BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 1 2012, pages 132–133
doi:10.1093/bioinformatics/btr627

Sequence analysis

Advance Access publication November 15, 2011

QuRe: software for viral quasispecies reconstruction from next-generation sequencing data

Mattia C. F. Prosperi* and Marco Salemi

Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, Emerging Pathogens Institute, University of Florida, Gainesville, FL 32610-3633, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: Next-generation sequencing (NGS) is an ideal framework for the characterization of highly variable pathogens, with a deep resolution able to capture minority variants. However, the reconstruction of all variants of a viral population infecting a host is a challenging task for genome regions larger than the average NGS read length. QuRe is a program for viral quasispecies reconstruction, specifically developed to analyze long read (>100 bp) NGS data. The software performs alignments of sequence fragments against a reference genome, finds an optimal division of the genome into sliding windows based on coverage and diversity and attempts to reconstruct all the individual sequences of the viral quasispecies—along with their prevalence—using a heuristic algorithm, which matches multinomial distributions of distinct viral variants overlapping across the genome division. QuRe comes with a built-in Poisson error correction method and a post-reconstruction probabilistic clustering, both parameterized on given error rates in homopolymeric and non-homopolymeric regions.

Availability: QuRe is platform-independent, multi-threaded software implemented in Java. It is distributed under the GNU General Public License, available at https://sourceforge.net/projects/qure/.

Contact: ahnven@yahoo.it; ahnven@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 15, 2011; revised on October 19, 2011; accepted on November 8, 2011

1 INTRODUCTION

Next-generation sequencing (NGS) techniques produce thousands to billions of sequence fragments (called *reads*), composed of tens to hundreds of nucleotide bases. The current applications of NGS include *de novo* sequencing, re-sequencing and metagenomics (Pareek *et al.*, 2011). Several technologies have been commercialized (Metzker, 2010) and while the purchase of a machine often is not possible by a clinical or research laboratory, NGS services are available at a reasonable price from other facilities for genetic studies or diagnostic purposes (Kingsmore and Saunders, 2011).

NGS assembly software has been designed to infer a unique genome from a read set (Miller *et al.*, 2010). One of the challenges of re-sequencing and *de novo* sequencing from NGS data is the reconstruction of a viral population, or viral swarm, also referred to as a quasispecies (Holmes, 2010; Más *et al.*, 2010). The current approach for quasispecies characterization uses clonal Sanger sequencing. However, NGS has the advantage to detect minority variants at a higher resolution than Sanger, at a competitive price.

Several theoretical approaches for the viral swarm reconstruction have been published, validated both via simulations and real NGS data (Eriksson *et al.*, 2008; Jojic *et al.*, 2008; Prabhakaran *et al.* 2010; Prosperi *et al.*, 2011; Westbrooks *et al.*, 2008; Zagordi *et al.*, 2010a; b). Three software have been implemented and released, ShoRAH (Zagordi *et al.*, 2011), PredictHaplo (http://www.cs.unibas.ch/personen/roth_volker/HivHaploTyper/) and ViSpA (Astrovskaya *et al.*, 2011), all for unix/linux platforms. An extensive review of methods has been recently published by Beerenwinkel and Zagordi (2011). We introduce here QuRe (Quasispecies Reconstruction algorithm), a platform-independent, multi-threaded Java implementation of the method by Prosperi *et al.* (2011). QuRe reconstructs automatically a set of error-free, full-gene or full-genome, variants from a collection of NGS reads. It is based on a heuristic algorithm that maps reads to a reference genome, and then matches distinct viral variants across a partition of the reference into overlapping intervals, using multinomial distributions. However, the performance of reconstruction algorithms may depend on many factors, including quasispecies diversity, rate of heterogeneity, read length and coverage; it has been shown that QuRe has advantages over other methods in reducing the chance to reconstruct false *in silico* recombinants (Prosperi *et al.*, 2011). The software has been designed for NGS technologies that produce long reads (>100 bp), such as Roche 454 Life Sciences (http://www.my454.com/).

2 METHODS AND IMPLEMENTATION

Both the reads generated by the NGS machinery and the reference genome are pre-processed to obtain all the substrings of length k (k-mers), stored in hashtables. Each k-mer set of a read (forward and reverse strand) is compared with the k-mer set of the reference in order to obtain an estimate of the mapping position of the read within the genome, as previously proposed by Archer *et al.* (2010). After a mapping position is found for a read, this position is elongated by the average read length plus 3 SDs. The reference is cut in this interval, and then aligned with the read using the Smith–Waterman–Gotoh local alignment algorithm (Gotoh, 1982). A distribution of quasi-random r alignment scores generated by aligning reads, whose characters have been shuffled, with random genome cuts is also calculated (r is set to 2000). A preliminary filter discards reads that do not exhibit an alignment score sufficiently high as compared to the random score distribution, using a z-test (Bacro and Comet, 2000). The significance is set to $P < 0.01$, corrected with

*To whom correspondence should be addressed.

the Benjamini–Hochberg procedure. Once all reads have been aligned to the reference, coverage and nucleotide content of each genome position are calculated, along with entropy. Positions that do not have coverage of at least c reads (30 by default) or below the fifth percentile of the overall coverage distributions are eliminated. A unique mapping interval is then inferred by extracting the largest set of consecutive positions with the specified coverage, and retained for the error correction procedure.

