BIOINFORMATICS
Vol. 26 ECCB 2010, pages i420–i425
doi:10.1093/bioinformatics/btq365

Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim

Susanne Balzer[1,2,*], Ketil Malde[1,*], Anders Lanzén[3,4], Animesh Sharma[2] and Inge Jonassen[2,3]

[1]Institute of Marine Research, PO Box 1870, N-5817, [2]Department of Informatics, University of Bergen, PO Box 7803, N-5020, [3]Computational Biology Unit, Bergen Center for Computational Science, Thormøhlensgate 55, N-5008 and [4]Department of Biology, University of Bergen, PO Box 7803, N-5020, Bergen

ABSTRACT

Motivation: The commercial launch of 454 pyrosequencing in 2005 was a milestone in genome sequencing in terms of performance and cost. Throughout the three available releases, average read lengths have increased to ∼500 base pairs and are thus approaching read lengths obtained from traditional Sanger sequencing. Study design of sequencing projects would benefit from being able to simulate experiments.

Results: We explore 454 raw data to investigate its characteristics and derive empirical distributions for the flow values generated by pyrosequencing. Based on our findings, we implement Flowsim, a simulator that generates realistic pyrosequencing data files of arbitrary size from a given set of input DNA sequences. We finally use our simulator to examine the impact of sequence lengths on the results of concrete whole-genome assemblies, and we suggest its use in planning of sequencing projects, benchmarking of assembly methods and other fields.

Availability: Flowsim is freely available under the General Public License from http://blog.malde.org/index.php/flowsim/

Contact: susanne.balzer@imr.no; ketil.malde@imr.no

1 INTRODUCTION

During the last few years novel sequencing technologies have been introduced. The platforms that are currently commercially available are marketed by Roche (454), Illumina (Solexa/Genome Analyzer), and Applied Biosystems (SOLiD), and they give new challenges for bioinformatics due to data volumes, short read lengths, and difference in errors and quality compared to traditional Sanger sequencing. So far, most bioinformatics methods available have been developed for Sanger sequencing data.

In this article, we characterize the data produced by the 454 system and in particular by its latest version named GS FLX Titanium (referred to as Titanium in the rest of the article). We analyze Titanium data sets from genomes for which the sequence has been determined. Specifically, we map each Titanium read to the reference and derive empirical distributions for the flowgram data obtained (see below; Table 1). This provides an improved basis for analysis and algorithm design, e.g. for base calling and alignment. In this article, we present a simulator that generates realistic flowgram data for any chosen DNA sequence.

The article is structured as follows: in the rest of Section 1, we briefly summarize pyrosequencing, specialized methods for analyzing pyrosequencing data (operating in 'flowspace', see Section 1.2), and simulations. Section 2 follows the results obtained from characterizing pyrosequencing data at the flow level, and in Section 3, we present the Flowsim simulator and some results obtained from comparing simulated and real data sets. Finally, in Section 4 a discussion is given.

1.1 Pyrosequencing

The 454 pyrosequencing technology is based on sequencing-by-synthesis and consists in the cyclic flowing of nucleotide reagents (repeatedly flowing T, A, C, G) over a PicoTiterPlate™. The plate consists of approximately one million wells, and each well contains at most one bead carrying a copy of a unique single-stranded DNA fragment to be sequenced. When the flowed nucleotide is complementary to the template strand in a well, the existing DNA strand in this well is extended with additional nucleotide(s) by a polymerase. This hybridization results in a reaction that generates an observable light signal which is recorded by a camera. The light intensity is converted into a 'flow value', a two-decimal non-negative number that is proportional to the length of a homopolymer run, i.e. it designates the number of nucleotides included in the flow, estimated by simply rounding the number to the closest integer (Margulies *et al*., 2005).

The term 'noise flow values' (in literature sometimes referred to as 'negative flow values', in practical terms being between 0 and 0.49) means that the light signal—although existing—is weak and judged not to result from a chemical reaction. A 'positive flow value' thus indicates incorporation of at least one base, and the number of bases (the homopolymer length) is determined from the flow value. Flow values for one bead (one read) can be used to plot a flowgram (Fig. 1a) from which the associated sequence can be determined.

