Novel Tools and Methods for Exploring Pyrosequencing Data

Including Quality Assessment and Simulation

Susanne Balzer[1,2], Ketil Malde[1], Inge Jonassen[2,3]

1 - Institute of Marine Research, Bergen. 2 - Department of Informatics, University of Bergen. 3 - Computational Biology Unit, Bergen Center for Computational Science, University of Bergen.

Objectives

Get to know the data used for assembly.
Improve base-calling.
Identify criteria for a "good read".
Rank reads after quality.
Provide a more straightforward assembly.
Benchmark assembly methods.

454 Pyrosequencing Technology

454 Sequencing, based on sequencing-by-synthesis, involves the following steps[1]:

- Nucleotides are flowed across the plate sequentially in a fixed order ("TACG") during a sequencing run.
- Up to a million beads each carrying millions of copies of a unique single-stranded DNA molecule are sequenced in parallel. Each bead produces one read.
- If a nucleotide complementary to the template strand is flowed into a well, the polymerase extends the existing DNA strand by adding nucelotide(s).
- Addition of one (or more) nucleotide(s) results in a reaction that generates a light signal.
- The signal strength is proportional to the number of nucleotides incorporated in a single nucleotide flow. It thus designates the length of a homopolymer run.
- For each bead, signal intensities are recorded by a camera over time. They can be used to plot a flowgram (fig. 1).

There are three 454 generations: GS20 (2005): ca. 100 base pairs (bp) per read. GS FLX (2007): ca. 200-300 bp. FLX Titanium (2008): ca. 400-500 bp.

[1] http://www.454.com

Flowgrams

Each 454 flowgram output file (.sff format) includes all reads from one run/plate. It comprises:

Flow values (light signal intensities), quality-clipping positions, quality values for all flow values, and finally the nucleotide sequences.

Around 500 base pairs per read, 1 million reads per plate, 2 GB per output file.

