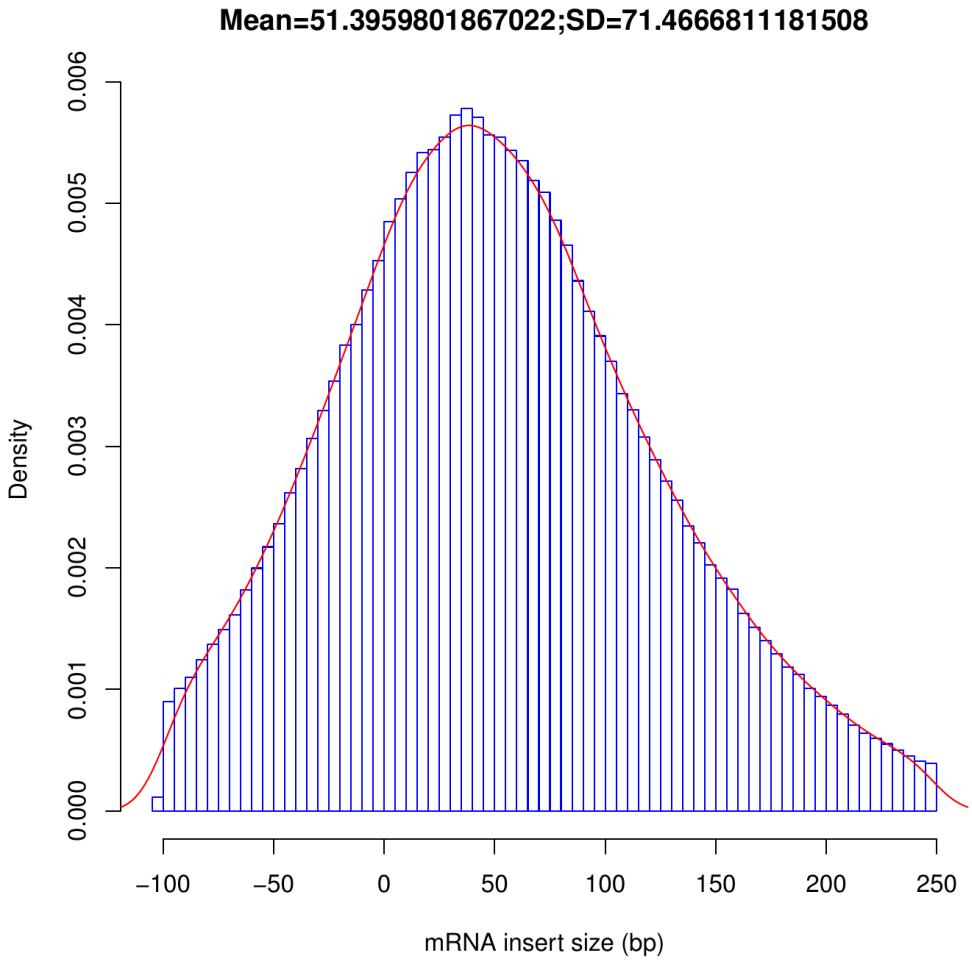


4.4 Calculate inner distance between read pairs

The module is used to estimate the inner distance distribution between paired reads.This is an important parameter when using RNA-seq data to detect structure variation or aberrant splicing.

The view is the result of calculating inner distance between read pairs.

The abscissa indicates mRNA insrt size(bp) and the ordinate indicates Density.And the calculation results of the mean value and the variance are displayed at the heading position.