The cyclically flowed nucleotides and the corresponding flow values build the basis for not only base calling, but also per-base quality score calculation (integrated in Titanium output). Obviously, the key to a correct base calling lies in the accuracy of the light signals. The 454 methodology differs from traditional Sanger sequencing in that substitution errors are a lot less frequent than insertions or deletions. Data properties have slightly changed over the three 454 generations (Roche Applied Science, 2008). We focus on the Titanium technology for all further calculations.

1.2 Use of flow values in data analysis

Although 454 sequences can be analyzed as Fasta files with standard bioinformatics tools, the flow values contain information that is not available in the pure nucleotide sequence. Consequently, several

*To whom correspondence should be addressed.

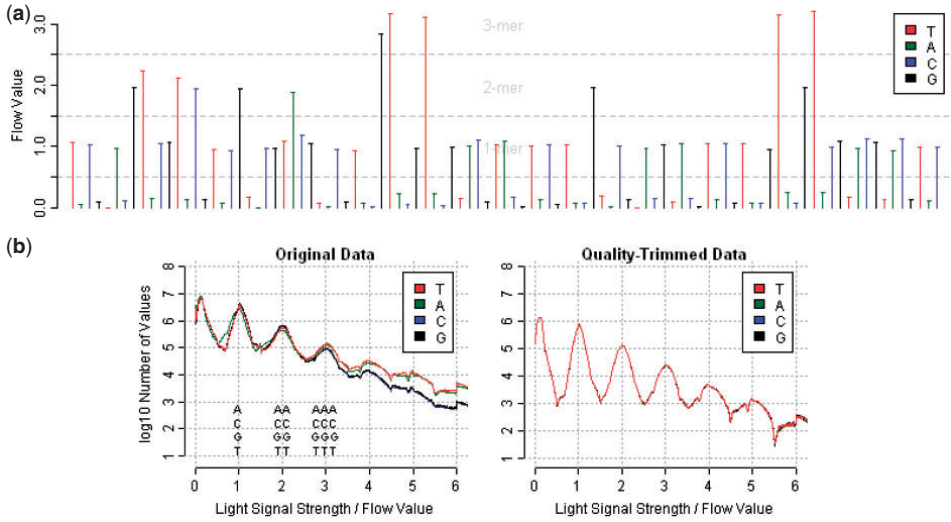

Fig. 1. (**a**) A 454 flowgram: cyclic flowing during one read. The light signal strengths (flow values) are directly translated into homopolymer runs. (**b**) Absolute frequencies of flow values (*E.coli*). Left: original data, no quality-trimming; right: quality-trimmed. The trimming algorithm enhances the separation of the homopolymer length distributions and levels out discrepancies between the nucleotides such that the curves for the four nucleotides are nearly identical.

groups have proposed algorithms to utilize flow values directly. This approach is referred to as operating in 'flowspace' as opposed to 'nucleotide space' and inhibits information loss. For example, the PyroNoise method (Quince *et al*., 2009) uses a maximum likelihood approach to decide whether a set of flowgrams is likely to result from one or several distinct underlying biological sequences. In an analogous manner, using Bayesian statistics, the PyroBayes method (Quinlan *et al*., 2008) determines the length of each homopolymer run as the most likely number of bases given the observed flow value. If the probability for an extra base exceeds a certain threshold, the extra base is added to the homopolymer run. This increases the number of insertion errors, but decreases the number of deletions and substitutions since it is intrinsic to 454 pyrosequencing that substitution errors can only arise from coherent over- and undercalls. This tendency to call more bases in homopolymer runs thus enables a higher SNP identification rate.

For small RNA discovery, direct mapping of flowgrams against a target genome ('FLAT', flowgram alignment tool) has been proved to be an efficient method (Vacic *et al*., 2008). It is also possible to achieve higher per-base accuracy rates in sequence assembly by building consensus sequences in flowspace from highly oversampled data (Huse *et al*., 2007; Margulies *et al*., 2005). Metagenomics is another field where the quality of 454-pyrosequenced data has received much attention (Gomez-Alvarez *et al*., 2009; Huson *et al*., 2007; Quince *et al*., 2009).