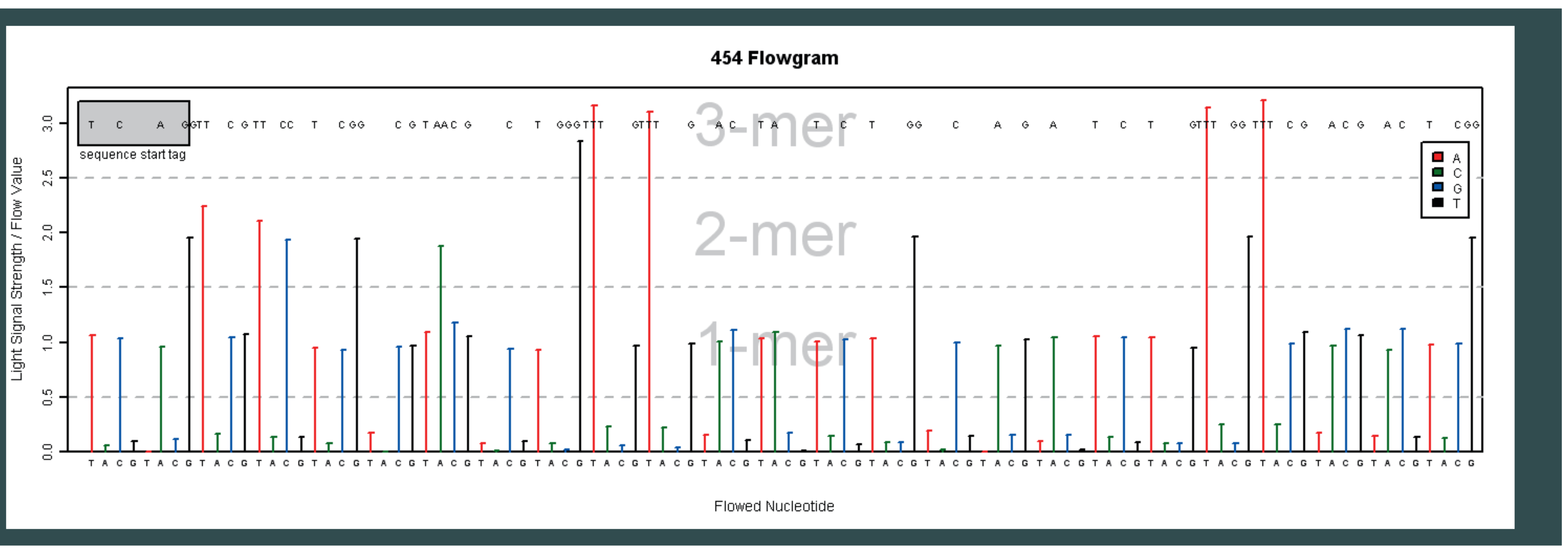

Figure 1: Flowgram. Light signal intensities (flow values) are visualized as vertical bars. The base sequence is read from left to right. Each flowgram starts with the TCAG tag which does not form part of the actual sequence.

Results

Data basis: 1.25 plates of the escherichia coli K-12 strain MG1655, FLX Titanium generation.

