Genome scan

Next Generation Sequencing

Bioinformatics data Quality Control

Guidelines

genomescan.nl

GenomeScan's Guidelines for Data Quality Assessment on NGS Data

Using our proprietary and validated Data Analysis Pipelines

Dear customer,

As of the beginning of 2015 ServiceXS became a trademark of GenomeScan B.V. GenomeScan focuses exclusively on Molecular Diagnostics whereas our ServiceXS trademark is intended for your R&D projects.

GenomeScan is dedicated to help you design and perform Next Generation Sequencing (NGS) experiments that generate high quality results. This guide provides information for our data analysis services and resources and tools for further analysis of your sequencing data. NGS experiments result in vast amounts of data and therefore data analysis can be challenging. Our ability to assist in the analysis of your results can be the key factor leading to a successful project.

The following document provides an initial starting point to evaluate the quality of your next generation sequencing data. Keep in mind that each sequencing platform generates a specific type of data with its own characteristics, quality, and error rates. Therefore, certain quality metrics may not apply to a particular platform.

GenomeScan provides a comprehensive package of bioinformatics services for our next-generation sequencing customers, which enable them to utilise all the applications that are possible with billions of bases of sequence data per run. GenomeScan can advise and assist you in every step of the data analysis. Do not hesitate to contact us if you have any questions after reading this guideline!

On behalf of the GenomeScan team,

Thomas Chin-A-Woeng
Project Manager

Document Outline

Changes to Previous Version (1.0)
-Layout changes

Chapter 1 Introduction

Next-generation sequencing (NGS, also high-throughput sequencing) makes it possible to generate vast amounts of sequencing data in comparison to the Sanger sequencing method. The large amount of data generated and new types of applications based on this technology offer novel challenges to scientists and bioinformaticians. One run with a next-generation sequencer typically generates several megabases to hundreds of gigabases of DNA sequences.

A major factor determining the quality of the final outcome of a next-generation sequencing project is proper quality checking and filtering of the data. Based on these quality metrics we decide whether the data set and what part of the data set is optimal for further processing. Quality checks and quality scores at each level of the work flow ensures minimal errors in the final outcome. Summary information of the general quality and distribution of errors are generated in comprehensive charts with clear descriptions how to interpret the results.

GenomeScan has many years of experience in data analysis for its portfolio of services. As one of world's first service providers for next generation sequencing GenomeScan has gathered a vast amount of knowledge and tools to handle the whole process of transforming raw sequence data obtained from sequencers such as Illumina, Pacific Biosciences, and Life Technologies, into easily intelligible reports and interpretable file formats.

Visit our website, download our detailed guidelines or contact us for more information.

1.1 Quality score analysis

Most of the current sequencers produce output in Sanger FASTQ format. The FASTQ files contain all the bases and a quality score (or Q-score) attached to each base. These quality scores are in a certain format, in which a character represents its corresponding Q-score.
A Q-score expresses an error probability. In particular, it serves as a convenient and compact way to convey very small error probabilities.

Given an assertion, A, the probability that A is not true, P(~A), is expressed by a quality score, Q(A), according to the relationship: Q(A) =-10 log10(P(~A)) where *P(~A)* is the estimated probability of an assertion A being wrong. The relationship between the quality score and error probability is demonstrated with the following table:

Table 1. Relationship between quality score and error probability

Quality score, Q(A)	Error probability P(~A)
10	0.1
20	0.01
30	0.001

Chapter 2 Quality Metrics

After sequence files are generated, the statistics of the quality of the sequence data is summarised per sample. The following section describes some general quality metrics that can be used to evaluate and characterise the sequence data. Not all metrics are relevant for the different sequencing platforms.

2.1 Quality scores per cycle

The Quality Score Distribution shows an overview of the range of quality values across all bases at each cycle in the read (base position in the FASTQ file). The x-axis shows the cycle number (position in read) and the y-axis the mean or median of the quality scores. The higher the score the better the quality of the base calls.