Studies have shown that there are several artifacts that heavily influence the processing of data for different purposes (Gomez-Alvarez *et al*., 2009; Huse *et al*., 2007), and especially methods that do not directly use flow values are sensitive to the characteristics of pyrosequencing data. For example, when matching 454 sequences with an indexing approach one can collapse all homopolymer subsequences to length one since pyrosequencing is likely to introduce errors in homopolymer lengths (Miller *et al*., 2008).

Especially for long homopolymers, many errors are caused by broad and overlapping signal distributions leading to ambiguous base calls, although there has also been work on improving 454 sequencing from the chemical aspect (Margulies *et al*., 2005). In addition to the correct determination of homopolymer lengths, the under- or over-calling of bases is especially critical for weak light signals (i.e. noise flow values). A flow value of 0.49 is treated as noise by the 454 base caller although it is almost as likely to originate from a single base call.

1.3 Simulating shotgun data

With Genfrag (Engle and Burks, 1994) and celsim (Myers, 1999), there have been earlier attempts to simulate shotgun read data, but, to the best of our knowledge, MetaSIM (Richter *et al*., 2008) is the only simulator that allows for generating 454 pyrosequencing data. MetaSIM targets Metagenomics. Internally, it uses parametric models for simulating flow values, but its output is Fasta files, and thus it is of limited use for applications that operate in flowspace.

2 FLOW VALUE DISTRIBUTIONS

One of the main challenges in 454 pyrosequencing is the correct determination of homopolymer lengths from flow values. The latter originate from a mixture of overlapping distributions. This is illustrated in Figures 1b and 3, where each distribution is assigned to one homopolymer length and one distribution to noise values. Incorrect homopolymer lengths lead to insertions and deletions during base calling (relative to the underlying biological sequence),

Table 1. Data basis for building the empirical distributions

SFF files	*Escherichia coli*	*Dicentrarchus labrax*	Total
Number of reads[a]	1 176 344	1 270 325	2 446 669
Average read length[a]	534.1	532.8	533.4
Number of bases[a]	92 924 311	85 822 587	178 746 898
Number of flow values[a]	142 361 278	130 621 280	272 982 558
Reference Genome	***Escherichia coli***	***Dicentrarchus labrax***	**Total**
Number of bases[b]	4 639 675	13 213 695	–
Empirical distributions	***Escherichia coli***	***Dicentrarchus labrax***	**Total**
Number of flow values in noise distributions	280 763 949	285 227 582	565 991 531
Number of flow values in homopolymer distributions[c]	314 495 947	278 127 101	592 623 048

[a]After 454 quality-trimming; [b]without *N*'s; [c]homopolymer lengths 1–5, equals to number of homopolymer runs in BLAST results.

and, when an over-call follows an under-call or vice versa, to a perceived substitution error. Therefore, if the distributions did not overlap, this would mean an error-free sequencing. An improved understanding of these distributions also improves the basis for designing algorithms that target the analysis of 454 pyrosequencing data.

2.1 Parametric versus empirical approaches

In earlier studies one has approximated flow values by normal, log-normal (Margulies *et al*., 2005) or non-central student's *t* distributions (Quinlan *et al*., 2008). However, for our data the fit of these distributions is not satisfying (Fig. 3). An alternative is to use non-parametric empirical distributions estimated from real Titanium data for which reference sequences are available. By mapping 454 data to the originating genome, we characterize the distributions of flow values coming from each homopolymer length.

2.2 Sequence comparisons

After having compared Titanium raw data from two different species, *Escherichia coli* and seabass (*Dicentrarchus labrax*, referred to as *E.coli* and *D.labrax,* respectively in the rest of the article), we decided to combine them—equally weighted—into one empirical distribution per homopolymer length. However, we also decided to include the four different nucleotide types in the same distributions since they appear to give rise to very similar distributions. In order to find the distribution of flow values that arises from one particular homopolymer length, we mapped Titanium flowgrams to a reference genome for the same organism, based on one Titanium plate each for an *E.coli* K-12 strain (Blattner *et al*., 1997) and *D.labrax* (Kuhl *et al*., 2010). We used BLAST (Altschul *et al.*, 1990) to identify the location of reads that could be aligned unambiguously to one location on the genome, with default BLAST parameters, except for gap open and extend penalties, which were set to 1.