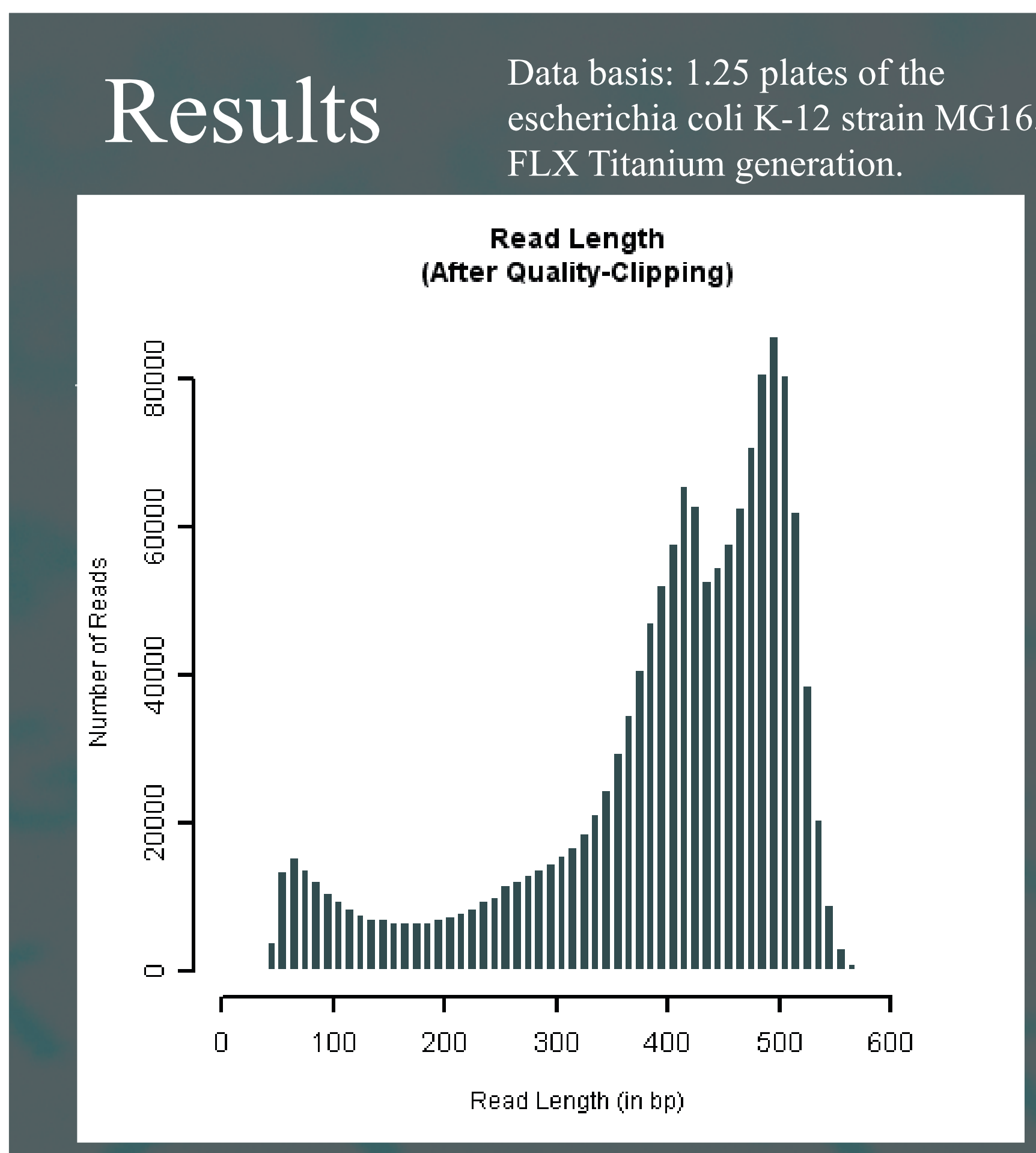

Figure 2: Distribution of read lengths after quality-clipping. Typical Titanium read lengths follow a three-modal distribution. Long reads containing around 500 base pairs are the heart piece of 454 pyrosequencing. They make de-novo whole-genome sequencing possible, which shows that pyrosequencing has become an important alternative to traditional Sanger sequencing.