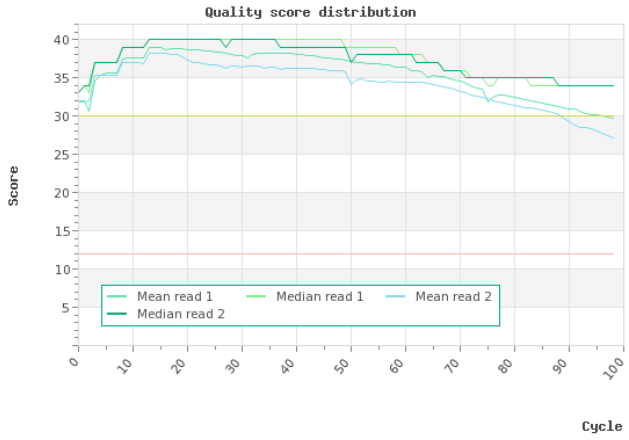

Fig. 1. Quality score distribution

2.2 Mean Probabilities of Error

The Mean Probabilities of Error plot shows the mean probabilities of base calling error by cycle (position in the read). The x-axis shows the cycle number and the y-axis the cumulative mean of the probability of error. The lower the score the better the quality of the reads.

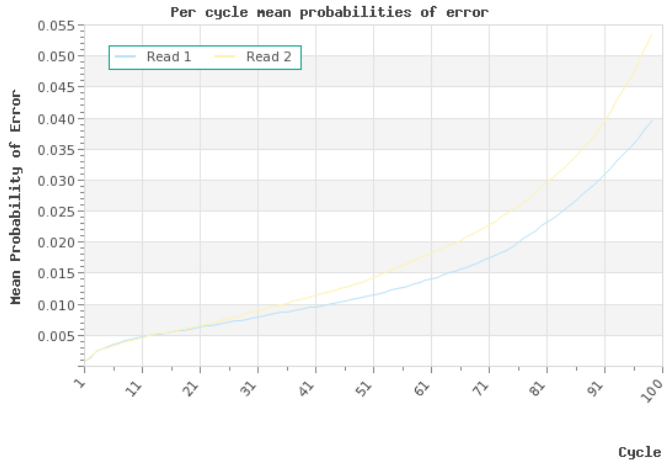

Fig. 2. Mean probability of error

2.3 Per Sequence Qualities

The Per Sequence Quality Distribution indicates whether a subset of the reads have low average quality values. This might be the case if part of the flow cell was poorly imaged due to stains, bubbles or other technical issues. The number of reads with low quality should represent only a small percentage of the total number of reads.

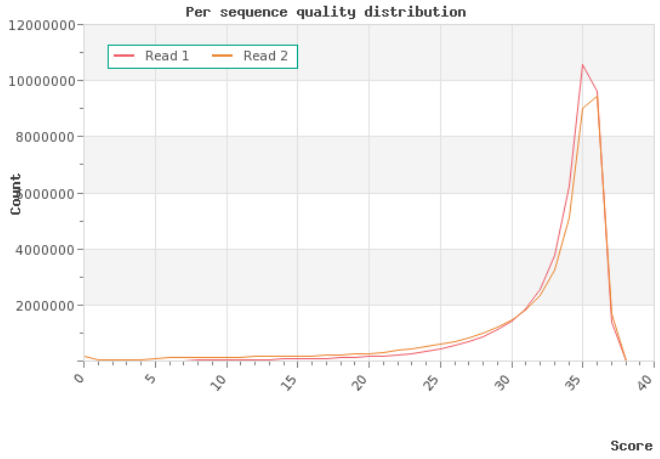

Fig. 3. Per sequence quality distribution

2.4 Per Cycle Base Composition

The Per Cycle Base Composition figure shows the proportion of each base per cycle (position in the read). In a random library one would expect little to no difference between the cycles in the read and the lines in this plot should run parallel with each other. The relative amount of each base should reflect the overall base composition in your genome. Strong biases usually indicates an overrepresented sequence which is contaminating the library. A bias which is consistent across all bases either indicates that the original library was sequence biased, or a systematic problem during the sequencing of the library.

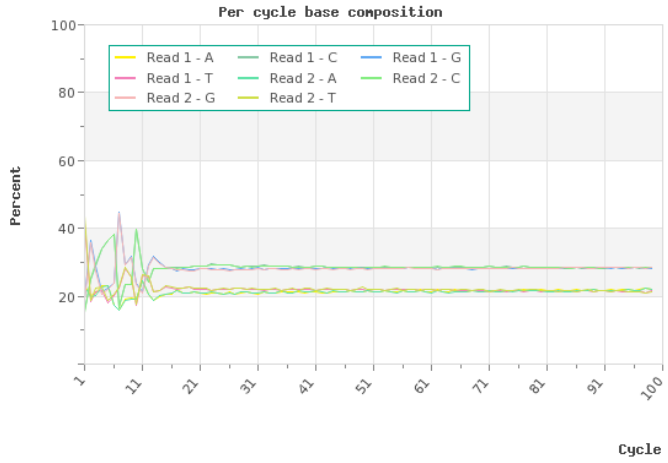

Fig. 4. Per cycle base composition

2.5 Per Cycle %GC

The Per Base %GC figure shows the GC content during each cycle (base position in read). In a random library one would expect little to no difference between the different bases of a sequence run, so