Table 2. Parameters of the empirical distributions

Homopolymer length	Mean	Standard deviation
0	0.1230	0.0737
1	1.0193	0.1227
2	2.0006	0.1585
3	2.9934	0.2188
4	3.9962	0.3168
5	4.9550	0.3863
Linear regression for $n \geq 6$	n[a]	$0.03494 + n \cdot 0.06856$[a]

[a]Normal distribution. Mean and standard deviation of normal distribution around homopolymer lengths of 6, 7 etc.

To distinguish sequencing errors from true biological variation, we used a bit score threshold of 200 and only the best match for each sequence. Furthermore, we discarded all those matches that had a corresponding second best match with a bit score <5% worse than the best match, i.e. two matches with bit scores that were approximately equally high.

For *E.coli*, there were uncertainties in terms of which reference genome to choose, as none of the available reference genomes gave us >97% identity with the pyrosequencing data, but the match filtering mentioned above should account for these problems.

2.3 Calculation of empirical distributions

We aligned the flowgrams to the matching genomic region, assigning each flow value to the corresponding true homopolymer length as known from the reference genome. Thus, we collected the flow values assigned to each homopolymer length distribution from 0 to 5, as shown in Figure 3.

For homopolymer lengths greater than 5, our data is sparse, and it is therefore better to approximate the real distributions by extrapolating parametric distributions from the shorter homopolymer lengths. Table 2 shows the observed mean and standard deviation of the empirical distributions for homopolymer lengths 0 to 5, and the linear regression for these parameters based on normal distributions fitted to homopolymer lengths 1 to 5.

2.4 Degradation and Noise

We find our resulting empirical distributions to be almost symmetrical around the corresponding integers, with relatively low standard deviation for short homopolymer runs. However, when analyzing data from the three 454 generations, we also found that the degree of symmetry varies between them. Quinlan *et al.* (2008) report a significantly higher insertion than deletion rate, which is consistent with an asymmetry in the tails of the distributions, but we found the asymmetry to decrease towards newer generation data.

Nevertheless, we can clearly observe two kinds of degradation: since standard deviation increases for increasing homopolymer lengths, these belong to broader distributions with overlapping tails, where the latter generally means a higher risk of over- and under-calls.

Second, analysis of the flow values associated with sequence parts that have been trimmed off (during standard 454 quality-trimming) indicates that 454 quality-filtering and -trimming calibrates discrepancies between the four nucleotides and increases the separations of the distributions, involving deeper valleys

