Supplementary Information Table of Contents

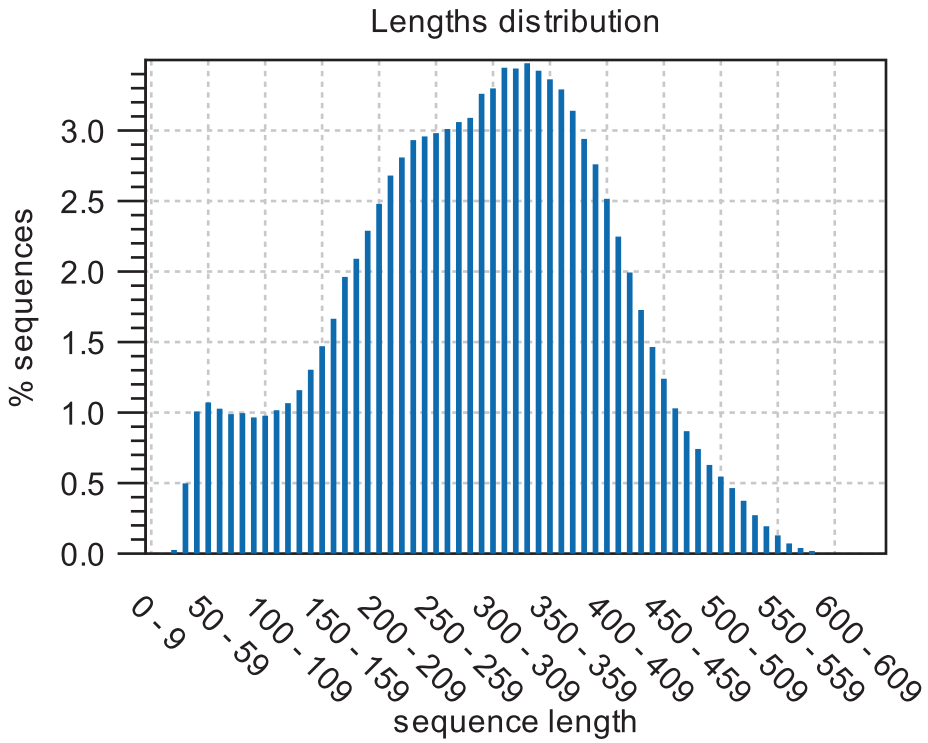
454 shotgun data Quality Report

1. Summary

Creation date:	Fri Jan 09 12:50:31 EST 2015
Generated by:	uks
Software:	CLC Genomics Workbench 7.5.1
Based upon:	1 data set
454_Shotgun:	462,052 sequences
Total nucleotides in data set	133,682,912 nucleotides

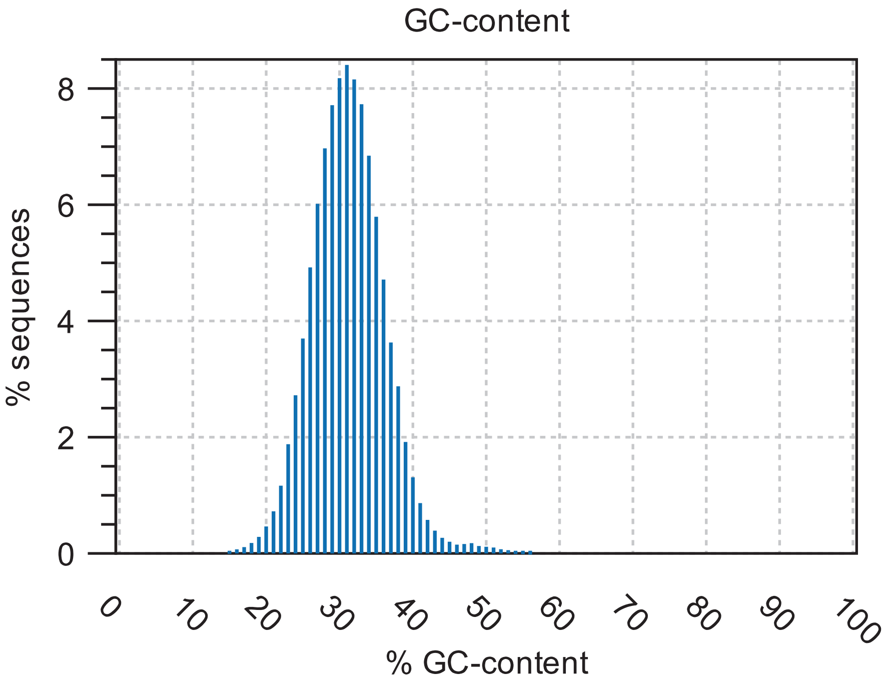
2. Per-sequence analysis

2.1 Lengths distribution



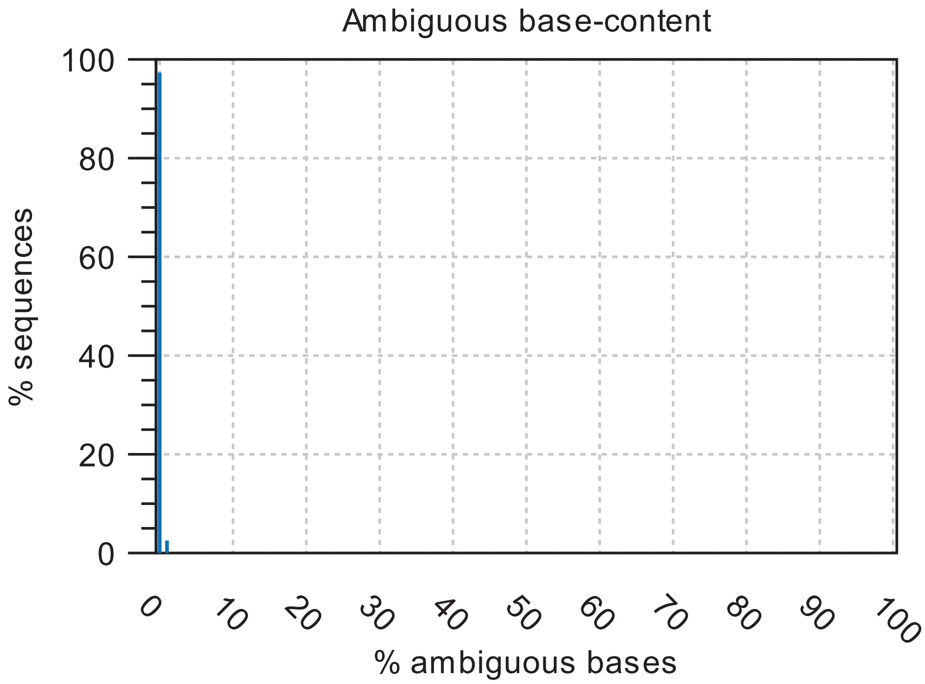
Distribution of sequence lengths. In cases of untrimmed Illumina or SOLiD reads it will ju st contain a single peak.
x: sequence length in base-pairs
y: number of sequences featuring a particular length normalized to the total number of seq uences

2.2 GC-content



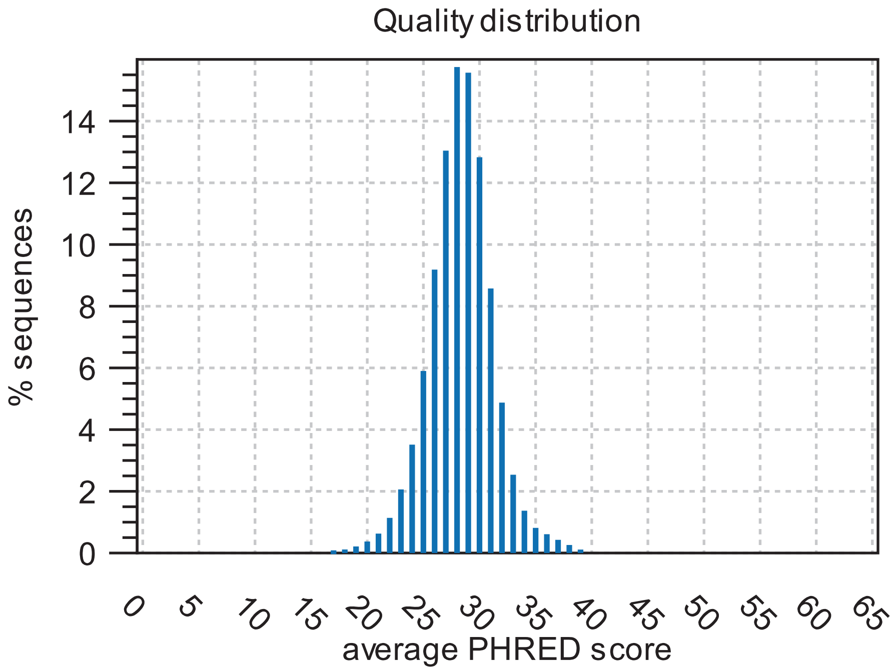
Distribution of GC-contents. The GC-content of a sequence is calculated as the number of G C-bases compared to all bases (including ambiguous bases).
x: relative GC-content of a sequence in percent
y: number of sequences featuring particular GC-percentages normalized to the total number of sequences

2.3 Ambiguous base-content



Distribution of N-contents. The N-content of a sequence is calculated as the number of amb iguous bases compared to all bases.
x: relative N-content of a sequence in percent
y: number of sequences featuring particular N-percentages normalized to the total number o f sequences

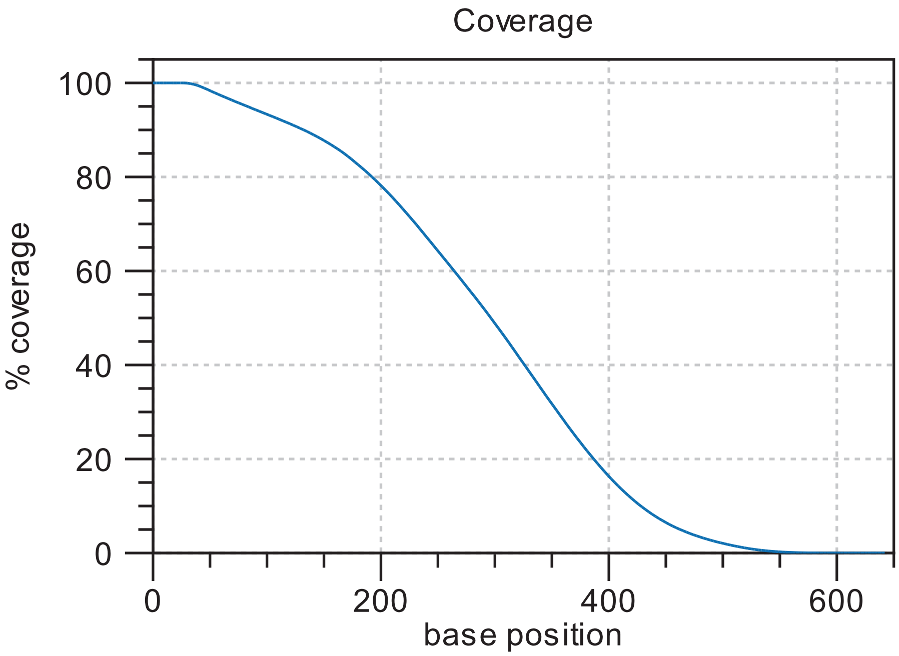
2.4 Quality distribution



Distribution of average sequence qualitie scores. The quality of a sequence is calculated as the arithmetic mean of its base qualities.
x: PHRED-score
y: number of sequences observed at that qual. score normalized to the total number of sequ ences