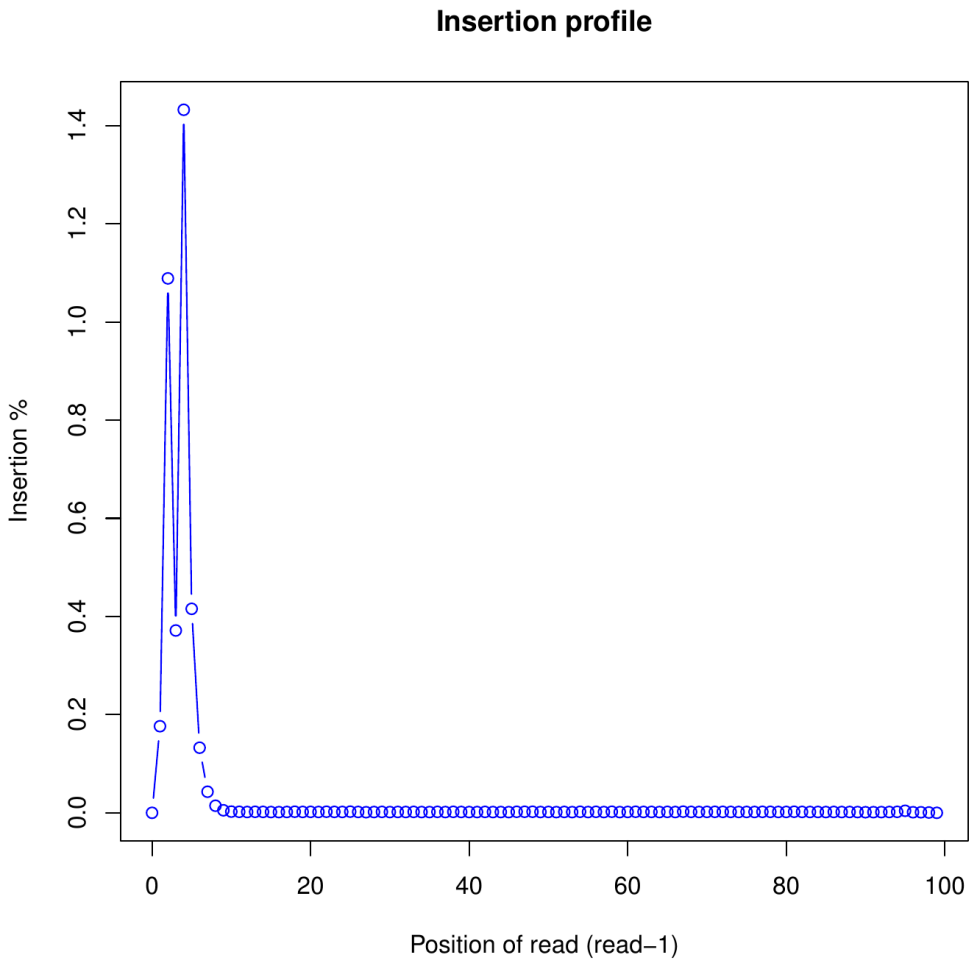


4.5 Calculate the distributions of inserted nucleotides across reads

The module is used to calculate the distributions of inserted nucleotides across reads.

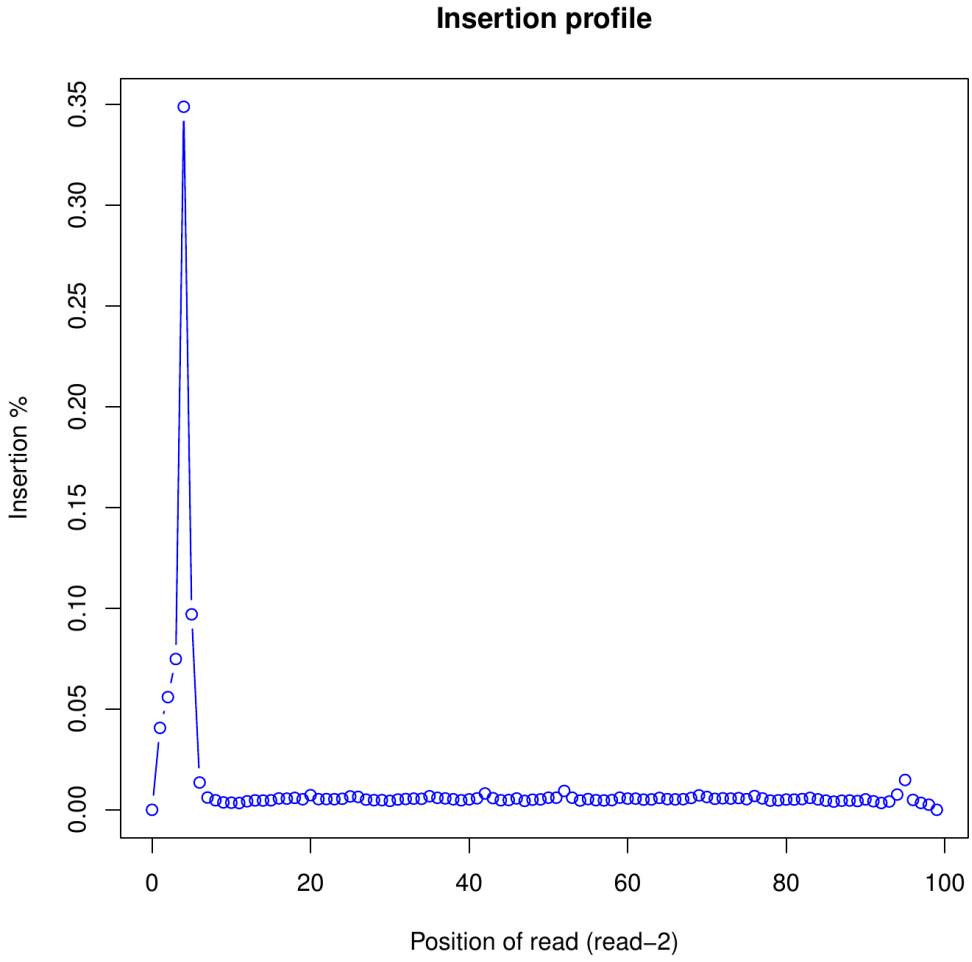
The view is a curve of insertion profile of reads1.