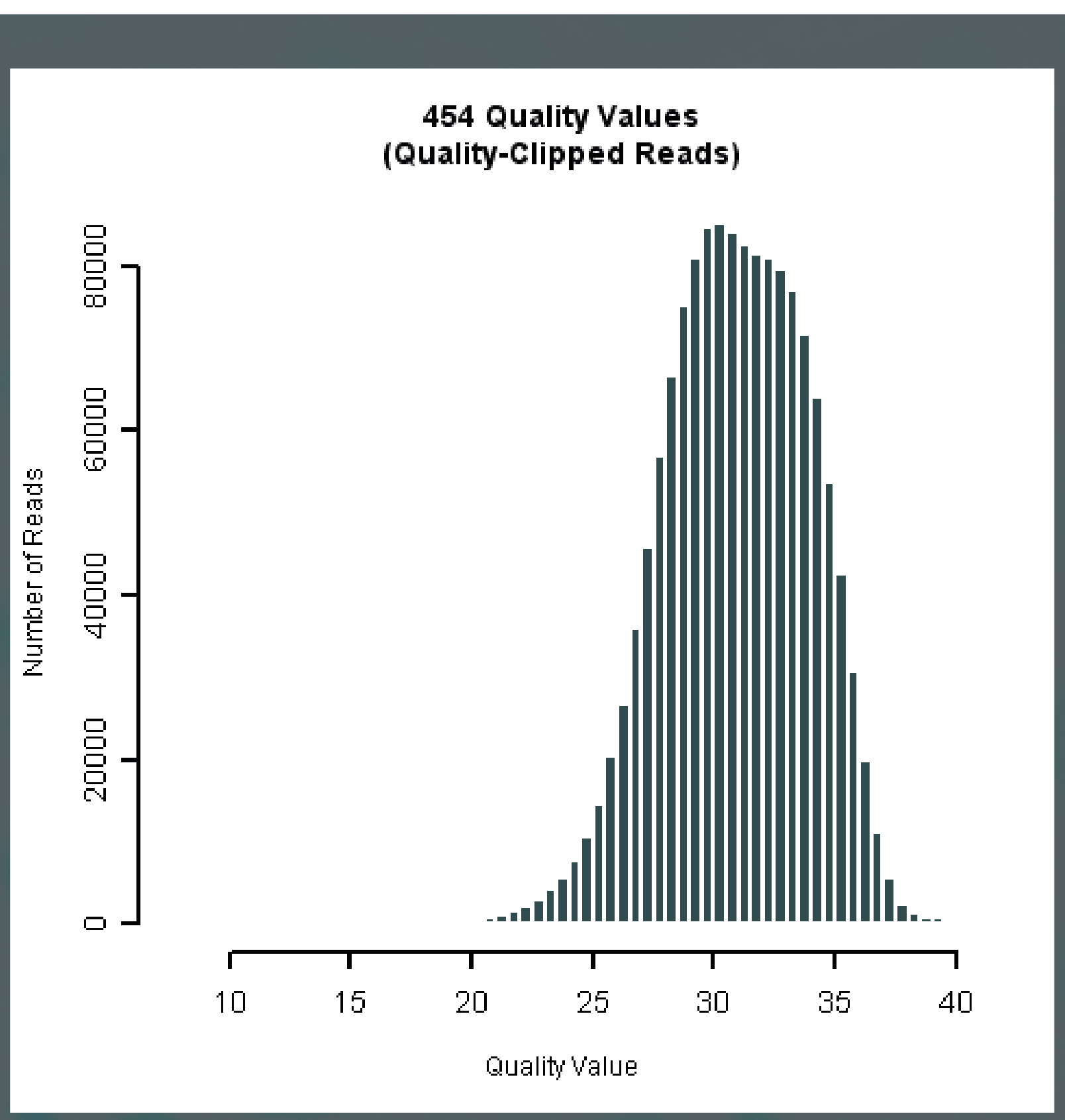

Figure 3: Distribution of 454 quality values (avg. per read). The algorithm of 454 quality value calculation incorporates the following factors: Base position in the sequence, flow value, flow value one cycle before/after, local variation of the flowgram signals in a window surrounding the flow. Within a homopolymer run, quality values are often identical. In earlier 454 releases (GS20 and GS FLX), values used to decrease towards the end of a homopolymer run. A high value indicates good quality.