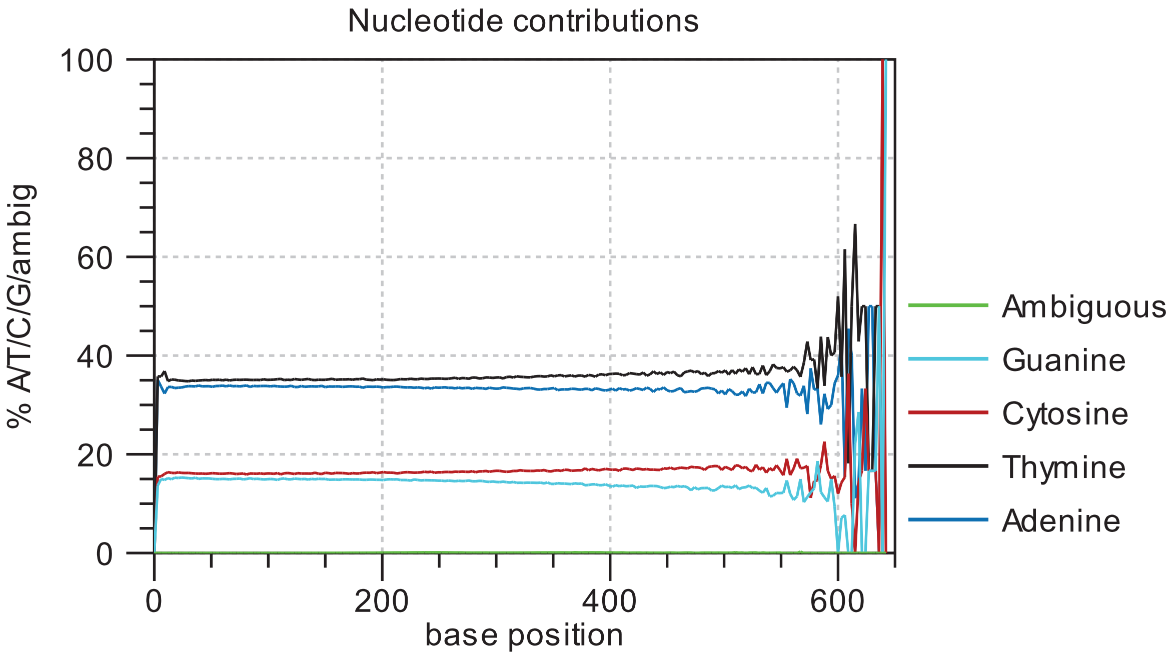
3. Per-base analysis

3.1 Coverage



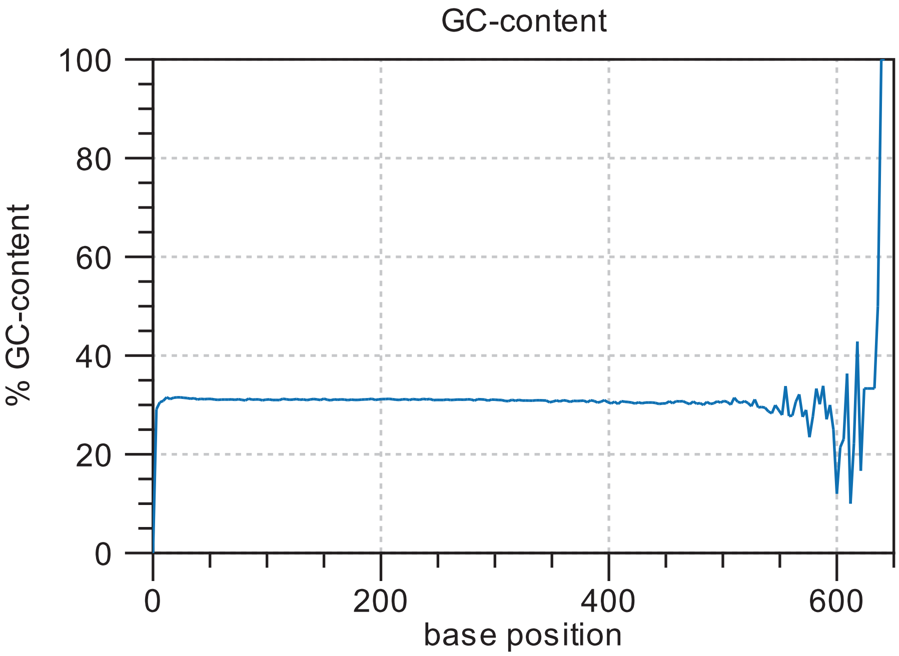
The number of sequences that support (cover) the individual base positions. In cases of un trimmed Illumina or SOLiD reads it will just contain a rectangle.
x: base position
y: number of sequences covering individual base positions normalized to the total number o f sequences

3.2 Nucleotide contributions



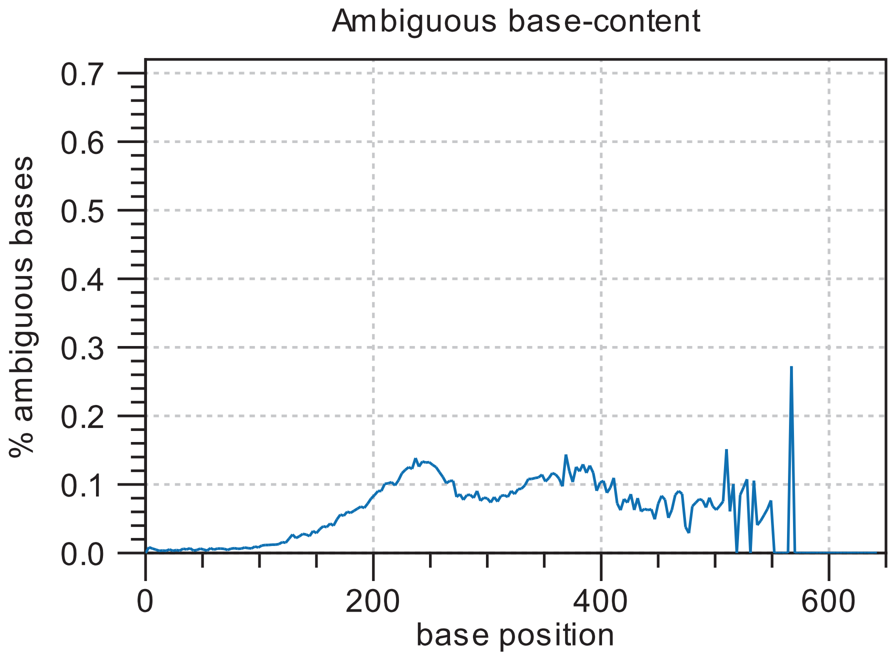
Coverages for the four DNA nucleotides and ambiguous bases.
x: base position
y: number of nucleotides observed per type normalized to the total number of nucleotides o bserved at that position

3.3 GC-content



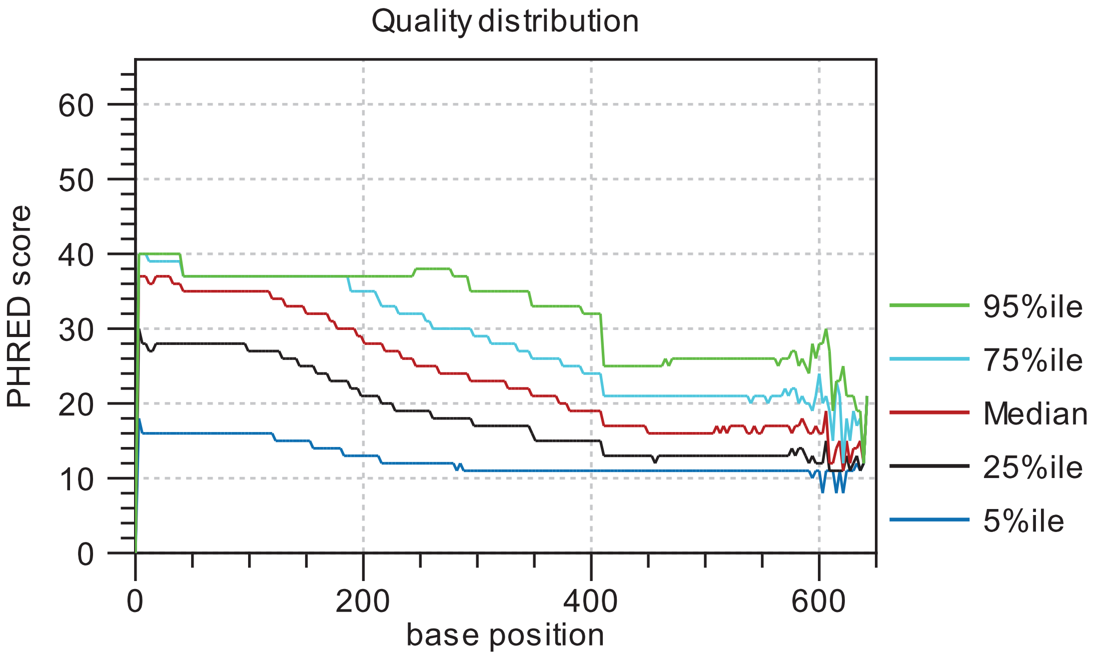
Combined coverage of G- and C-bases.
x: base position
y: number of G- and C-bases observed at current position normalized to the total number of bases observed at that position

3.4 Ambiguous base-content



Combined coverage of ambiguous bases.
x: base position
y: number of ambiguous bases observed at current position normalized to the total number o f bases observed at that position

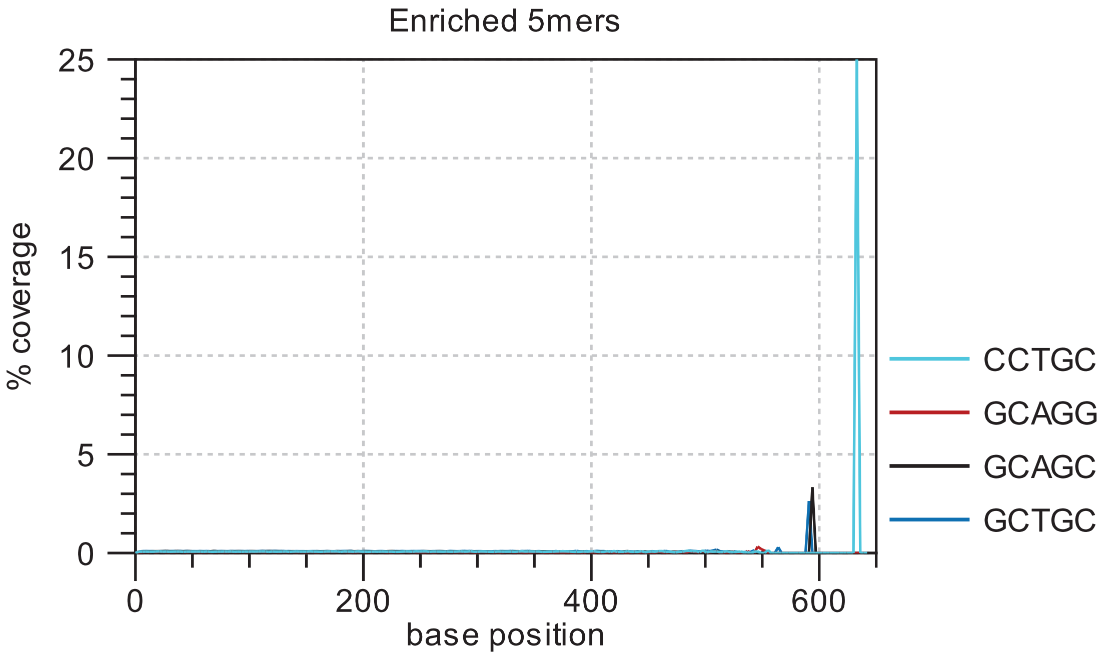
3.5 Quality distribution



Base-quality distribution along the base positions.
x: base position
y: median & percentiles of quality scores observed at that base position

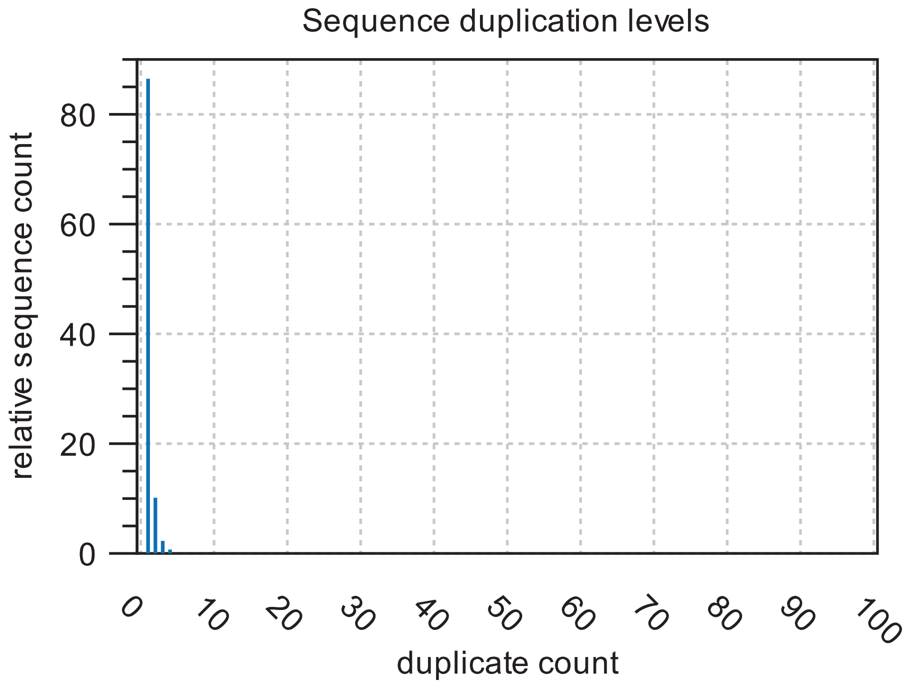
4. Over-representation analyses

4.1 Enriched 5mers



The five most-overrepresented 5mers. The over-representation of a 5mer is calculated as th e ratio of the observed and expected 5mer frequency. The expected frequency is calculated as product of the empirical nucleotide probabilities that make up the 5mer. (5mers that contain ambiguous bases are ignored)
x: base position
y: number of times a 5mer has been observed normalized to all 5mers observed at that posit ion