The abscissa indicates position of read1 and the ordinate indicates the ratio of insertion.



The view is a curve of insertion profile of reads2.

The abscissa indicates position of read2 and the ordinate indicates the ratio of insertion.



4.6 The detected splicing sites were compared with the reference gene model

The module is used to separate all detected splice junctions into 'known', 'complete novel' and 'partial novel' by comparing them with the reference gene model.

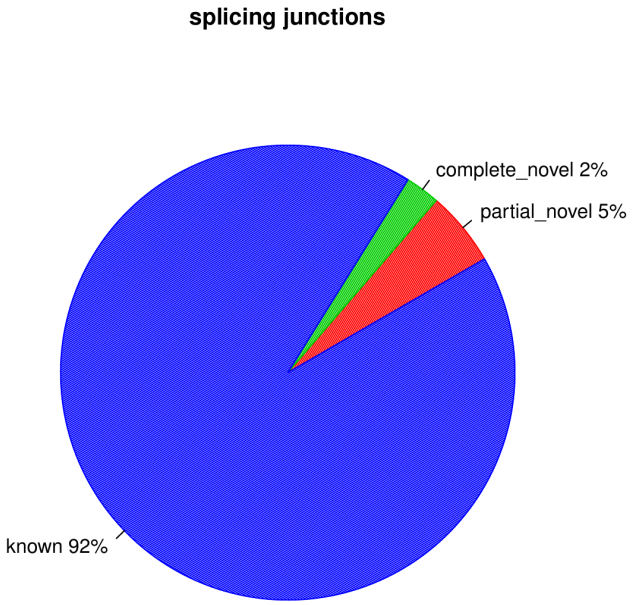
Splicing annotation is performed in two levels: splice event level and splice junction level.

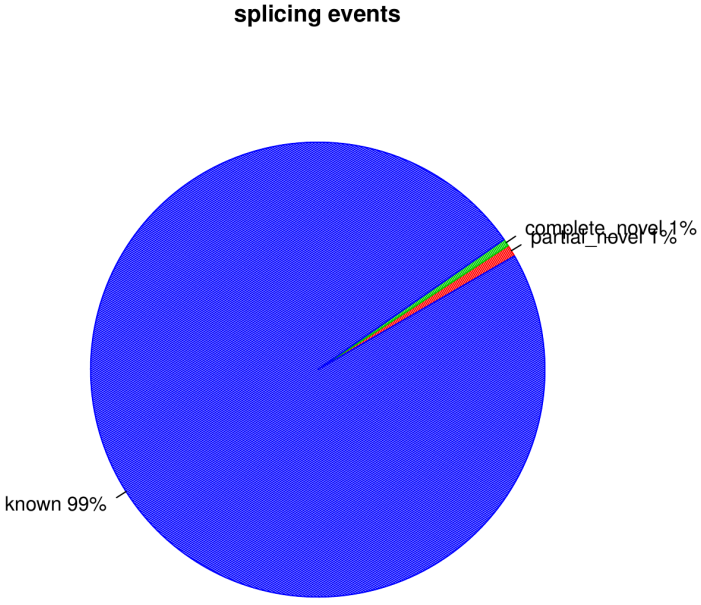
- splice event: An RNA read, especially long read, can be spliced 2 or more times, each time is called a splicing event; In this sense, 100 spliced reads can produce >= 100 splicing events.
- splice junction: multiple splicing events spanning the same intron can be consolidated into one splicing junction.