the line in this plot should run horizontally across the graph. The overall GC content should reflect the GC content of the organism. A GC bias which changes in different bases could indicate an overrepresented sequence which is contaminating the library. A bias which is consistent across all bases either indicates that the original library was sequence biased, or that there was a systematic problem during the sequencing of the library.

2.6 Per Sequence %GC

The Per Sequence %GC measures the GC content across the read. In a normal random library one expects a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the organism. 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.

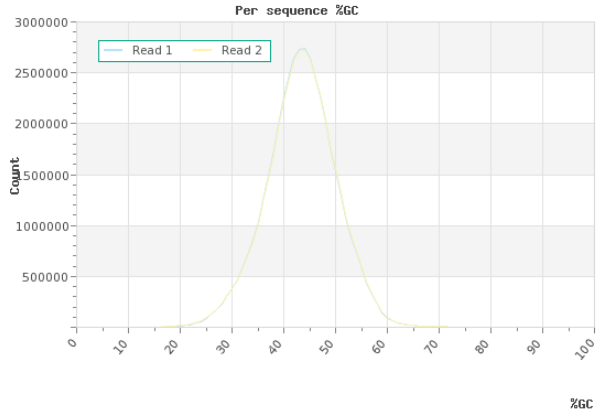

Fig. 5. Per sequence %GC.

2.7 Per Cycle N bases

If a base call cannot be made with sufficient confidence an N will be substituted. The Per Cycle N bases figure shows the percentage of base calls at each cycle or position for which an N was called.

2.8 Quality Score Distribution

The Quality Score Distribution plot shows the distribution of quality scores of the bases in the entire set of reads. The x-axis shows the Q-score and the y-axis the absolute read count. More instances plotted in the outer right part of the graph mean better quality of the data set.

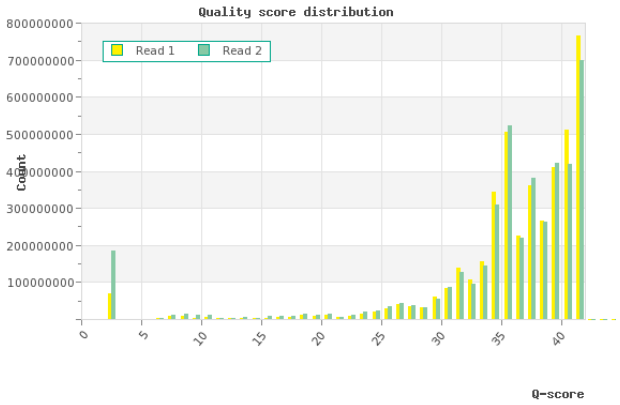

Fig. 6. Quality score distribution.

2.9 Read Score Distribution

The Read Score Distribution plot shows the distribution of the average quality score of reads in the entire set of reads. The x-axis shows the Q-score and the y-axis the absolute read count. More reads plotted in the outer right part of the graph mean better quality reads.

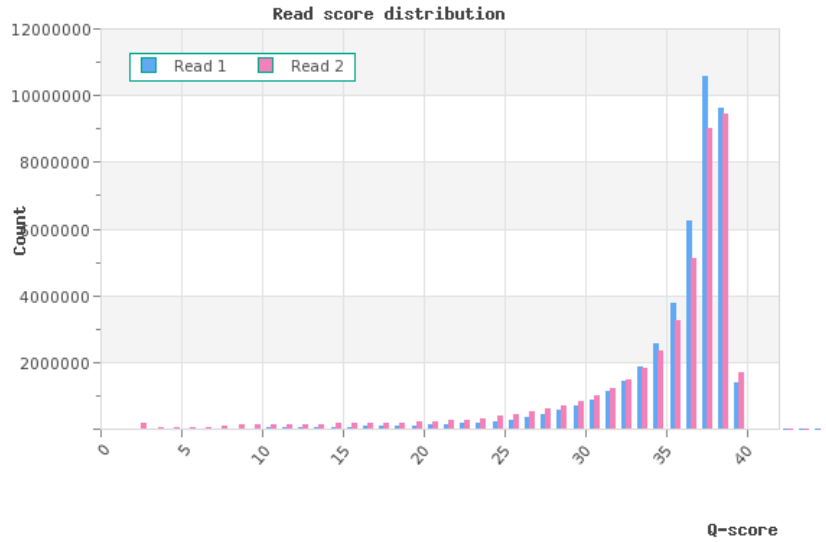

Fig. 7. Read score distribution.