4.2 Sequence duplication levels



Duplication level distribution. Duplication levels are simply the count of how often a par ticular sequence has been found.
x: duplicate count
y: number of sequences that have been found that many times normalized to the number of un ique sequences

4.3 Duplicated sequences

A table of over-represented sequences is given in the supplementary report

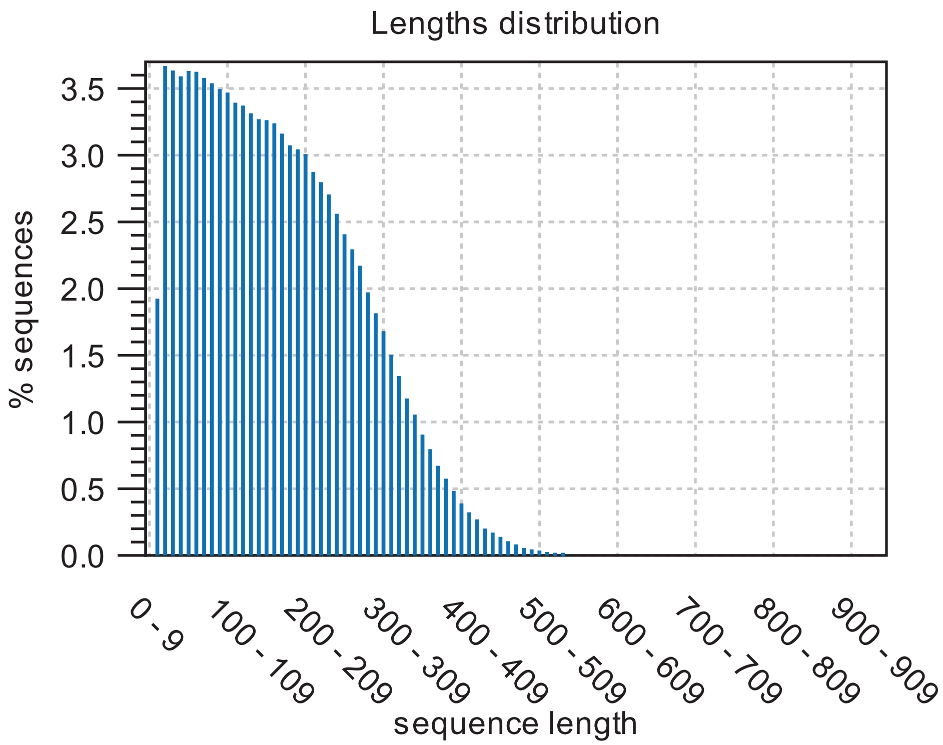
454 3KB data Quality Report

1. Summary

Creation date:	Thu Dec 11 14:06:09 EST 2014
Generated by:	uks
Software:	CLC Genomics Workbench 7.5.1
Based upon:	2 data sets
GQW19BL01 (single):	128,856 sequences
GQW19BL01 (paired):	764,756 sequences in pairs
Total nucleotides in data sets	150,922,863 nucleotides

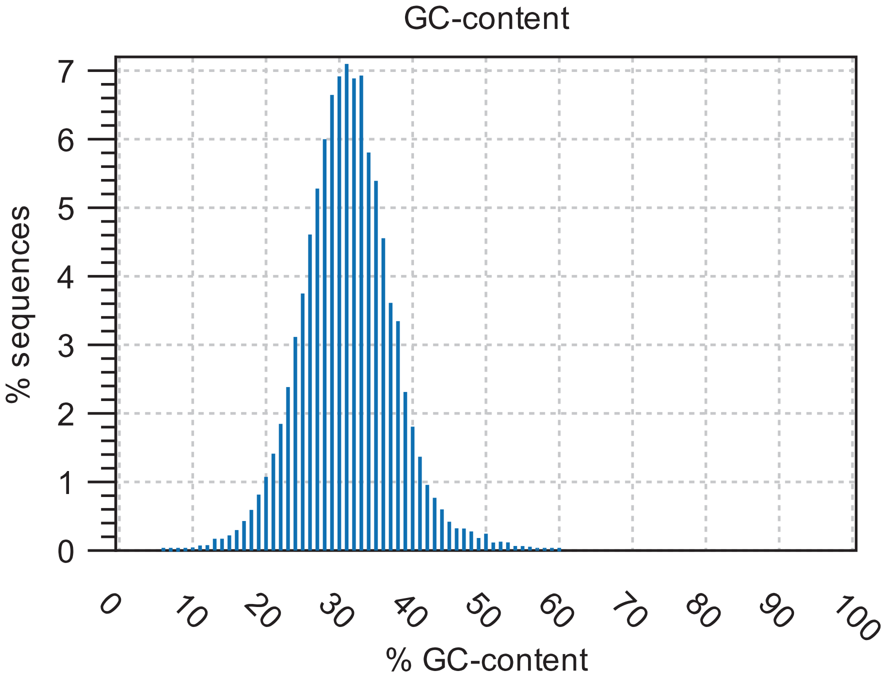
2. Per-sequence analysis

2.1 Lengths distribution



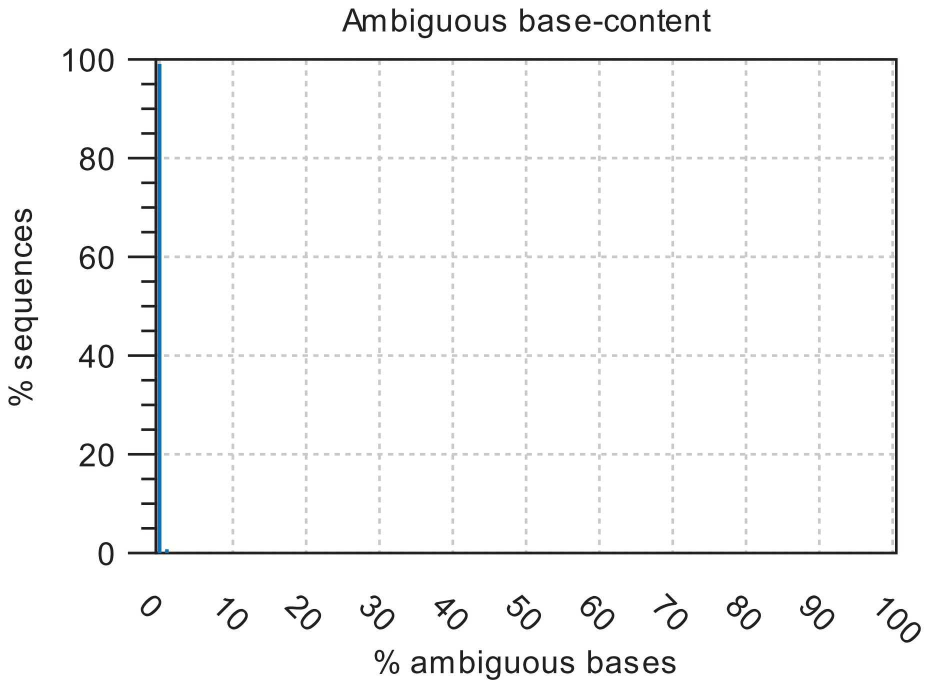
Distribution of sequence lengths. In cases of untrimmed Illumina or SOLiD reads it will ju st contain a single peak.
x: sequence length in base-pairs
y: number of sequences featuring a particular length normalized to the total number of seq uences

2.2 GC-content



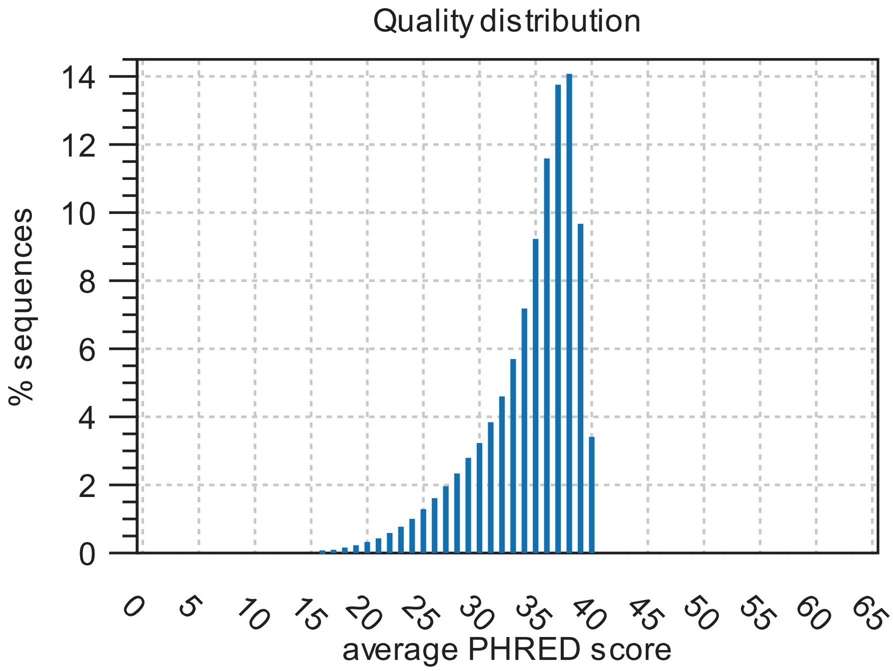
Distribution of GC-contents. The GC-content of a sequence is calculated as the number of G C-bases compared to all bases (including ambiguous bases).
x: relative GC-content of a sequence in percent
y: number of sequences featuring particular GC-percentages normalized to the total number of sequences

2.3 Ambiguous base-content



Distribution of N-contents. The N-content of a sequence is calculated as the number of amb iguous bases compared to all bases.
x: relative N-content of a sequence in percent
y: number of sequences featuring particular N-percentages normalized to the total number o f sequences

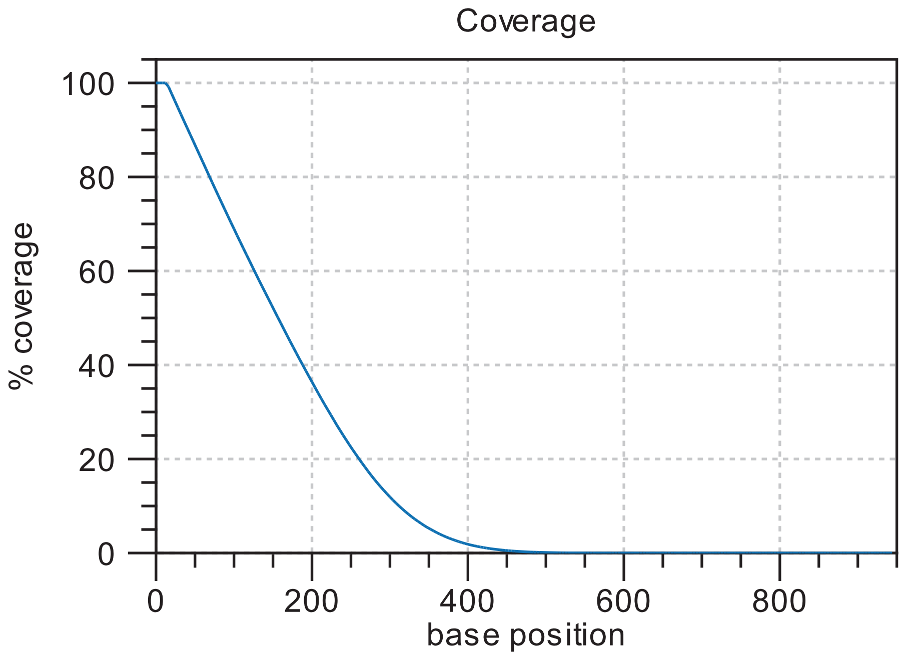
2.4 Quality distribution



Distribution of average sequence qualitie scores. The quality of a sequence is calculated as the arithmetic mean of its base qualities.
x: PHRED-score
y: number of sequences observed at that qual. score normalized to the total number of sequ ences