All detected junctions can be grouped to 3 exclusive categories:

- Annotated: The junction is part of the gene model. Both splice sites, 5' splice site (5'SS) and 3'splice site (3'SS) can be annotated by reference gene model.
- complete-novel: Complete new junction. Neither of the two splice sites cannot be annotated by gene model.
- partial-novel: One of the splice site (5'SS or 3'SS) is new, while the other splice site is annotated (known).

In the figure, blue represents known, red represents partial-novel, and green represents complete-novel.

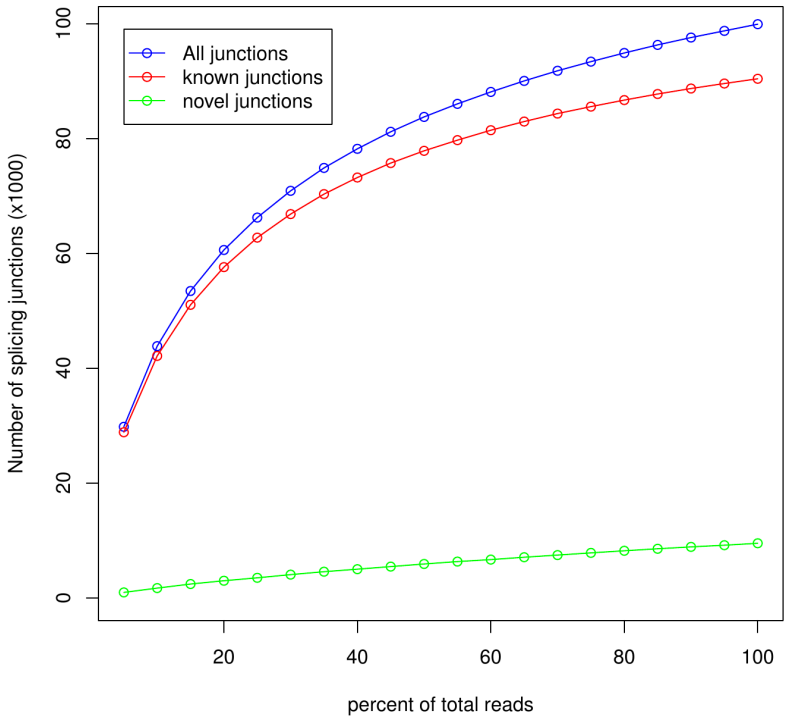

splicing junctions
complete_novel 2%
partial_novel 5%
known 92%


splicing events
complete_novel 1%
partial_novel 1%
known 99%

4.7 Splicing sites of each subset were detected and compared with the reference gene model

The module is used to determine if the current sequencing depth is sufficient to perform alternative splicing analyses. For a well annotated organism, the number of expressed genes in particular tissue is almost fixed so the number of splice junctions is also fixed. All (annotated) splice junctions should be rediscovered from a saturated RNA-seq data, otherwise, downstream alternative splicing analysis is problematic because low abundance splice junctions are missing.

In the result view, red curve represents known junctions, blue curve represents all junctions, green curve represents novel junctions. And the abscissa indicates percent of total reads and the ordinate indicates number of splicing junctions(X1000). If the number of "known junction" reaches a plateau, current sequencing depth is almost saturated. In other words, nearly all "known junctions" have already been detected, and deeper sequencing will not likely to detect additional "known junction" and will only increase junction coverage . Other curve explanations are similar to the above, and will not be described again.