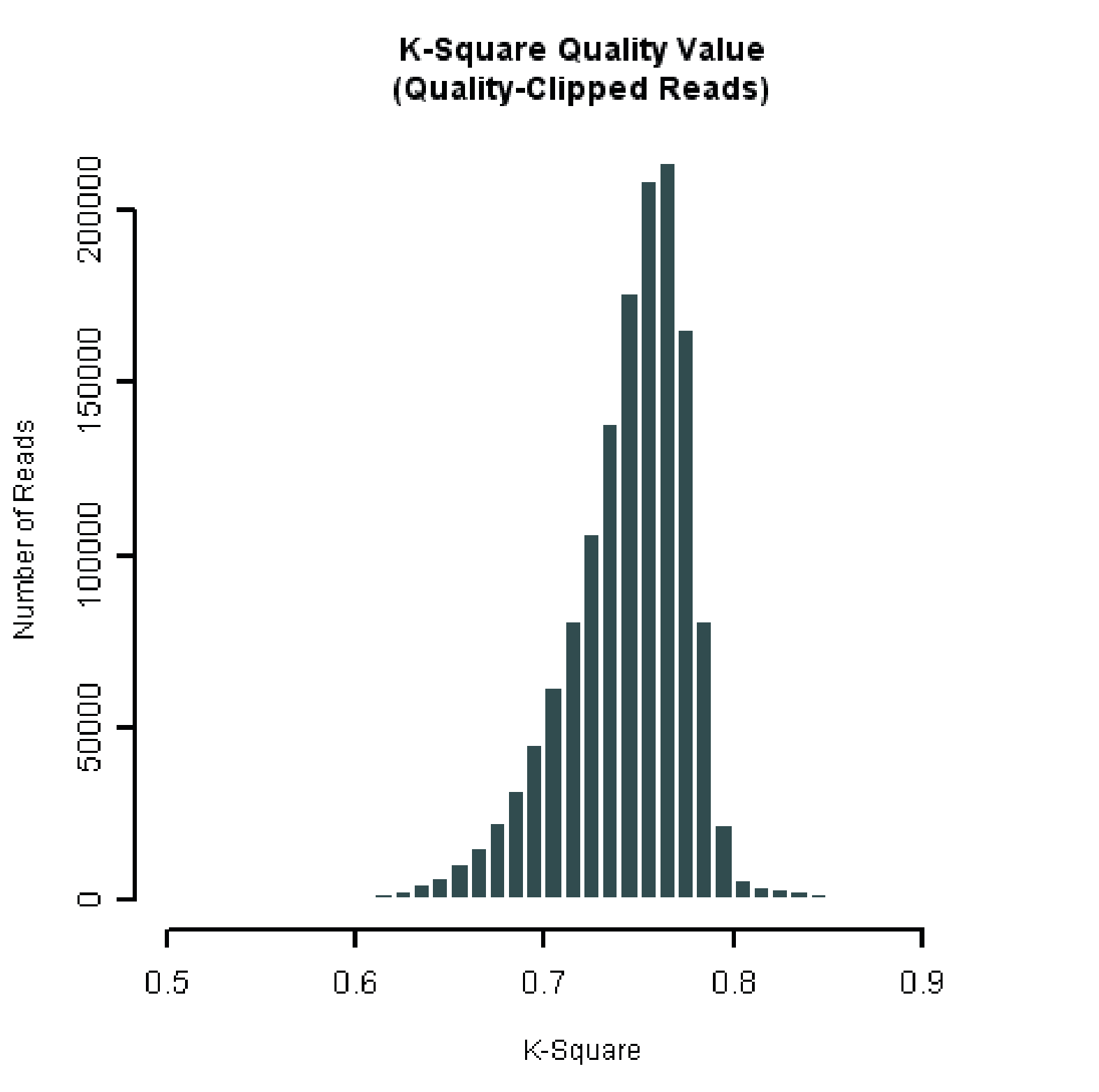

Figure 4: Distribution of K-square quality values (per read). K-square (see formula below) measures the read quality as a degree of deviation from the theoretically optimal flow value. As homopolymer lengths are derived from flow values rounded to integers, flow values close to integers indicate high quality. While 454 quality values are not computed for noise flow values (<0.5), K-square includes them. A flow value of 0.49 is more likely to indicate an existing base than a value of 0.02. A high K-square value indicates good quality.