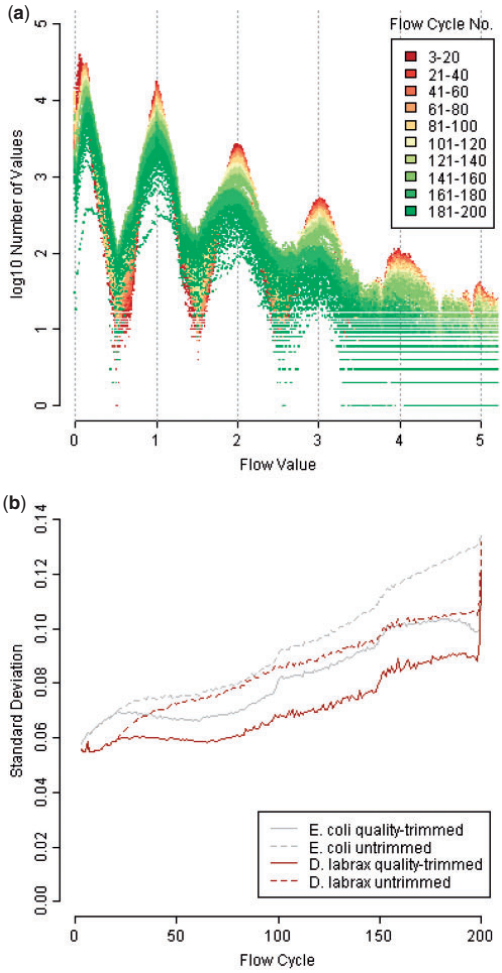

Fig. 2. (**a**) Absolute frequencies of flow values by flow cycle. A total of 200 flow cycles of a Titanium run correspond to $200 \times 4 = 800$ flows. The first two flow cycles contain the TCAG tag and are omitted here. Towards the end of a run, flow values tend to lie further away from their ideal values (integers), but are obviously less in number because many values from later flow cycles have been trimmed away. (**b**) Standard deviation of flow values (difference in relation to their closest integer), by flow cycle. Standard deviation increases almost linearly. Only flow values <5.5 were included.

between them (Figs 1b and 2a). We therefore use only the subsequences retained after quality-trimming to estimate the empirical distributions, thus being able to treat the nucleotides equally. Also for quality-trimmed raw data, we can see that both read and flow position of a base have a remarkable influence on the accuracy of flow values. We have observed a clear degradation in accuracy over the length of a run, i.e. when comparing earlier to later flow cycles, by measuring for each flow cycle how much the difference between a flow value and its ideal counterpart (i.e. the closest integer) varies (Fig. 2b).

2.5 Read lengths

The length of un-trimmed reads in 454 pyrosequencing is limited by either the number of flows (168 in GS20, 400 in GS FLX and 800 in GS FLX Titanium) or the length of the clones. The longest reads are thus obtained when the clone length exceeds the number of flows, such that the DNA strands in the well are extended until the very last flow cycle.

As quality decreases towards the end of a read, several filters are applied on the reads, which again gives a different read length distribution. We can thus distinguish between the distribution of clone lengths, the distribution of read lengths before filtering and quality-trimming and that after application of those filters. A detailed description of the filtering algorithms is given in the 454 manual (Roche Applied Science, 2008). As visible in Figure 1b, they eliminate (some of the) artifacts in the distributions by trimming low-quality flow values from the end of each read.

3 FLOWSIM—A SIMULATOR FOR 454 DATA

To take advantage of the empirical distributions, we implemented Flowsim, a simulator for pyrosequencing data.

3.1 Implementation of Flowsim

Given an input sequence in Fasta format, Flowsim selects substrings of this sequence with random position and strand, and generates a flowgram by converting the nucleotide sequences to sequences of homopolymer lengths. Each homopolymer length is then altered according to its flow distribution, where the latter is allowed to vary (degrade) with the flow position in the simulated read. To emulate degradation, we derived 20 different sets of empirical distributions from our mapping results (Fig. 3), where each of them represents 10 consecutive Titanium flow cycles, which sums up to 800 flow values.

The simulated flowgram is then analyzed to call nucleotide sequence and quality scores. Finally, all generated information is stored in an SFF file, similar to the ones produced by the 454 software.

One can further specify the number of desired output reads and also incorporate user-defined empirical distributions, either position-specific (degrading) or not.

3.2 Quality scores

It is crucial to assign a quality score to each called base, since sequenced bases are not filtered individually during quality-filtering and -trimming, but rather in the context of their reads. Quality scores are e.g. useful for assembly projects, although some assemblers do not use them. If they do, however, they might rely on them for incorporating Sanger reads since 454 quality scores are expressed as a phred equivalent (Margulies *et al.*, 2005; Roche Applied Science, 2008). On the other hand, scores can also be used by assemblers built for Sanger sequences when assembling 454 sequences.

