An efficient simulator of 454 data using configurable statistical models

Fredrik Lysholm1*, Björn Andersson2 and Bengt Persson1,2

Abstract

Background: Roche 454 is one of the major 2nd generation sequencing platforms. The particular characteristics of 454 sequence data pose new challenges for bioinformatic analyses, e.g. assembly and alignment search algorithms. Simulation of these data is therefore useful, in order to further assess how bioinformatic applications and algorithms handle 454 data.

Findings: We developed a new application named 454sim for simulation of 454 data at high speed and accuracy. The program is multi-thread capable and is available as C++ source code or pre-compiled binaries. Sequence reads are simulated by 454sim using a set of statistical models for each chemistry. 454sim simulates recorded peak intensities, peak quality deterioration and it calculates quality values. All three generations of the Roche 454 chemistry (‘GS20’, ‘GS FLX’ and ‘Titanium’) are supported and defined in external text files for easy access and tweaking.

Conclusions: We present a new platform independent application named 454sim. 454sim is generally 200 times faster compared to previous programs and it allows for simple adjustments of the statistical models. These improvements make it possible to carry out more complex and rigorous algorithm evaluations in a reasonable time scale.

Introduction

The introduction of 2nd generation sequencing techniques has resulted in a myriad of new large DNA sequencing projects, many of which have posed new challenges for bioinformatics. Roche 454 is one of the major 2nd generation sequencing platforms. The 454 instruments use a type of pyrosequencing chemistry, where the complementary strand is elongated through repeated cycles, each using one of the four nucleotides (flows). The complementary strand is elongated in the absence of a terminator, and homopolymer lengths are estimated by the light intensity recorded. As a consequence, homopolymer length uncertainties sometimes occur, especially at long homopolymer stretches [1]. This specific aspect of 454 data poses new challenges for downstream bioinformatic applications, such as assembly and alignment search algorithms [2-4]. Simulation of sequence data is an important and extensively used tool for assessing how bioinformatic applications and algorithms handle sequence data. The first software that was developed for simulation of 454 data was MetaSIM [5], which provided simulated data using the statistical distributions suggested for 454 data by Margulies et al. [1]. However, MetaSIM did not produce the raw data that many algorithms utilize and did not model the ‘Titanium’ chemistry. As a response to these limitations, Flowsim [6] was created. Flowsim produced 454 raw data and allowed for ‘Titanium’ reads to be produced. Flowsim accepted parameters that modify a particular setting, but all 454 chemistry generations supported by Flowsim were written in source code, which made small modifications complicated. Flowsim and MetaSIM did not produce detailed information of the simulation run, which would for example allow evaluation of correct homopolymer identification in an efficient manner. In order to overcome these limitations, we now introduce 454sim.

Methods and implementation

454sim generates simulated 454 data from input sequences in FASTA format. The algorithm models positive flows (a flow interpreted as one or more bases) with a normal distribution and negative flows (a flow interpreted as no base of that type, i.e. a noise flow)
Table 1 454sim performance evaluation

| Application       | Reads/sec(*) | Wall time     | CPU time       |
|-------------------|--------------|---------------|----------------|
| 454sim 8-threads  | 10,100       | 1 min 39 sec  | 11 min 30 sec  |
| 454sim 1-thread   | 2,730        | 6 min 6 sec   | 5 min 59 sec   |
| Flowsim 0.2.7     | 503          | 5 hrs 31 min 40 sec | 5 hrs 30 min 20 sec |

Evaluation details for the ‘454sim’ and ‘flowsim-0.2.7’ binary in simulation of 1,000,000 ‘Titanium’ reads. The simulation was performed on a workstation equipped with an “Intel Core i7 920” processor, HT enabled, running Linux 64-bit. * Reads/sec is derived from wall time.

Discussion

As the amount of sequence data produced keeps increasing, many downstream bioinformatic programs are already adapted to rapidly process large amounts of reads. In order to establish rigorous methods for the evaluation of these programs, it is important to also be able to simulate reads efficiently. Furthermore, as 454 data quality can vary, for example between metagenomic and genomic sequencing, it is also important to be able to modify the statistical models by which data is simulated. 454sim was constructed to meet these demands. The statistical models describing the Roche 454 platform chemistries are imported from separate text files. These can easily be modified by editing the file using a text editor. Entirely new models can also be added by creating an additional text file. To further facilitate analysis, 454sim also produces optional detailed output where the simulation of each base is described. This enables new types of evaluation such as correct homopolymer indel identification, which is not possible with previous tools. 454sim is at least two orders of magnitude faster than flowsim and reduced the run-time of a ‘Titanium’ chemistry simulation from 5 1/2 hours down to less than 2 minutes, on an Intel Core i7 920. 454sim is platform independent and available as C++ source code [see Additional file 1] or the project homepage.

Availability and requirements

Project name: 454sim

Project home page: https://sourceforge.net/p/bioinfo-454sim/

Operating system(s): Platform independent

Programming language: C++

Other requirements: –

License: GNU General Public License

Any restrictions to use by non-academics: –

Additional material

Additional file 1: Binaries and source code

The file contains Linux 32 and 64-bit binaries, a Windows 32-bit binary as well as source code. See the enclosed README for more information.

Acknowledgements and funding

We gratefully acknowledge financial support from the Swedish Research Council, the Research School of Medical Bioinformatics supported by the...
Knowledge Foundation Sweden, Karolinska Institutet and Linköping University.

Author details
1IFM Bioinformatics and SeRC (Swedish e-Science Research Centre), Linköping University, S-581 83 Linköping, Sweden. 2Department of Cell and Molecular Biology, Science for Life Laboratory, Karolinska Institutet, S-171 77 Stockholm, Sweden.

Authors’ contributions
FL has implemented the software and written the manuscript. BP and BA have helped design the study and draft the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 13 July 2011 Accepted: 26 October 2011

References
1. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Izyk GP, Jando SC, Alenquer ML, Jarvie TP, Jarman KB, Kim JR, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, et al: Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005, 437:376-380.
2. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 2007, 8: R143.
3. Gomez-Alvarez V, Teal TK, Schmidt TM: Systematic artifacts in metagenomes from complex microbial communities. ISME J 2009, 3:1314-1317.
4. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT: Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 2009, 6:639-641.
5. Richter DC, Ott F, Auch AF, Schmidt R, Huson DH: MetaSim: a sequencing simulator for genomics and metagenomics. PloS One 2008, 3:e3373.
6. Balzer S, Malde K, Lanzén A, Sharma A, Jonassen I: Characteristics of 454 pyrosequencing data - enabling realistic simulation with flowsim. Bioinformatics 2010, 26:420-425.
7. Marsaglia G, Tsang WW: The Ziggurat Method for Generating Random Variables. Journal of Statistical Software 2000, 5:1-7.
8. Matsumoto M, Nishimura T: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans Model Comput Simulat 1998, 8:3-30.

doi:10.1186/1756-0500-4-449
Cite this article as: Lysholm et al.: An efficient simulator of 454 data using configurable statistical models. BMC Research Notes 2011 4:449.