2.10 Trimmed Lengths

The Trimmed Length or Longest Segment plot displays the proportion of all reads that produce a segment of a given length when the longest contiguous section of each read in which the quality of each base is better to the cutoff of p = 0.05 (Q13) is taken.

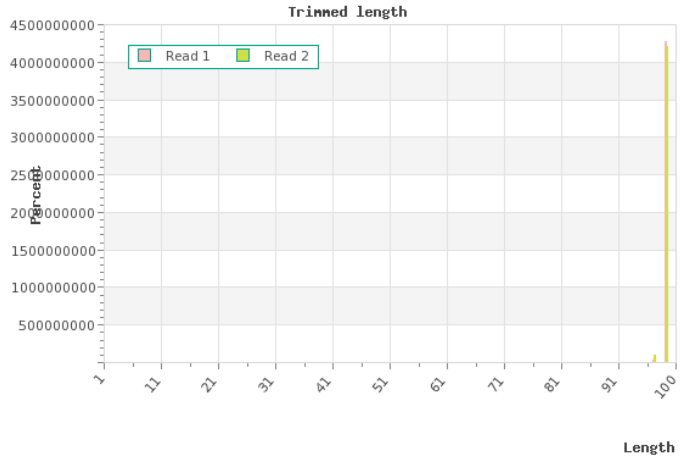

Fig. 8. Trimmed lengths.

Chapter 3 Data Filtering

Each next-generation sequencer generates data of different characteristics. Depending on the type of sequence data and application it may be necessary to further polish the sequencing data on top of the basic quality filter available via the manufacturer of sequencer. Often, the default sequencing software removes the worst sequences from the data set during primary analysis, but depending on the application (e.g. mutation analysis or expression analysis) additional quality filtering may be required. We perform quality assurance on sequencing data using in-house techniques and software tools while keeping quality measures for a particular purpose in mind. Data filtering may also differ per sequencing platform. E.g. Illumina/Solexa technology typically generates reads with a biased first base and quality/reliability quickly drops after certain read lengths.

3.1 Quality filtering

GenomeScan uses in-house tools to pre-process sequence reads prior to its downstream data analysis pipelines. FASTQ files are read, checked for integrity and a new set of filtered data files is generated meeting the filtering criteria defined for a particular type of analysis. Data quality can differ significantly between runs.

Reads with excessive low quality bases or uncalled bases (Ns) can be removed from your data set according certain filtering algorithms and thresholds by trimming reads at both the 5' and/or 3' end or discarding the read. Reads can be discarded if they are too short, or when a percentage of quality scores is lower than the indicated threshold. More advanced trimming allows separate thresholds for each quality score distribution (e.g.100%>Q10, 80%>Q20, 70%>Q30 for a read) to be defined.

3.2 Barcode sorting

When using custom bar coding reads can be demultiplexed into separate file based on the barcode in the sequence read or index read. Both standard barcoding as well as in-read barcodes and dual barcoding is supported.

3.3 Adapter identification

Primer sequences due to adapter-adapter ligations or read-through of the insert can be trimmed from the reads. Our algorithms can determine adapter sequences accurately at single base level assuring that not a single base of adapter sequence can interfere with downstream processing (e.g. assembly). Known adapter sequences can be removed from both 5' and 3' ends.

3.4 Size trimming

Fixed size trimming of reads at both 5' and 3' ends can be performed. Quality score-based trimming at 5' and 3' end leaving the longest part of the read of good quality is the most common means of filtering for our pipelines. Minimum lengths for the remaining sequence read and minimum base quality scores can be defined.

3.5 Error correction

Using coverage information errors in reads can be repaired. Roche 454/IonTorrent/IonProton reads contain unreliable polymeric tracts that can be repaired using this coverage information. DNA sequences can also be improved by combining multiple sequencing platforms.

3.6 Statistics

In addition to filtering the sequence files, general statistics are gathered and calculated in terms of the number of accepted or rejected reads and bases, quality and length distribution, and total counts. All is provided in a number of tables and charts. Summary information of the general quality and distribution of errors can be generated.

Notes

Caring for your future