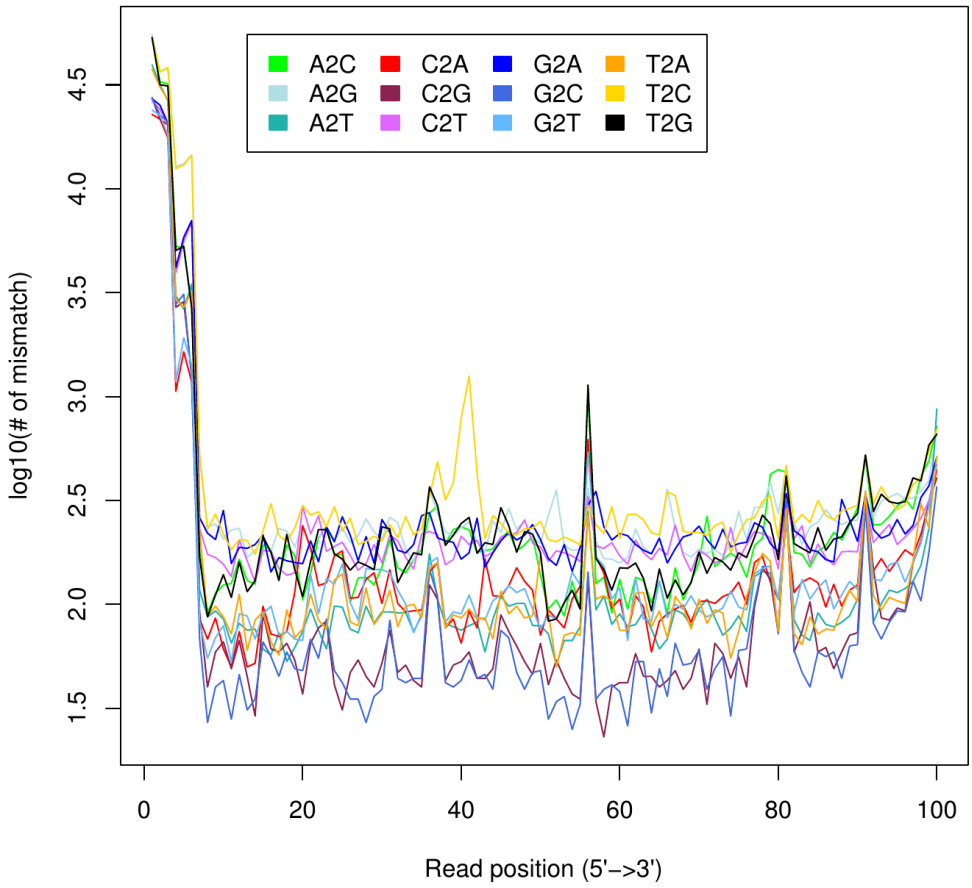


4.8 Calculate the distribution of mismatches across reads

The view is used to display the result of calculating the distribution of mismatches across reads.

The abscissa indicates positions of reads(from 5 to 3 end) and the ordinate indicates log10(# of mismatch).

And in the result view is included the label box that to display the specific meaning of the different color curves.



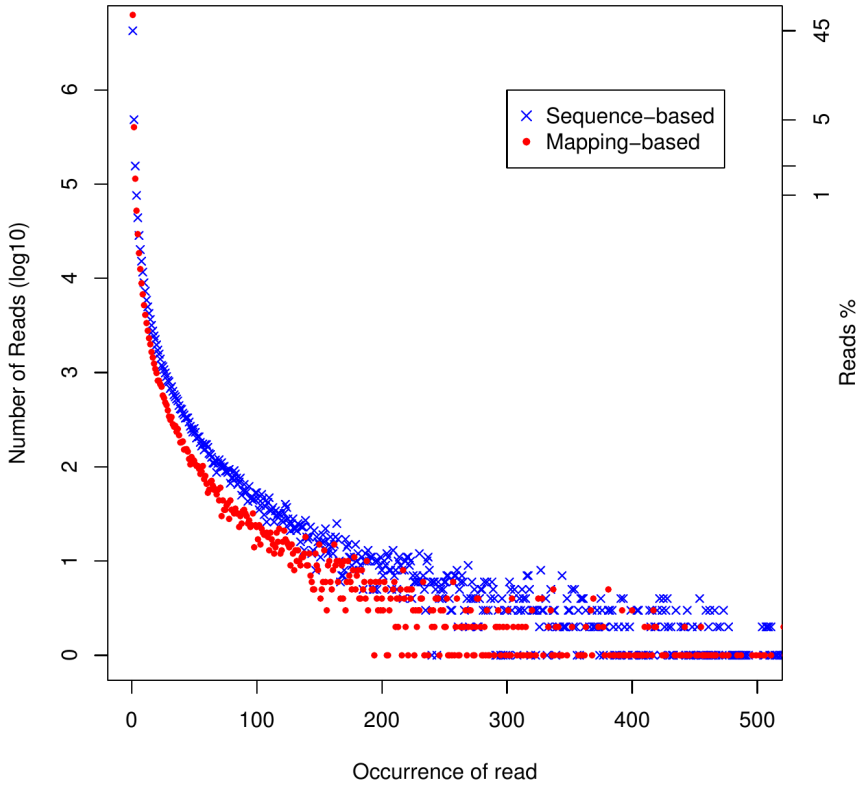
4.9 Two strategies were used to determine reads duplication rate