Although the method for determining quality has been described both for GS20 (Margulies *et al.*, 2005) and Titanium (Brockman *et al.*, 2008), the exact parameters are not known. Instead, Flowsim

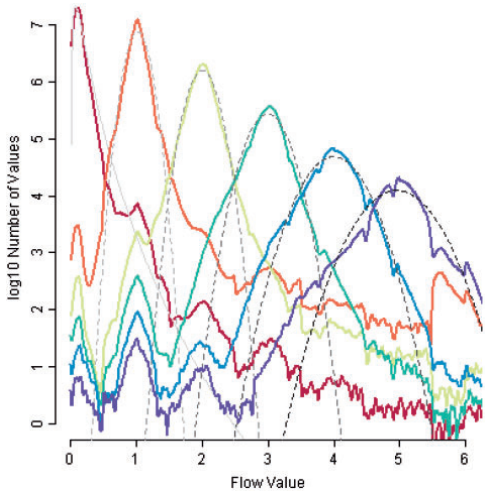

Fig. 3. Empirical distributions (smoothed average of *E.coli* and *D.labrax*) on logarithmic scale. In gray: fitted (log-) normal distributions.

calculates the error probability ('the base in question is an over-call'), using Bayes' Theorem, and transfers it into a phred equivalent. Thus, the quality score corresponds to the true quality of the simulated base call, rather than to the quality the 454 software would produce for the same flowgram.

Flowsim currently supports two quality calling methods based on Bayesian statistics. One produces decreasing quality scores for the bases in a homopolymer, similar to GS20. The second produces a series of identical values for each base in a homopolymer, as in Titanium, but otherwise builds on the same Bayesian approach as the GS20 algorithm. Compared to the quality scores assigned to Titanium by the Roche analysis pipeline, our quality scores are lower. As GS20 appears to use a fixed table mapping each flow value to a set of qualities, there is also a third option of assigning qualities from a table derived from GS20 data.

Bayes' theorem requires both the prior probability for each homopolymer length and the conditional probability for a flow value given a certain homopolymer length. In contrast to Margulies *et al*., we use both empirical priors (from the input Fasta file) and empirical conditional probabilities (from our empirical distributions). This allows us to assess the quality of our simulated data as accurately as possible. When position-specific empirical distributions are used in Flowsim, we also use these for quality score calculation.

3.3 Simulating data sets

We used Flowsim to generate synthetic data sets, using our empirical distributions as the flow model. Each of the 20 distributions was used for 10 flow cycles (40 flows), giving a realistic degradation of quality along the sequence. We also simulated data sets using 400 flow cycles, simulating a hypothetical 454 generation with twice the read length of the current Titanium generation. The *E.coli* genome (K-12 strain, GenBank ID: 49175990) was used as the input genome.

Table 3. *De novo*-based and reference-based N50 for *E. coli*

Coverage	Real	200 cycles (simulated)	400 cycles (simulated)
De novo-based N50 for *E.coli*			
1	649	651	995
5	2406	7045	7623
10	23 613	132 913	104 012
15	67 231	173 592	178 129
20	86 902	172 127	203 060
25	95 348	176 747	207 011
30	97 821	171 819	207 011
Reference-based N50 for *E. coli*			
1	895	1093	1681
5	8305	31 730	40 321
10	76 687	207 827	2 343 849
15	110 013	207 856	2 496 857
20	118 387	207 740	2 497 013
25	161 266	207 899	2 497 058
30	177 489	207 845	2 724 990

3.4 Simulation results

We have performed both de-novo and reference-based assembly using Newbler assembler version 2.3 (Roche), approximating various coverage (1×, 5×, 10×, 15×, 20×, 25× and 30×). A simulation with 200 flow cycles shows ∼1% inferred error, while 400 flow cycles result in an error rate of ∼0.8%, which is the same as for the real data (Titanium, i.e. 200 flow cycles).

Our results indicate that Flowsim can be useful to estimate the quality of an assembly that can be expected from using Titanium to shotgun sequence a genome. However, the assemblies resulting from our simulations were consistently better in terms of contig sizes (through the N50 summarizing statistic, see Table 3) for the simulated data sets than for the real ones. This may partly be due to all simulated reads coming from the reference genome and thus avoiding strain-specific discrepancies, which leads to the fact that 100% of the reads for 200 and 400 flow cycle simulations can be mapped back to genome, while real data reach only ∼98.7% for all studied coverage values. There may also be other factors such as possible biases in terms of genome coverage in the experimental protocols used to generate the shotgun libraries for Titanium sequencing. Further work will include exploring such biases and other sources of variability as well as characterizing their influence on the simulation accuracy of Flowsim. Also Flowsim will be extended to include simulation of paired-reads, which will be of high value for simulation and planning of projects for de-novo whole-genome sequencing.