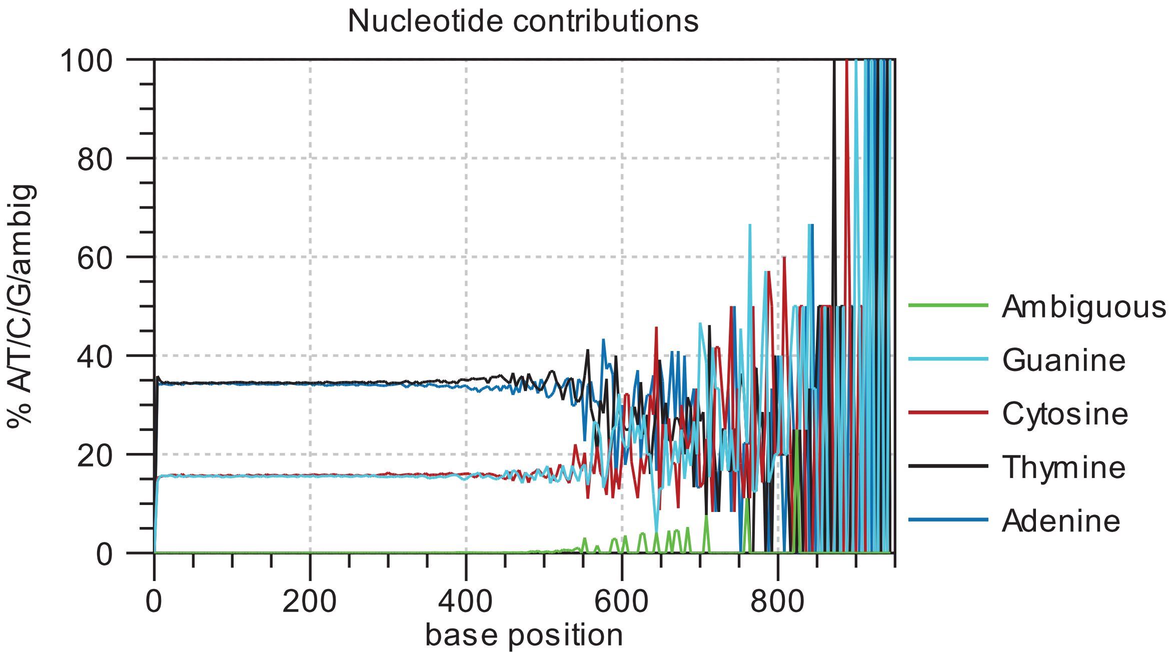
3. Per-base analysis

3.1 Coverage



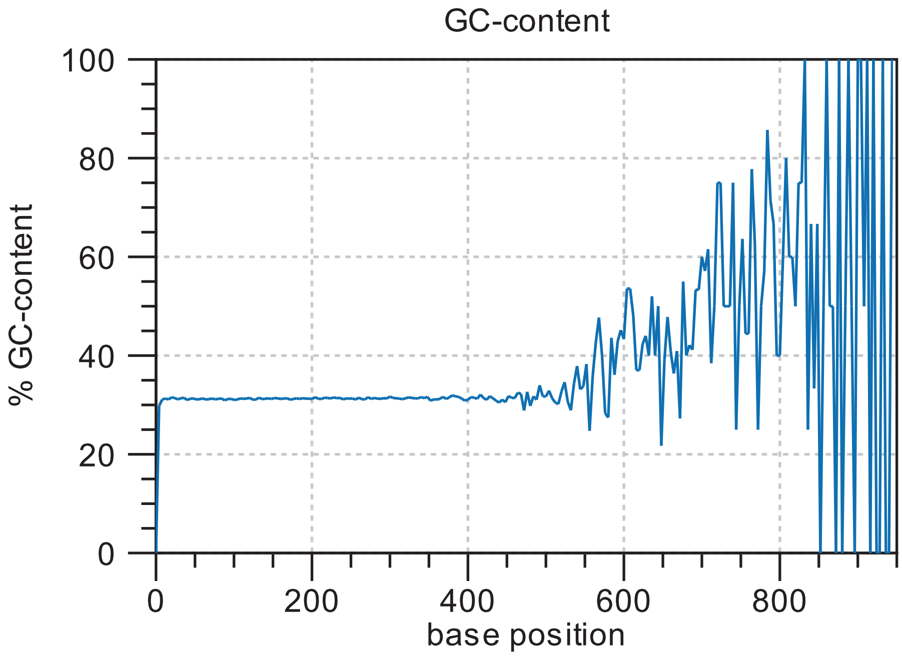
The number of sequences that support (cover) the individual base positions. In cases of un trimmed Illumina or SOLiD reads it will just contain a rectangle.
x: base position
y: number of sequences covering individual base positions normalized to the total number o f sequences

3.2 Nucleotide contributions



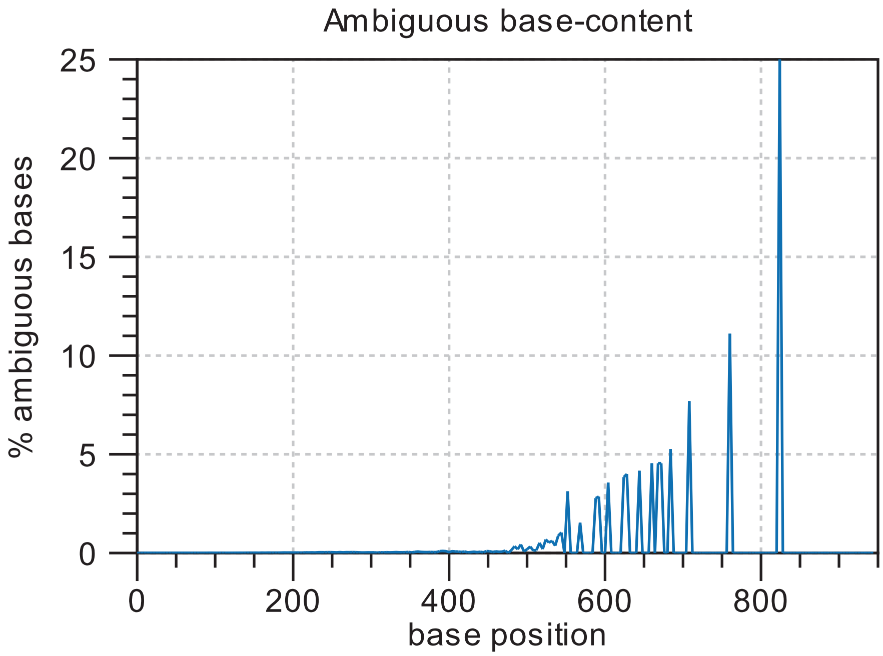
Coverages for the four DNA nucleotides and ambiguous bases.
x: base position
y: number of nucleotides observed per type normalized to the total number of nucleotides o bserved at that position

3.3 GC-content



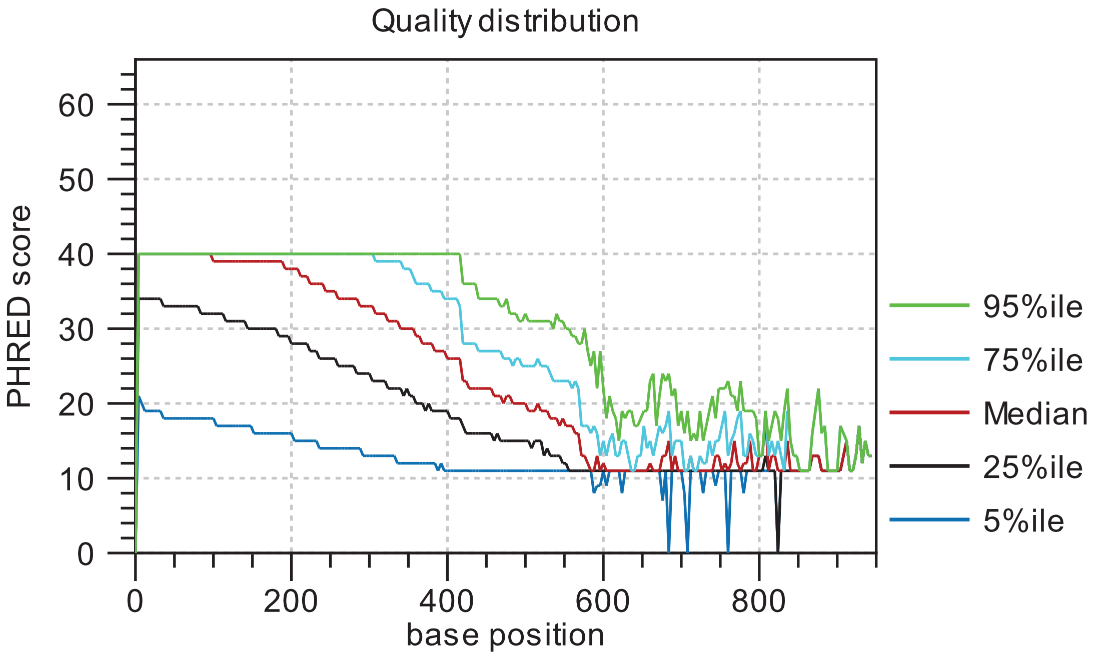
Combined coverage of G- and C-bases.
x: base position
y: number of G- and C-bases observed at current position normalized to the total number of bases observed at that position

3.4 Ambiguous base-content



Combined coverage of ambiguous bases.
x: base position
y: number of ambiguous bases observed at current position normalized to the total number o f bases observed at that position

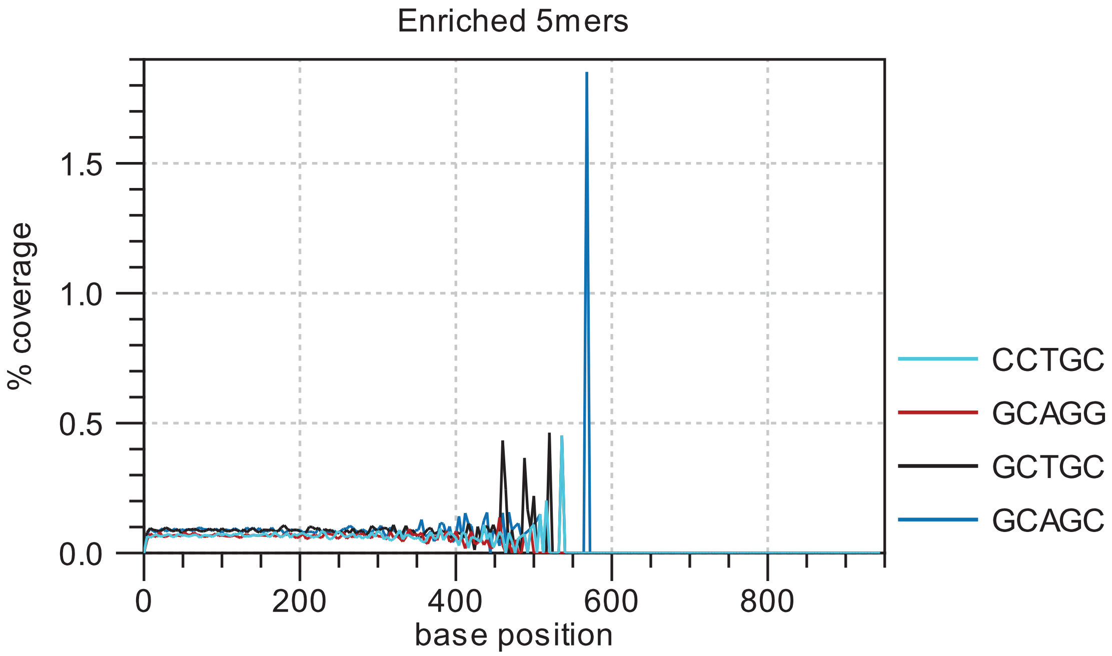
3.5 Quality distribution



Base-quality distribution along the base positions.
x: base position
y: median & percentiles of quality scores observed at that base position

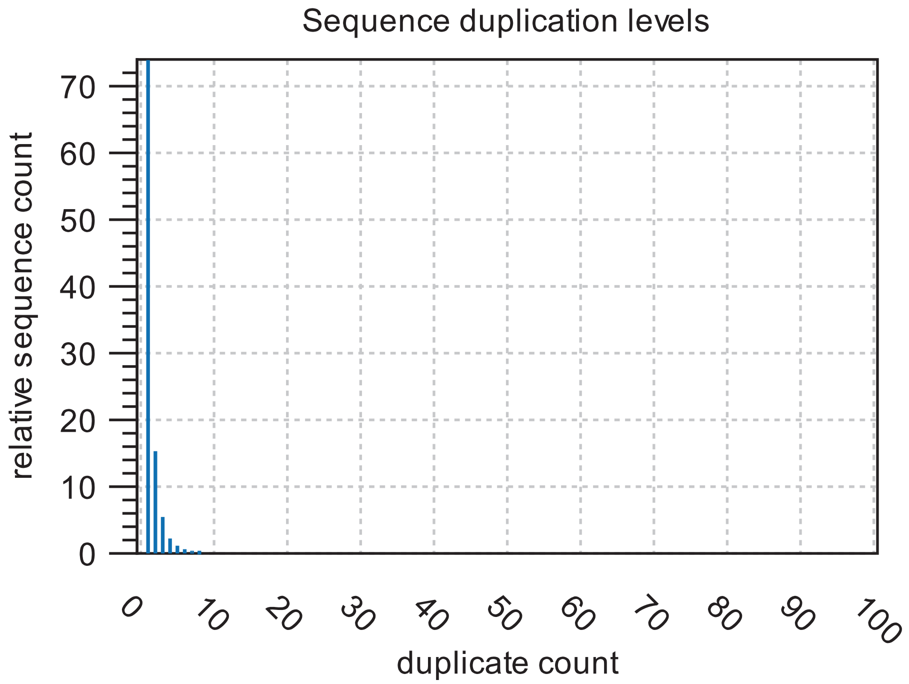
4. Over-representation analyses

4.1 Enriched 5mers



The five most-overrepresented 5mers. The over-representation of a 5mer is calculated as th e ratio of the observed and expected 5mer frequency. The expected frequency is calculated as product of the empirical nucleotide probabilities that make up the 5mer. (5mers that contain ambiguous bases are ignored)
x: base position
y: number of times a 5mer has been observed normalized to all 5mers observed at that posit ion