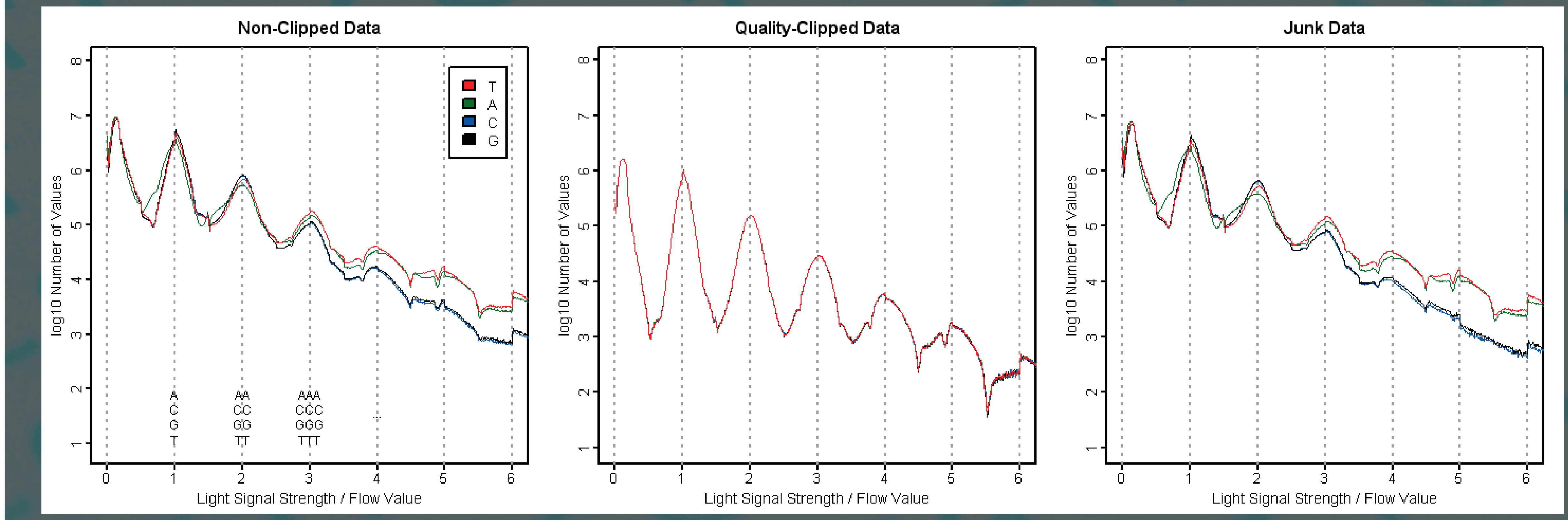

Figure 6: Distribution of flow values (number of occurences per flow value) before and after quality clipping. Signaling noise is partly removed by 454 quality-clipping. Differences between the nucleotides vanish, and valleys between the peaks become deeper, i.e. the probability of an over- or under-call decreases. Nevertheless, homopolymer runs of length four and more cannot be identified in an equally reliable way as shorter lengths: The distributions get broader, i.e. the variance increases. Flow values have been reported to originate from mixed Gaussian or Student's t distributions, but the fit is rather poor for either model.