4 DISCUSSION

This study aims to sketch the opportunities that arise from analyzing pyrosequencing raw data, culminating in the use of empirical distributions. The empirical distributions give us a very realistic picture of the underlying characteristics of the light signal values that are later translated into DNA sequences. In contrast, earlier approaches to modeling flow data have built on parametric

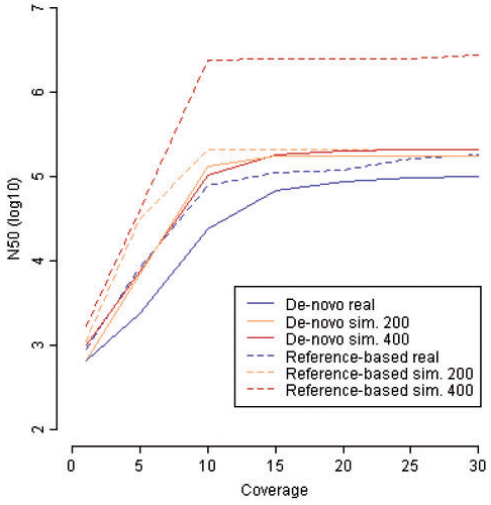

Fig. 4. *De novo* and reference-based N50 for *E.coli*. Both real and simulated 454 data were assembled using Newbler v2.3.

distributions, and the same distributions were used for whole reads, without respect to flow or read positions.

Our findings and the empirical distributions are based on large amounts of data from three different species (*E.coli*, *D.labrax*, *Gadus morhua*), four sequencing labs, both shotgun and paired-end reads with different gap sizes. The empirical flow value distributions are very similar, and we have not observed any factors which influence the shape of the distributions apart from the 454 generation. Thus, we have a good reason to believe that the distributions used in Flowsim are representative.

The flow values that result from 454 sequencing exhibit many interesting characteristics and artifacts, and we do not address them all here. Some of these are generation-specific, some of them have remained stable over the years, and some of them only appear on one certain plate, for one certain species or in one lab. One known artifact, exact or almost-exact duplicates, has been not only described for metagenomics in the literature (Gomez-Alvarez *et al*., 2009), but we also observed them in shotgun sequences for *E.coli* and *D.labrax*.

We do emulate the degradation in empirical flow distributions, and we also calculate the corresponding quality scores. In contrast, we neglect some of the artifacts that we have observed in the empirical distributions, but are not able to interpret properly yet, such as for example: shifts in peaks that lead to systematic over- or under-calls, jumps, neighboring peaks, i.e. subpeaks around the next or preceding integer. These are particularly strong for the noise distribution (with a neighboring peak around 1) and the 1-distribution (with neighboring peaks around 0.1 and 2), but the values causing these peaks are not many in number. Analyzing the corresponding data including the related alignments we found that the subpeaks are likely to be caused by real biological differences. This will be explored further in a separate study. In this context, we also performed a weak smoothing process that helped to reduce subpeaks and jumps.

Furthermore, the 454 image analysis software implements a set of quality filters that sets trimming coordinates to identify the high-quality part of each read. In addition, some reads are eliminated entirely based on quality metrics. Although these filters are documented (Roche Applied Science, 2008), the documentation is not sufficient to re-implement them, and the current version of Flowsim does not attempt to simulate them. We hope to address this in a future release (Fig. 4).

In conclusion, our simulator produces sufficiently realistic 454 files as we model all important phenomena that we have observed. Furthermore, Flowsim allows the user to specify many of its parameters, making it adaptable to new real or hypothetical 454 generations.

ACKNOWLEDGEMENTS

The authors wish to thank Dr Alexander Sczyrba, DOE Joint Genome Institute, Dr Richard Reinhardt, Max Planck Institute for Molecular Genetics, Berlin, for kindly providing us with Titanium raw data and Dr Christopher Quince, University of Glasgow, for the fruitful discussions. Notur is acknowledged for access to the Titan cluster in Oslo.