4.2 Sequence duplication levels



Duplication level distribution. Duplication levels are simply the count of how often a par ticular sequence has been found.
x: duplicate count
y: number of sequences that have been found that many times normalized to the number of un ique sequences

4.3 Duplicated sequences

A table of over-represented sequences is given in the supplementary report

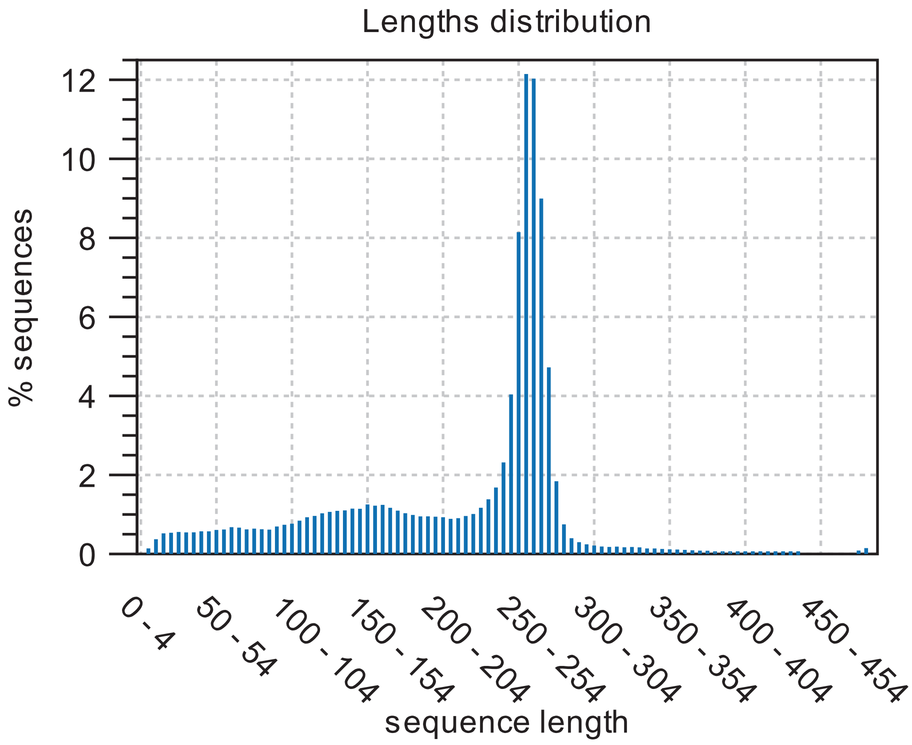
Ion Torrent data Quality Report

1. Summary

Creation date:	Fri Jan 09 12:50:19 EST 2015
Generated by:	uks
Software:	CLC Genomics Workbench 7.5.1
Based upon:	1 data set
Ion_Torrent:	453,686 sequences
Total nucleotides in data set	97,658,077 nucleotides

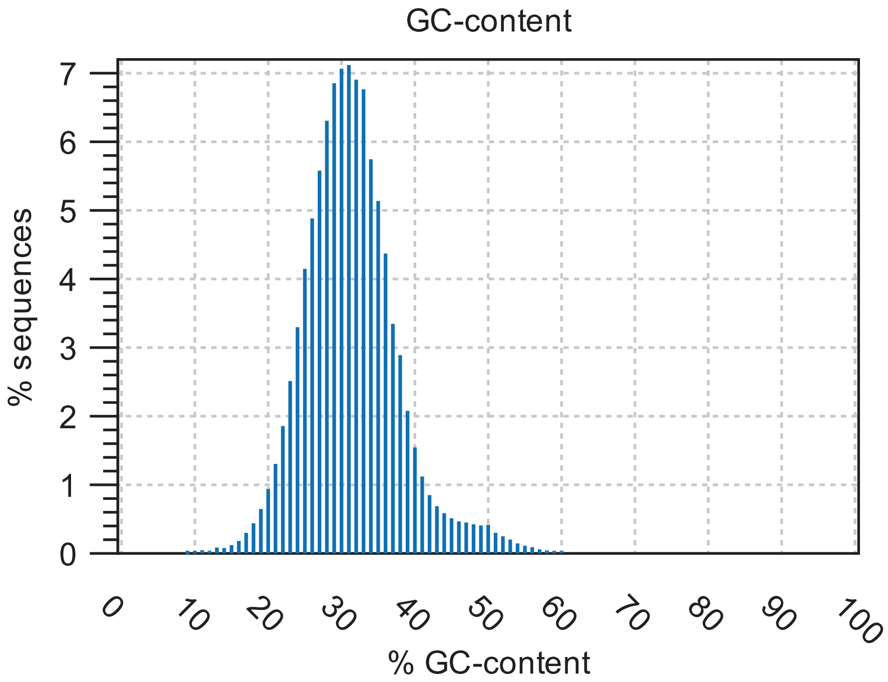
2. Per-sequence analysis

2.1 Lengths distribution



Distribution of sequence lengths. In cases of untrimmed Illumina or SOLiD reads it will ju st contain a single peak.
x: sequence length in base-pairs
y: number of sequences featuring a particular length normalized to the total number of seq uences

2.2 GC-content

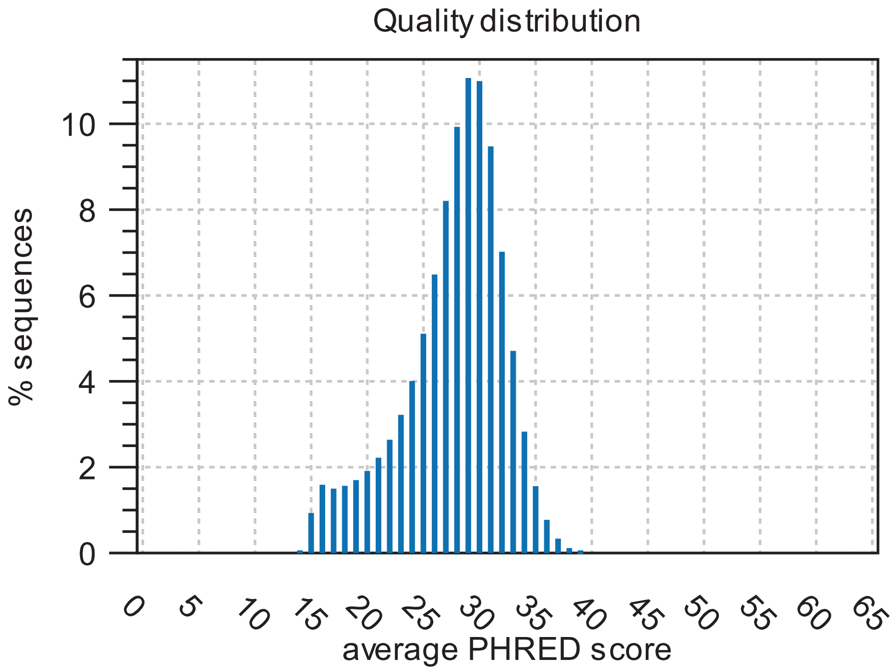


Distribution of GC-contents. The GC-content of a sequence is calculated as the number of G C-bases compared to all bases (including ambiguous bases).
x: relative GC-content of a sequence in percent
y: number of sequences featuring particular GC-percentages normalized to the total number of sequences

2.3 Ambiguous base-content

No ambiguous bases detected

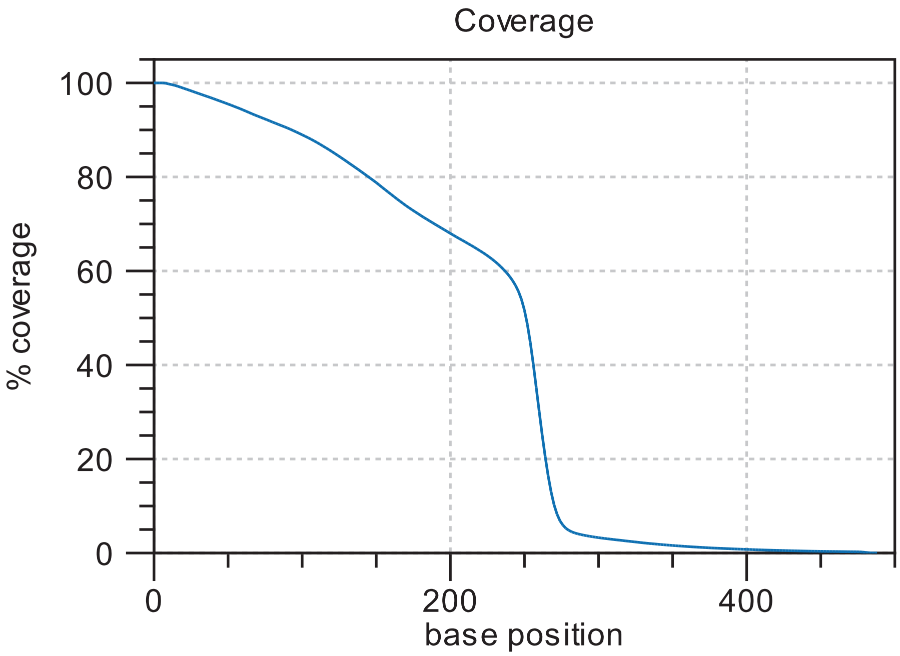
2.4 Quality distribution



Distribution of average sequence qualitie scores. The quality of a sequence is calculated as the arithmetic mean of its base qualities.
x: PHRED-score
y: number of sequences observed at that qual. score normalized to the total number of sequ ences