The module uses two strategies were used to determine reads duplication rate:

- Sequence based: reads with identical sequence are regarded as duplicated reads.
- Mapping based: reads mapped to the exactly same genomic location are regarded as duplicated reads.

For splice reads, reads mapped to the same starting position and splice the same way are regarded as duplicated reads.

In the result view is included two result, red dots for the results based on the mapping, and blue for the sequence-based results. The abscissa represents the incidence of read, the left ordinate represents the ratio of reads and the right ordinate represents the ratio of reads.

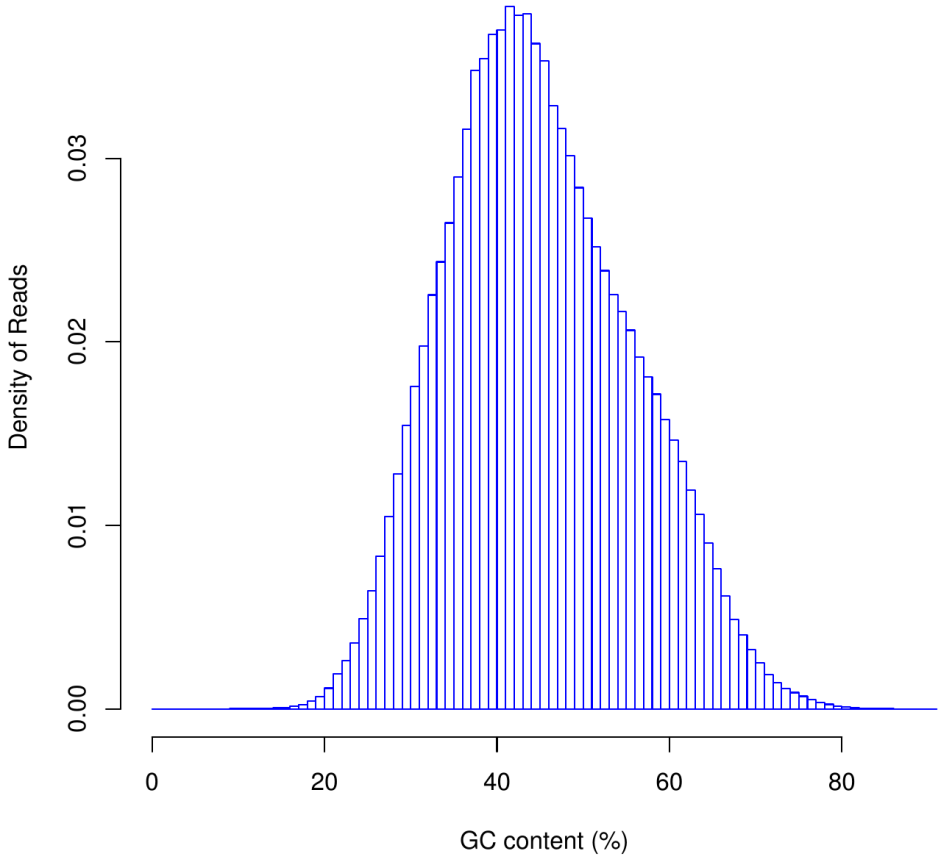


4.10 GC content distribution of reads

This view measures the GC content across the whole length of each sequence in a bam file.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.

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.

Values in X-axis represent the GC content(%), and the values in Y-axis represent the Density of Reads.