Funding: The National Program for Research in Functional Genomics in Norway (FUGE) in the Research Council of Norway.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Blattner,F.R. *et al*. (1997) The complete genome sequence of Escherichia coli K-12. *Science*, **277**, 1453–1462.

Brockman,W. *et al*. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.*, **18**, 763–770.

Engle,M.L. and Burks,C. (1994) GenFrag 2.1: new features for more robust fragment assembly benchmarks. *Comput. Appl. Biosci.*, **10**, 567–568.

Gomez-Alvarez,V. *et al*. (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J*., **3**, 1314–1317.

Huse,S.M. *et al*. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, **8**, R143.

Huson,D.H. *et al*. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

Kuhl,H. *et al*. (2010) The European sea bass Dicentrarchus labrax genome puzzle: comparative BAC-mapping and low coverage shotgun sequencing. *BMC Genomics*, **11**, 68.

Margulies,M. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Miller,J.R. *et al*. (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**, 2818–2824.

Myers,G. (1999) A dataset generator for whole genome shotgun sequencing. In *Proceedings of International Conference on Intelligent Systems of Molecular Biology*, pp. 202–210.

Quince,C. *et al*. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods*, **6**, 639–641.

Quinlan,A.R. *et al*. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods*, **5**, 179–181.

Richter,D.C. *et al*. (2008) MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One*, **3**, e3373.

Roche Applied Science. (2008) *Genome Sequencer Data Analysis Software Manual*, Software Version 2.0.00, Roche Diagnostics GmbH.

Vacic,V. *et al*. (2008) A probabilistic method for small RNA flowgram matching. In *Pacific Symposium on Biocomputing*, pp. 75–86.

Characteristics of 454 Pyrosequencing Data – Enabling Realistic Simulation with Flowsim

Susanne Balzer, Ketil Malde, Anders Lanzén, Animesh Sharma and Inge Jonassen

The authors would like to apologize for an error in the calculation of the number of bases, number of flow values and average read length. Our reads turned out to be a lot shorter than previously reported. None of these errors has implications on the method or the results. The corrected table is shown below.

Table 1. Data basis for building the empirical distributions

SFF files	*E. coli*	*D. labrax*	Total
Number of reads*	1,176,344	1,270,325	2,446,669
Average read length*	393.7	424.0	409.4
Number of bases*	463,133,786	538,607,063	1,001,740,849
Number of flow values*	710,777,022	819,636,576	1,530,413,598
Reference Genome	***E. coli***	***D. labrax***	**Total**
Number of bases**	4,639,675	13,213,695	-
Empirical Distributions	***E. coli***	***D. labrax***	**Total**
Number of flow values in noise distributions	280,763,949	285,227,582	565,991,531
Number of flow values in homopolymer distributions***	314,495,947	278,127,101	592,623,048

*after 454 quality-trimming **without N's ***homopolymer lengths 1-5, equals to number of homopolymer runs in BLAST results

The error also affects figure 1b, where the left part of the plot is to be compared with the right set of curves. The corrected figure is shown below.

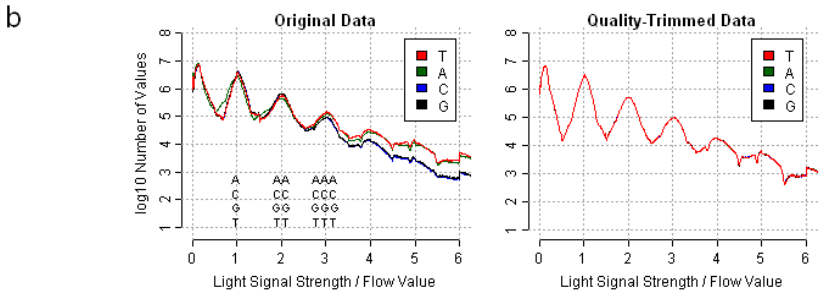

Fig. 1 (b) Absolute frequencies of flow values (*E. coli*). Left: Original data, no quality-trimming; right: quality-trimmed. The trimming algorithm enhances the separation of the homopolymer length distributions and levels out discrepancies between the nucleotides such that the curves for the four nucleotides are nearly identical.