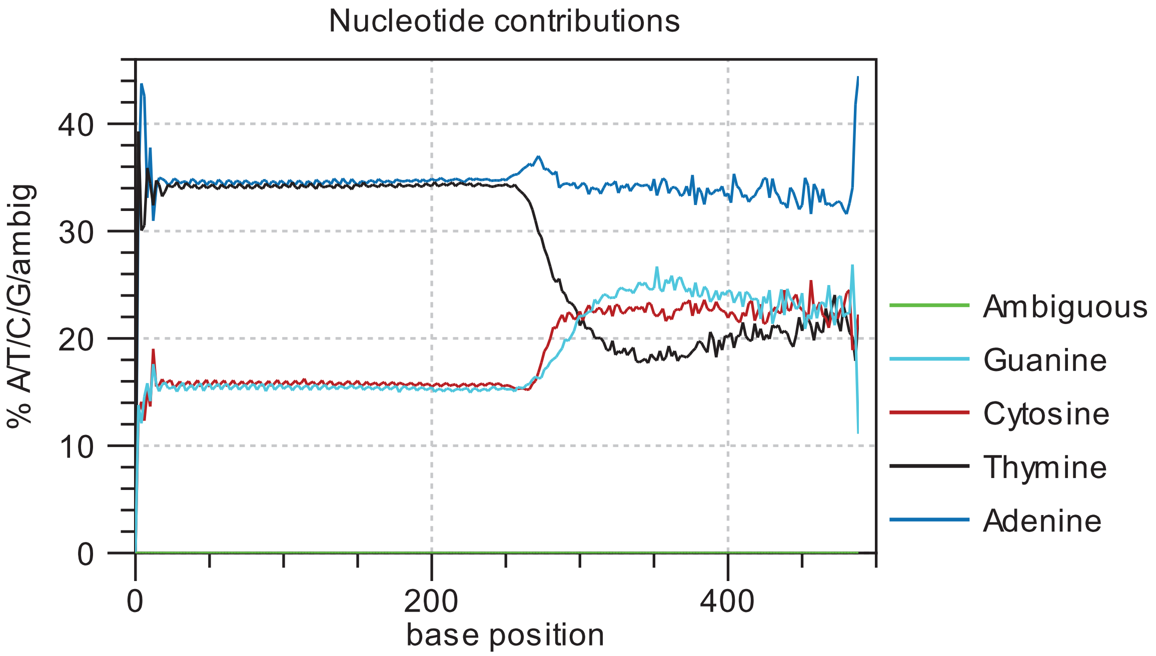
3. Per-base analysis

3.1 Coverage



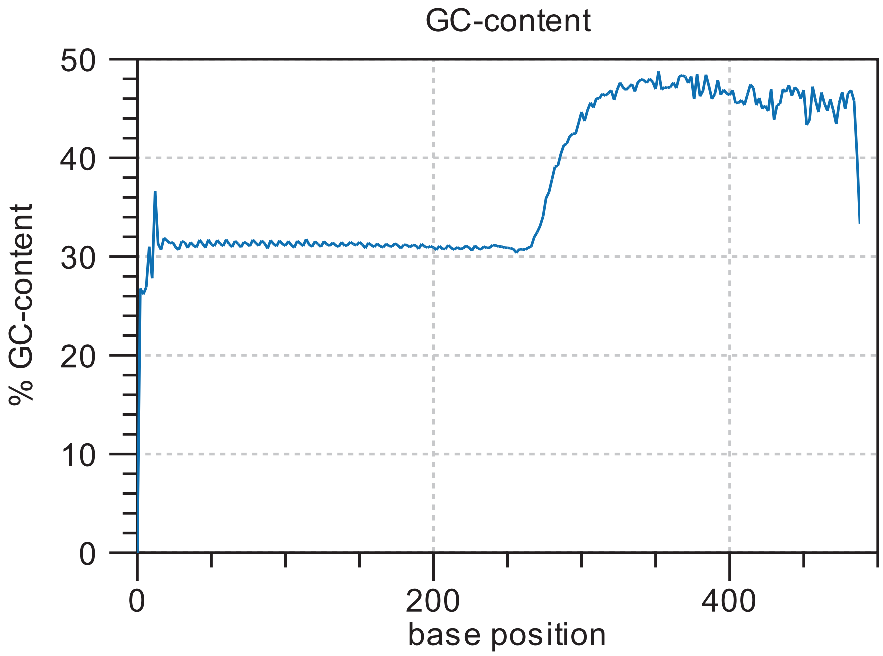
The number of sequences that support (cover) the individual base positions. In cases of un trimmed Illumina or SOLiD reads it will just contain a rectangle.
x: base position
y: number of sequences covering individual base positions normalized to the total number o f sequences

3.2 Nucleotide contributions



Coverages for the four DNA nucleotides and ambiguous bases.
x: base position
y: number of nucleotides observed per type normalized to the total number of nucleotides o bserved at that position

3.3 GC-content

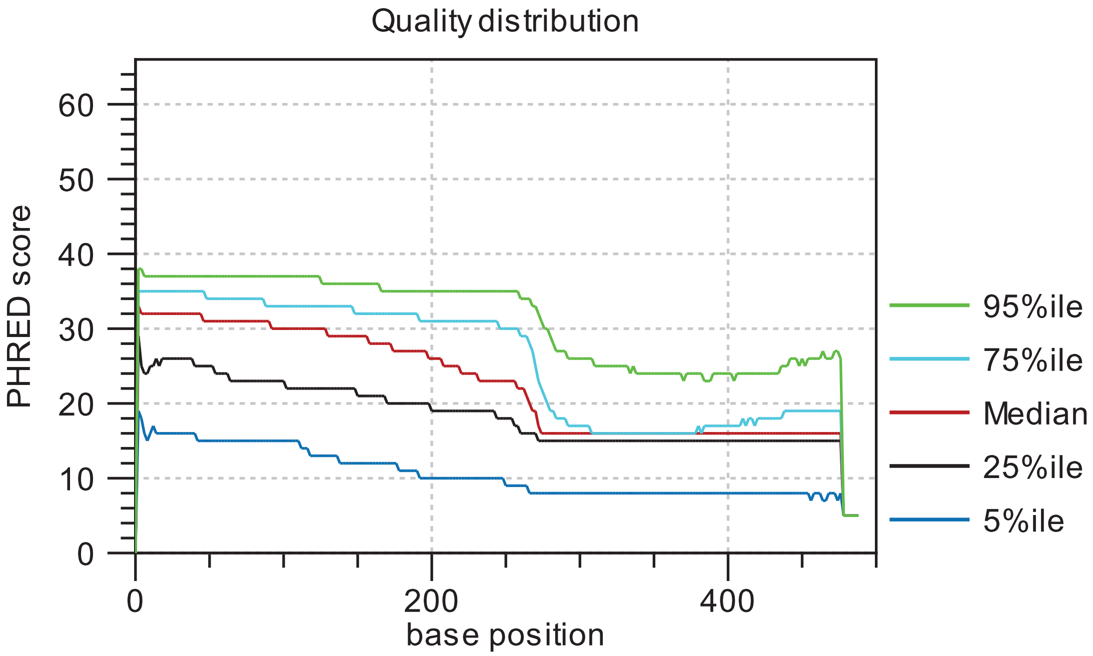


Combined coverage of G- and C-bases.
x: base position
y: number of G- and C-bases observed at current position normalized to the total number of bases observed at that position

3.4 Ambiguous base-content

No ambiguous bases detected

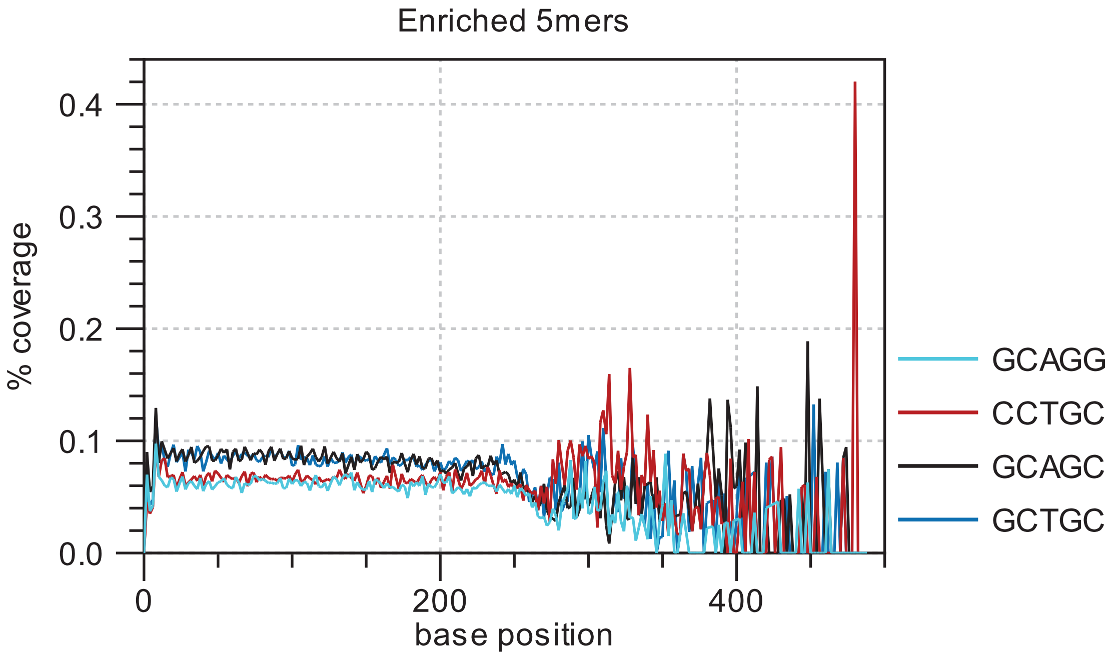
3.5 Quality distribution



Base-quality distribution along the base positions.
x: base position
y: median & percentiles of quality scores observed at that base position

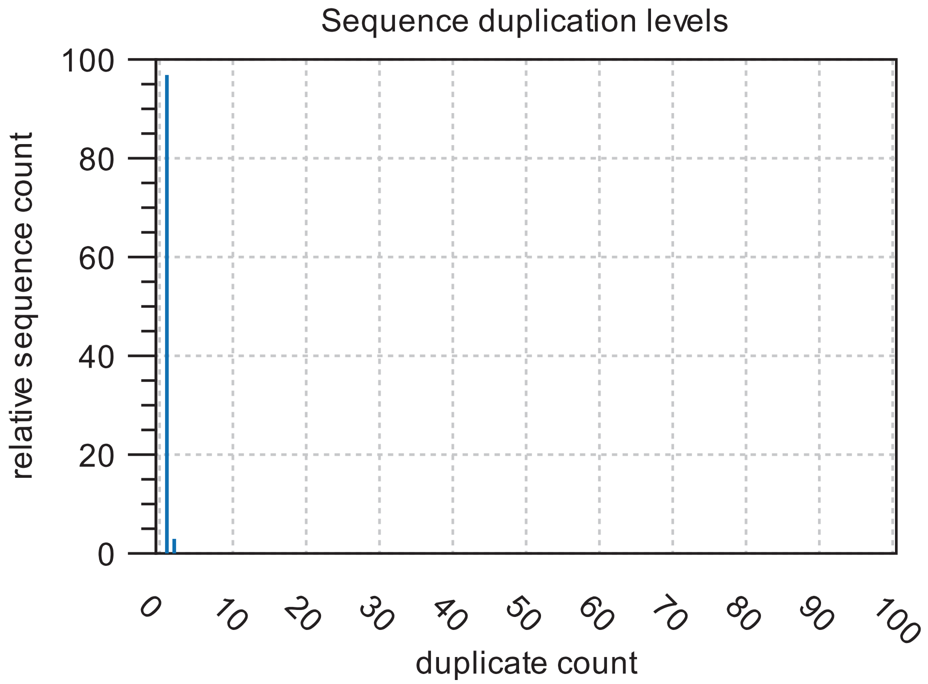
4. Over-representation analyses

4.1 Enriched 5mers



The five most-overrepresented 5mers. The over-representation of a 5mer is calculated as th e ratio of the observed and expected 5mer frequency. The expected frequency is calculated as product of the empirical nucleotide probabilities that make up the 5mer. (5mers that contain ambiguous bases are ignored)
x: base position
y: number of times a 5mer has been observed normalized to all 5mers observed at that posit ion

4.2 Sequence duplication levels



Duplication level distribution. Duplication levels are simply the count of how often a par ticular sequence has been found.
x: duplicate count
y: number of sequences that have been found that many times normalized to the number of un ique sequences

4.3 Duplicated sequences

A table of over-represented sequences is given in the supplementary report

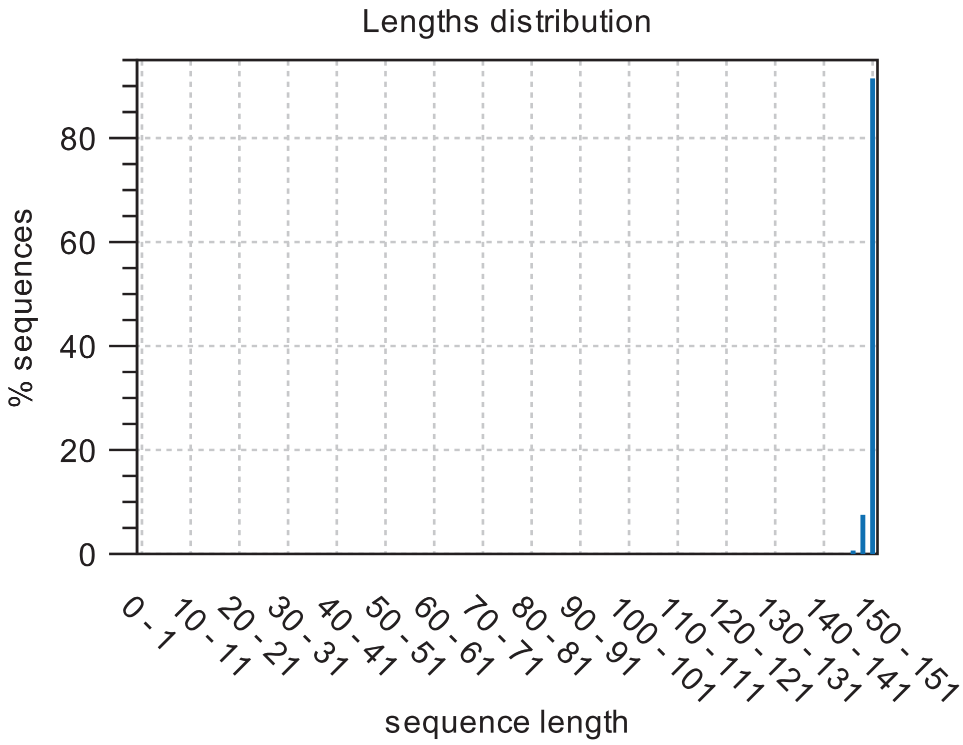
Illumina data Quality Report

1. Summary

Creation date:	Thu Dec 11 14:05:57 EST 2014
Generated by:	uks
Software:	CLC Genomics Workbench 7.5.1
Based upon:	1 data set
DSMZ_R1 (paired):	3,689,644 sequences in pairs
Total nucleotides in data set	554,448,529 nucleotides