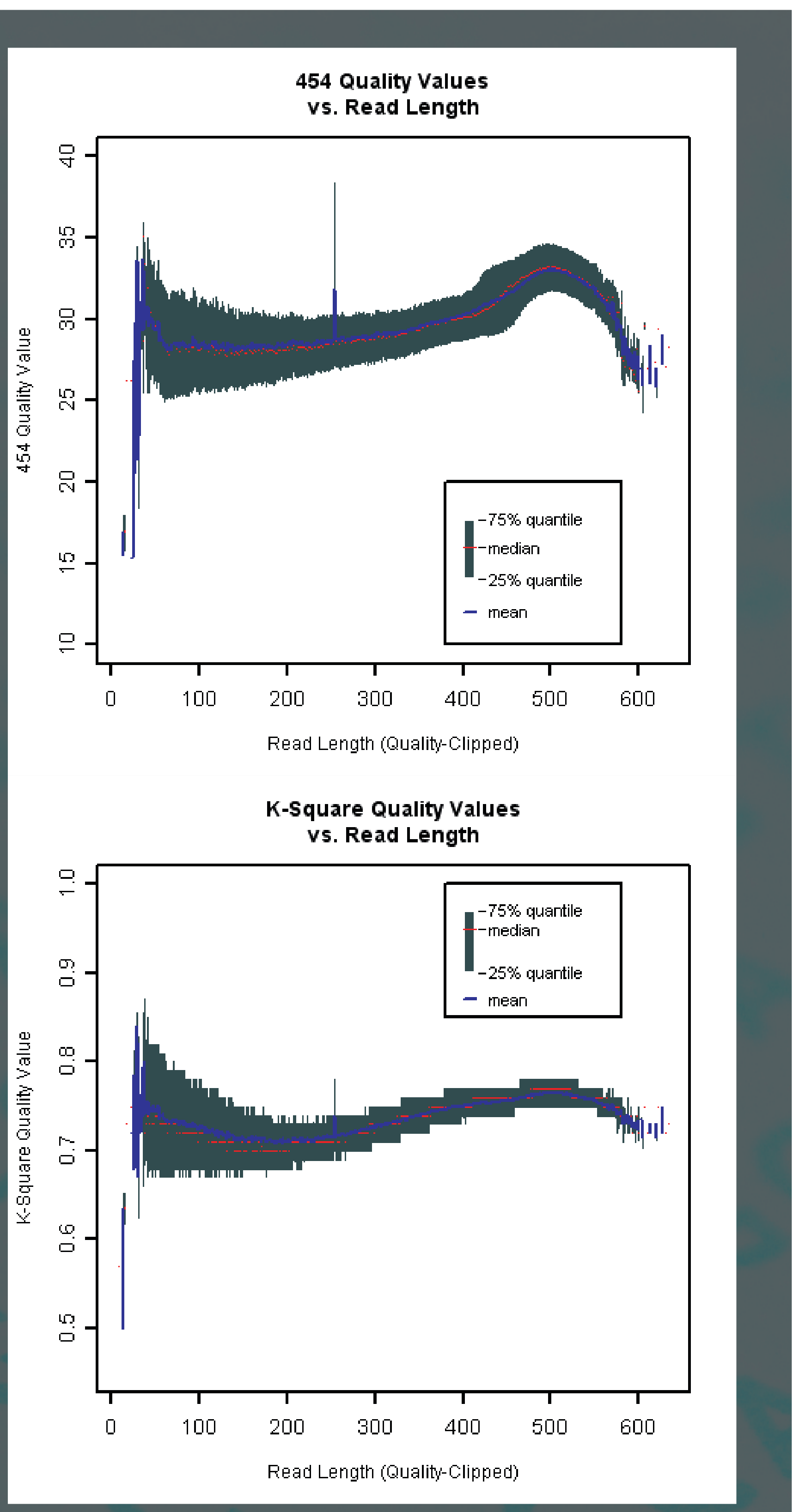

Figure 5: Quality related to read length.
Top: 454 quality values. Bottom: K-square quality values.
Long reads with high quality are the essence of a reliable base-calling. Especially for de-novo whole-genome sequencing, these reads are an excellent starting point.
K-square is insofar a good quality measure as it is both intuitive and simple, and it accounts for the fact that fuzzy base calls from the distribution valleys around x.5 are most likely to be over- or under-calls.

$$K^2 = 1 - 2 \cdot \sqrt{\frac{\sum_{i=1}^{n} (flow_i - round(flow_i))^2}{n}}$$

$flow_i$ flow value in flow position i
n maximum flow position (800 for Titanium)

New Software Tools

http://codgenome.no/software/

flower...

...is a fast and convenient data extraction tool for .sff flowgram files. Like sffinfo (a tool that comes with the 454 sequencing package), *flower* includes the output of raw data, fasta-formatted sequences, or quality values. In addition, *flower* extracts read-specific characteristics, like the XY position on the plate, quality-clipping information, and read length (either from original or quality-clipped data). Furthermore, *flower* computes aggregated values for both 454 and K-square quality (for either original or quality-clipped reads). *flower* supports both terminal and file output.

flowselect...

...allows a smart read selection after one or several factors. For a cumulative whole-genome assembly, one might want to choose reads which, for example, are no shorter than 400 base pairs and have an average K-square quality greater than 0.75. *flowselect* reduces large amounts of data to an arbitrarily smaller amount of high-quality data.

flowsim...

...is a simulator for flowgram data. It creates realistic .sff flowgram files from fasta-formatted nucleotide sequences. Random flow values are drawn from an approximation to the flow distributions (fig. 6). Currently, we use Gaussian distributions for simplicity and performance reasons, but we will integrate further findings into the algorithm and, if required, choose a different distribution model.
As we could show that the variance of flow values increases towards the end of a read, we modeled a degradation factor into the distribution variance.

Contact

Susanne.Balzer@imr.no
Ketil.Malde@imr.no
Inge.Jonassen@ii.uib.no

Bergen Center for Computational Science
COMPUTATIONAL BIOLOGY UNIT

INSTITUTE OF MARINE RESEARCH
HAVFORSKNINGSINSTITUTTET

A project funded by FUGE / The Research Council of Norway.