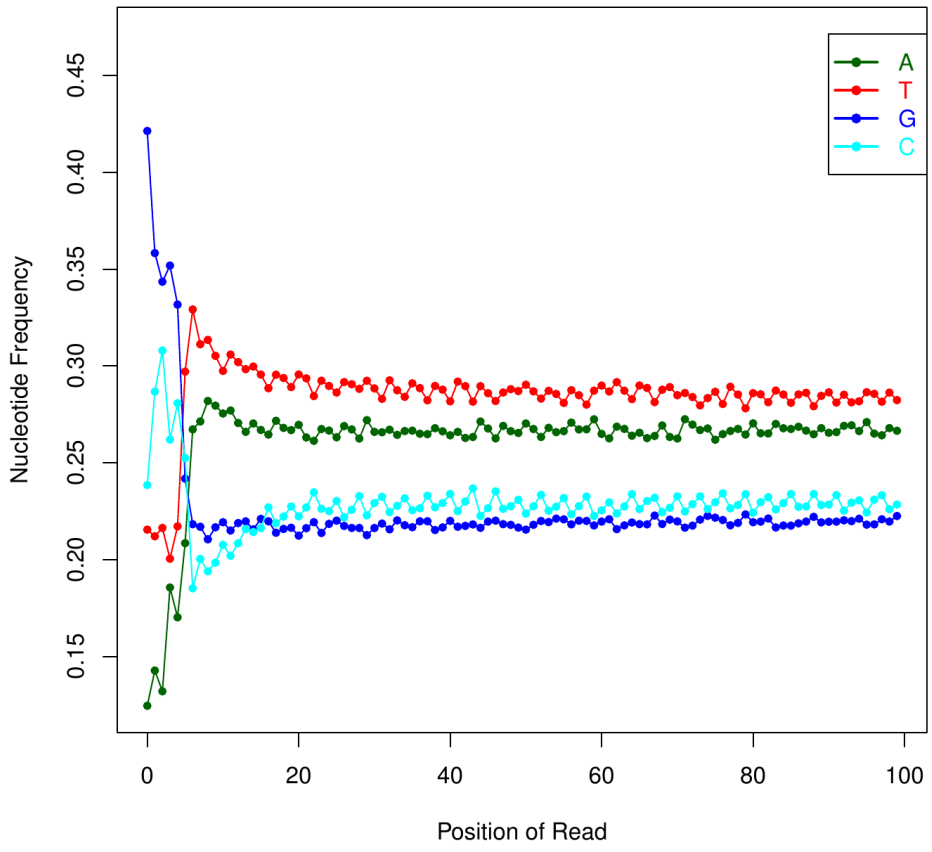


4.11 Check for nucleotide composition bias

This module is used to check the nucleotide composition bias. Due to random priming, certain patterns are over represented at the beginning (5'end) of reads. This bias could be easily examined by NVC (Nucleotide versus cycle) plot. NVC plot is generated by overlaying all reads together, then calculating nucleotide composition for each position of read (or each sequencing cycle).

In ideal condition (genome is random and RNA-seq reads is randomly sampled from genome), we expect A%=C%=G%=T%=25% at each position of reads.

Four curves of A, G, C, and T are included in the result view, and each curve represents the proportion of the corresponding base at each site. The abscissa indicates position of reads and the ordinate indicates nucleotide position. The green curve represents the A base, the red curve represents the T base, the dark blue represents the G base and the light blue represents the C base.