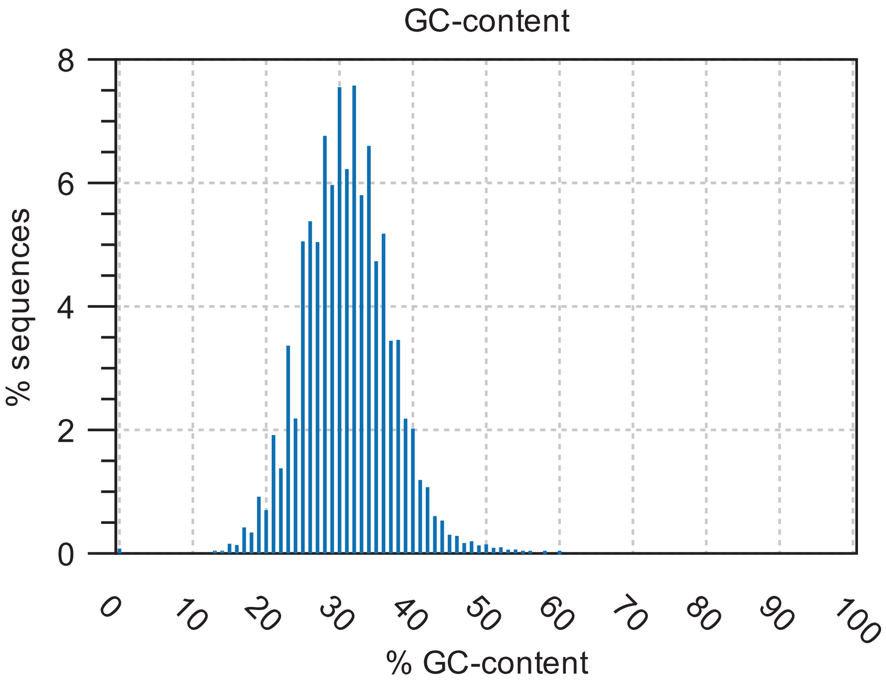
2. Per-sequence analysis

2.1 Lengths distribution



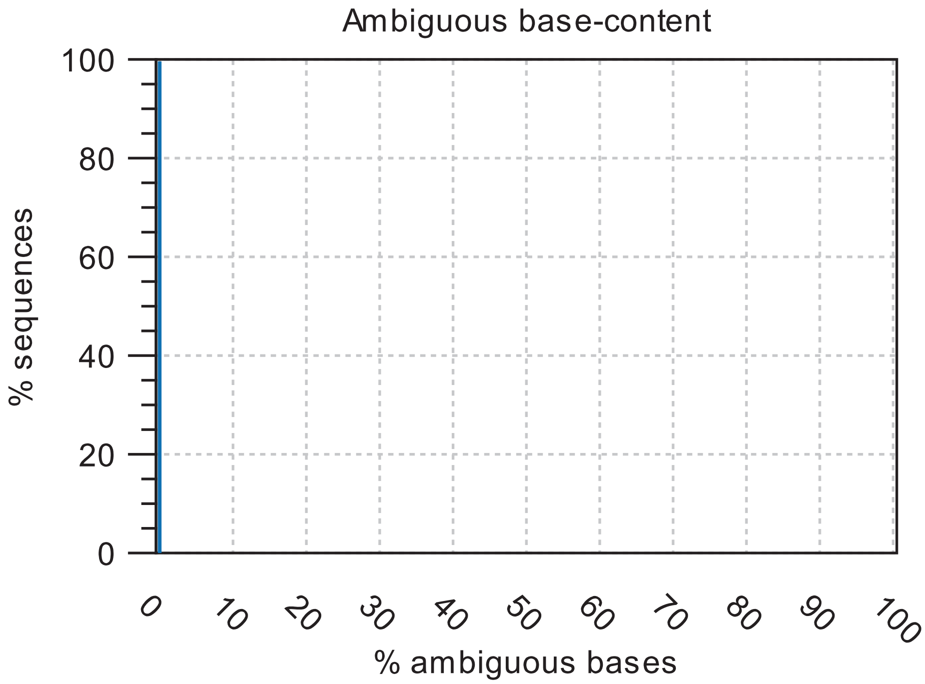
Distribution of sequence lengths. In cases of untrimmed Illumina or SOLiD reads it will ju st contain a single peak.
x: sequence length in base-pairs
y: number of sequences featuring a particular length normalized to the total number of seq uences

2.2 GC-content



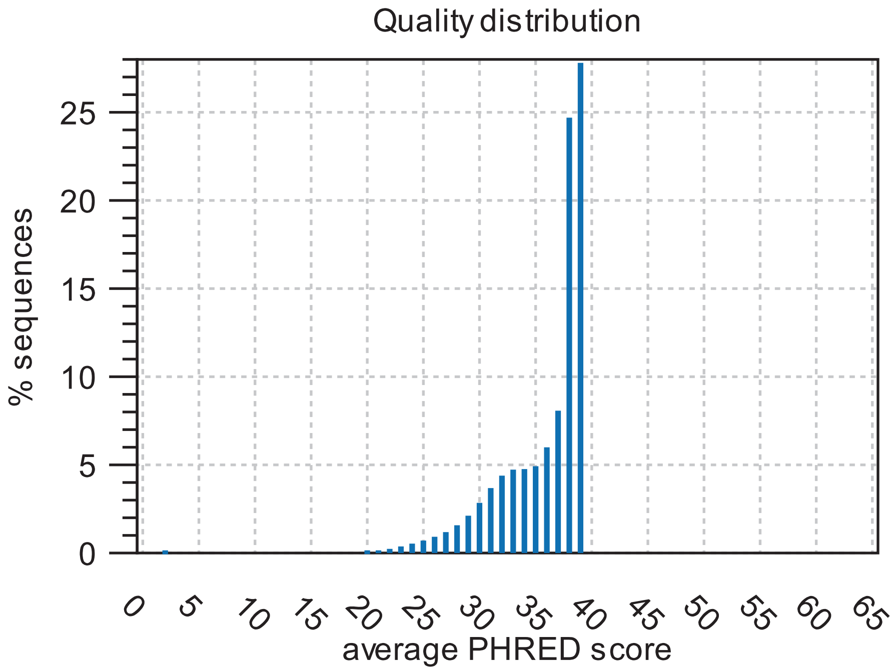
Distribution of GC-contents. The GC-content of a sequence is calculated as the number of G C-bases compared to all bases (including ambiguous bases).
x: relative GC-content of a sequence in percent
y: number of sequences featuring particular GC-percentages normalized to the total number of sequences

2.3 Ambiguous base-content



Distribution of N-contents. The N-content of a sequence is calculated as the number of amb iguous bases compared to all bases.
x: relative N-content of a sequence in percent
y: number of sequences featuring particular N-percentages normalized to the total number o f sequences

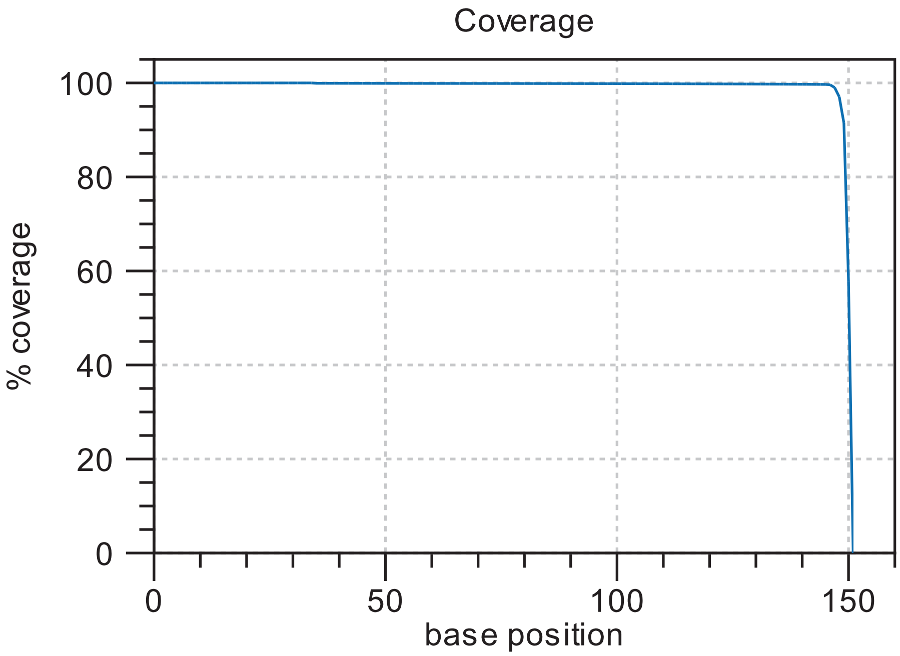
2.4 Quality distribution



Distribution of average sequence qualitie scores. The quality of a sequence is calculated as the arithmetic mean of its base qualities.
x: PHRED-score
y: number of sequences observed at that qual. score normalized to the total number of sequ ences

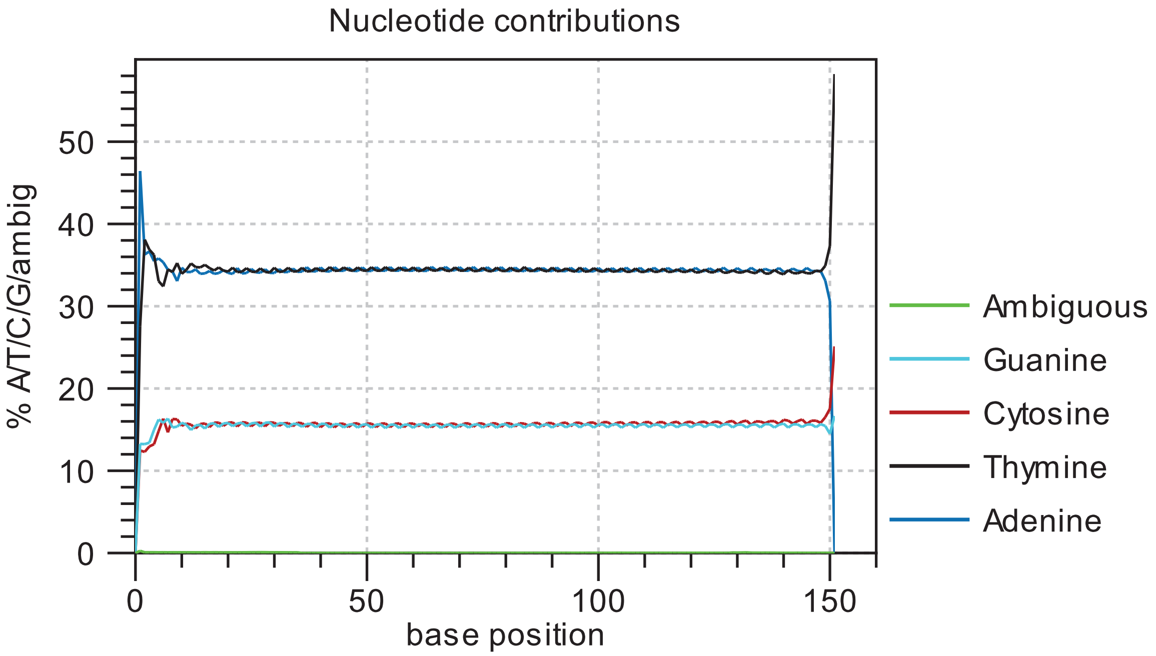
3. Per-base analysis

3.1 Coverage



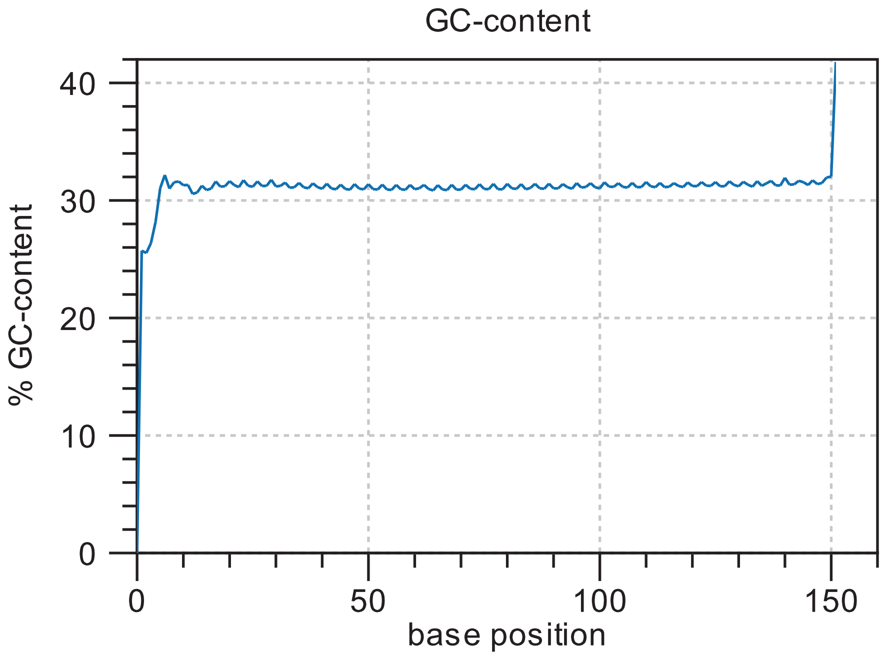
The number of sequences that support (cover) the individual base positions. In cases of un trimmed Illumina or SOLiD reads it will just contain a rectangle.
x: base position
y: number of sequences covering individual base positions normalized to the total number o f sequences

3.2 Nucleotide contributions



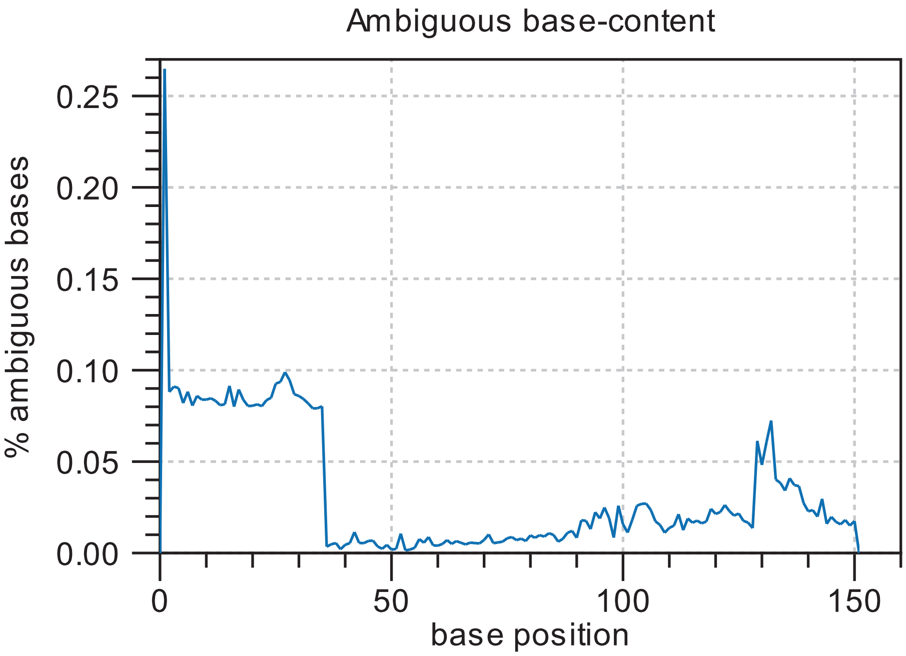
Coverages for the four DNA nucleotides and ambiguous bases.
x: base position
y: number of nucleotides observed per type normalized to the total number of nucleotides o bserved at that position

3.3 GC-content



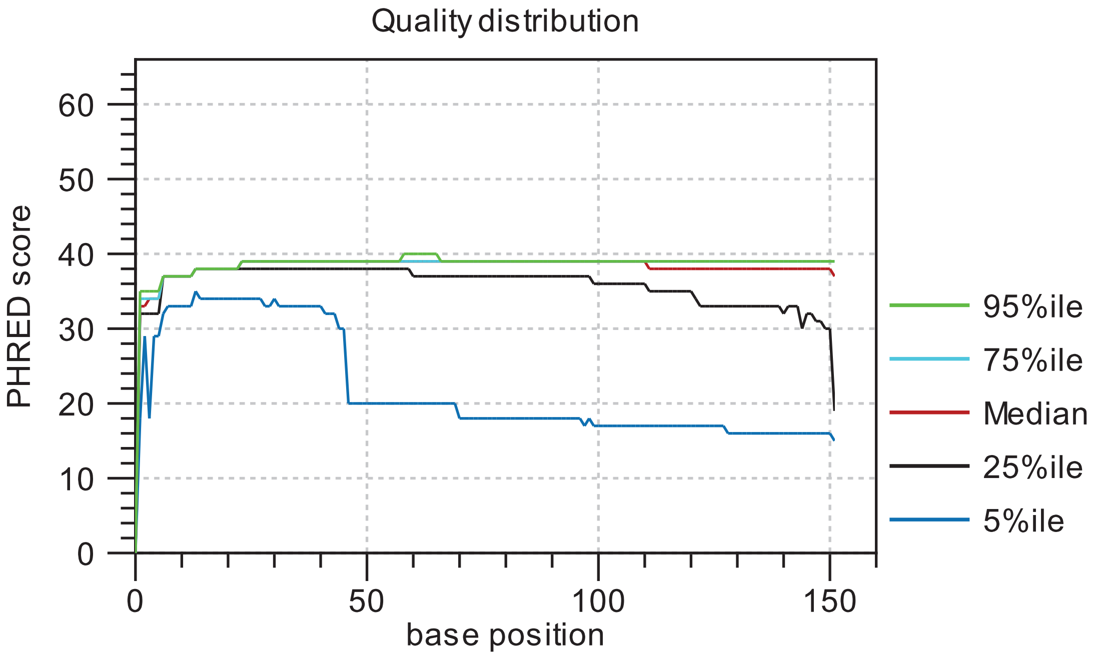
Combined coverage of G- and C-bases.
x: base position
y: number of G- and C-bases observed at current position normalized to the total number of bases observed at that position

3.4 Ambiguous base-content



Combined coverage of ambiguous bases.
x: base position
y: number of ambiguous bases observed at current position normalized to the total number o f bases observed at that position

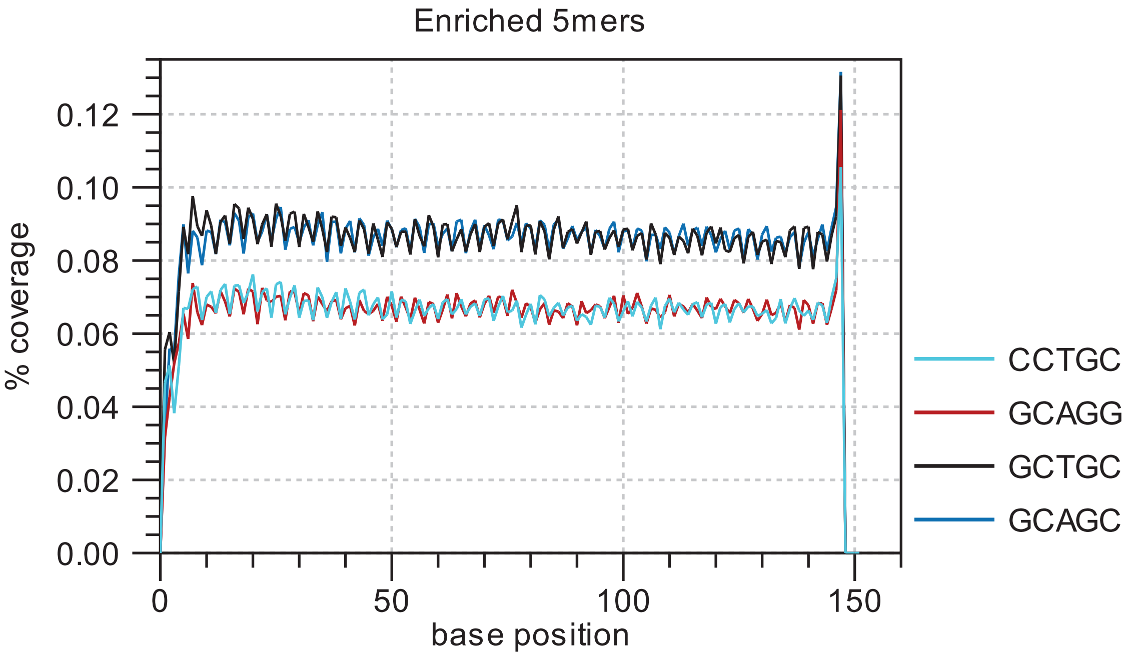
3.5 Quality distribution



Base-quality distribution along the base positions.
x: base position
y: median & percentiles of quality scores observed at that base position

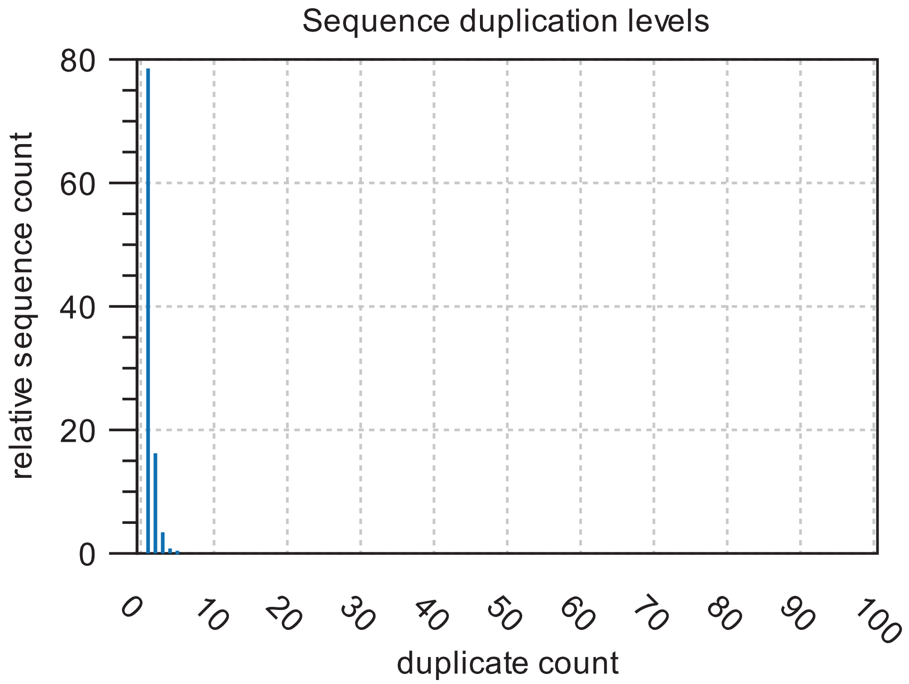
4. Over-representation analyses

4.1 Enriched 5mers



The five most-overrepresented 5mers. The over-representation of a 5mer is calculated as th e ratio of the observed and expected 5mer frequency. The expected frequency is calculated as product of the empirical nucleotide probabilities that make up the 5mer. (5mers that contain ambiguous bases are ignored)
x: base position
y: number of times a 5mer has been observed normalized to all 5mers observed at that posit ion

4.2 Sequence duplication levels



Duplication level distribution. Duplication levels are simply the count of how often a par ticular sequence has been found.
x: duplicate count
y: number of sequences that have been found that many times normalized to the number of un ique sequences

4.3 Duplicated sequences

A table of over-represented sequences is given in the supplementary report

Guide for downloading the described datasets using NCBI SRA Toolkit

Pre-requisites:

1. Appropriate version of NCBI SRA Toolkit Installed on the system.
2. Following downloading instructions are provided for the Linux based operating systems.

Steps:

1. Launch Linux terminal and navigate to NCBI SRA Toolkit installed location.
2. Navigate to /bin/ directory.
3. Respective commands for each dataset are mentioned below:

Download Illumina PE dataset with accession SRR989790

```
fastq-dump -I --split-files SRR989790
```

Download 454 3 Kb PE dataset with accession SRR989497

```
sff-dump SRR989497
```

Download 454 shotgun dataset with accession SRR1748017

```
sff-dump SRR1748017
```

Download Ion Torrent dataset with accession SRR1748018

```
sff-dump SRR1748018
```

Download PacBio RS II dataset with accession SRR989791

```
fastq-dump SRR989791
```

The dataset size and md5 checksum values for each dataset is provided in table below:

Table S1 – Dataset properties after download from SRA.

Data Type	Accession	File names	Size	md5 checksum
Roche 454 shotgun	SRR1748017	SRR1748017.sff	1.5 Gb	950975500428cc5cd91b9a03dbafd877
Roche 454 3 kb	SRR989497	SRR989497.sff	1.4 Gb	4d9f35e94ff9625980d853df63a1cb24
Illumina	SRR989790	SRR989790_1.fastq	669 Mb	750352f0fbccc593dce4c3eaacbae6e0
		SRR989790_2.fastq	669 Mb	fa8dfd498d3f16f94541ebfb3f9d5859
Ion Torrent	SRR1748018	SRR1748018.sff	858 Mb	a17df0c66161dea8d65bdeace023c135
PacBio RS II	SRR989791	SRR989791.fastq	1.5 Gb	0eb1b192bf1f9b6c141903787fa38bf9