4.12 RPKM saturation

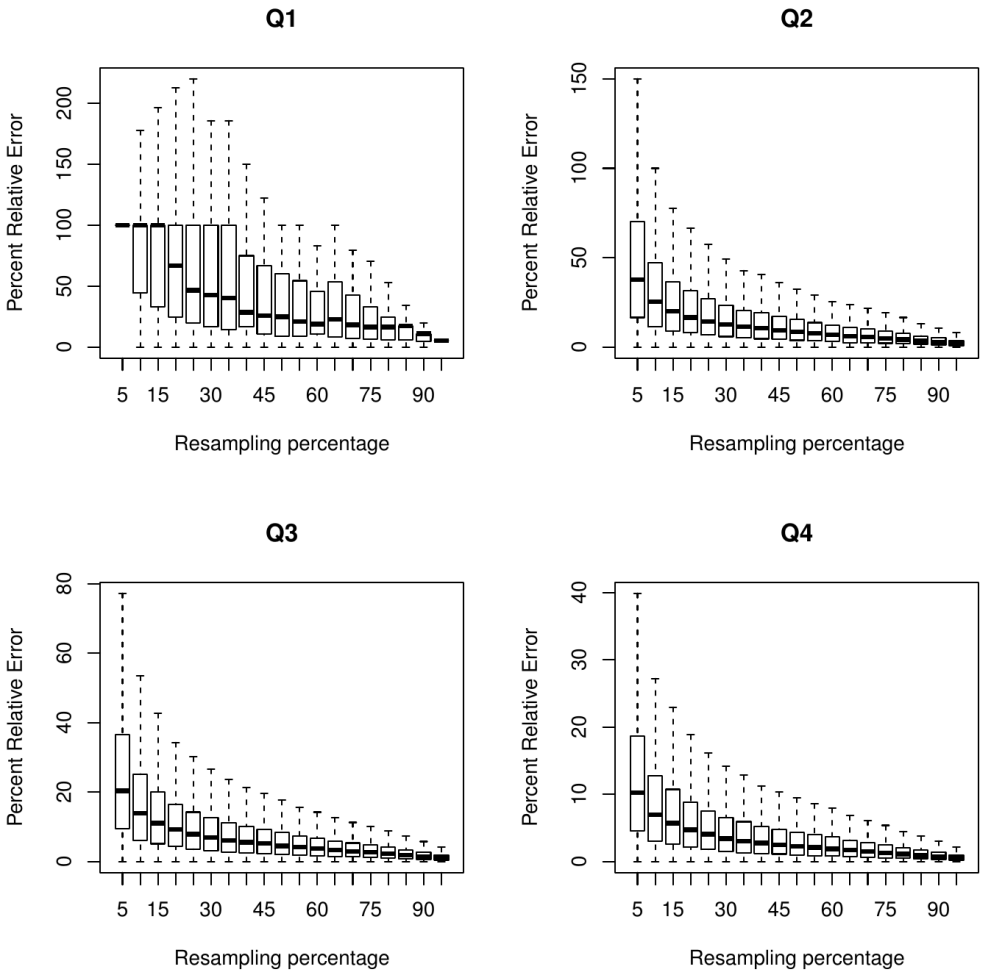
The precision of any sample statitics (RPKM) is affected by sample size (sequencing depth); ”resampling” or ”jackknifing” is a method to estimate the precision of sample statistics by using subsets of available data.

This module will resample a series of subsets from total RNA reads and then calculate RPKM value using each subset. By doing this we are able to check if the current sequencing depth was saturated or not (or if the RPKM values were stable or not) in terms of genes' expression estimation. If sequencing depth was saturated, the estimated RPKM value will be stationary or reproducible.

In the result figure, Y axis is ”Percent Relative Error” or ”Percent Error” which is used to measures how the RPKM estimated from subset of reads (i.e. RPKMobs) deviates from real expression level (i.e. RPKMreal). However, in practice one cannot know the RPKMreal. As a proxy, we use the RPKM estimated from total reads to approximate RPKMreal.

All transcripts were sorted in ascending order according to expression level (RPKM). Then they are divided into 4 groups:

- Q1 (0-25%): Transcripts with expression level ranked below 25 percentile.
- Q2 (25-50%): Transcripts with expression level ranked between 25 percentile and 50 percentile.
- Q3 (50-75%): Transcripts with expression level ranked between 50 percentile and 75 percentile.
- Q4 (75-100%): Transcripts with expression level ranked above 75 percentile.


Q1
Percent Relative Error
0
50
100
150
200
5
15
30
45
60
75
90
Resampling percentage
Q2
Percent Relative Error
0
50
100
150
5
15
30
45
60
75
90
Resampling percentage
Q3
Percent Relative Error
0
20
40
60
80
5
15
30
45
60
75
90
Resampling percentage
Q4
Percent Relative Error
0
10
20
30
40
5
15
30
45
60
75
90
Resampling percentage

reference

[1] Shifu Chen, Tanxiao Huang, Yanqing Zhou, Yue Han, Mingyan Xu, and Jia Gu. Afterqc: automatic filtering, trimming, error removing and quality control for fastq data. *Bmc Bioinformatics*, 18(3):80, 2017.

[2] L. Wang, S. Wang, and W. Li. Rseqc: quality control of rna-seq experiments. *Bioinformatics*.