The error correction model assumes a Poisson distribution of errors parameterized differently in homopolymeric and non-homopolymeric regions, following Wang *et al.* (2007). Nucleotide variations or insertions/deletions are then filtered according to their P-value (<0.01, using the Bonferroni correction), and a consensus sequence is generated. The corrected read set is input to the reconstruction algorithm.

The reconstruction algorithm is divided in three phases. The first phase infers and optimizes a division of the reference genome into a set of overlapping intervals (SOI) or sliding windows. An optimal SOI ensures a maximal read coverage in its intervals and maximal read diversity in its overlaps (see Supplementary Material for details). In fact, the higher the diversity in the overlaps, the higher the interval coverage, the better is the chance to reconstruct true variants (Prosperi *et al.*, 2011). For each interval (whose length is designed such that a read can span it entirely), a set of distinct variants and the corresponding prevalence is retained, i.e. the quasispecies is reconstructed 'locally' by using all reads spanning the interval (Beerenwinkel and Zagordi, 2011). The frequencies of those distinct variants form a multinomial distribution that is used in the 'global' quasispecies reconstruction, i.e. across all overlapping intervals. The second phase of reconstruction executes the heuristic algorithm previously described in detail by Prosperi *et al.* (2011), based on multinomial distribution matching, using a maximum-likelihood guide distribution. A set of reconstructed variants is generated as output. The third phase takes the set of reconstructed variants and clusters them using a variation of the probabilistic clustering algorithm introduced by Zagordi *et al.* (2010a). The difference from the original method is that the *a posteriori* probability inferred via the Dirichlet's process mixture prior and the Gibbs' sampling are substituted with the Bayesian information criterion and a random search.

QuRe has been implemented using the Java Development Kit Standard Edition 6 (http://www.oracle.com/technetwork/java), coupled with the open-source Java library JAligner (http://jaligner.sourceforge.net/). The software works from the command line and requires two input files: a read set file and a genome reference file, both in FASTA format. The main output is a FASTA file containing the reconstructed quasispecies, i.e. all distinct variants covering the reference genome with the associated prevalence (see the Supplementary Material and README file for optional input parameters and ancillary output files). QuRe is multithreaded and uses as many processors as are available on the computer where it is run. Multithreaded procedures include the random score estimation, read alignment, error correction, SOI optimization and final clustering. Running time for analyzing a NGS dataset of $\approx$20 000 reads (with an average read length of $\approx$450 bp, encompassing a region of $\approx$2000 bases over a reference sequence of $\approx$10 000 bases) is <10 min on an Intel® Core i7 740QM at 1.73 GHz, using $\approx$1.5 Gb of random access memory. The software is distributed under the GNU General Public License and is available at https://sourceforge.net/projects/qure/.

3 CONCLUSIONS

Ideally, NGS can outperform Sanger in quasispecies characterization, both in terms of resolution and length of sequences, but the reliability of the reconstruction method is crucial, as well as the availability of software for testing and comparing the methods in multiple contexts or scales, e.g. different viruses and/or different intra-host diversity. The core algorithm of QuRe has been shown to be superior to other approaches (Prosperi *et al.*, 2011) by reducing the chance to reconstruct false positives, i.e. *in silico* recombinants, as shown with both simulated and real NGS experiments (see also Supplementary Material).

In conclusion, QuRe combines a robust, fast read mapping and quasispecies reconstruction method with an easy-to-use, platform-independent implementation. Future improvements include the implementation of more advanced error correction methods besides the naïve Poisson model (Gilles *et al.*, 2011).

Funding: This work was supported in part by a National Intitutes of Health (NIH), National Center for Research Resources (NCRR), Clinical and Translational Science Awards (CTSA) award to the University of Florida (grant # UL1RR029890); National Institutes of Health (NIH), National Institute of Neurological Disorders and Stroke (NINDS) (grant R01 NS063897-01A2).

Conflict of Interest: none declared.

REFERENCES

Archer,J. *et al*. (2010) The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time-an ultra-deep approach. *PLoS Comput. Biol.*, **6**, e1001022.

Astrovskaya,I. *et al.* (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, **12** (Suppl. 6), S1.

Bacro,J.N. and Comet,J.P. (2000) Sequence alignment: an approximation law for the Z-value with applications to databank scanning. *Comput. Chem.*, **25**, 401–410.

Beerenwinkel,N. and Zagordi,O. (2011) Ultra-deep sequencing for the analysis of viral populations. *Curr. Opin. Virol.*; doi:10.1016/j.coviro.2011.07.008.

Eriksson,N. *et al.* (2008) Viral population estimation using pyrosequencing. *PLoS Comput. Biol.*, **4**, e1000074.

Gilles,A. *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **12**, 245.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Holmes,E.C. (2010) The RNA virus quasispecies: fact or fiction? *J. Mol. Biol.*, **400**, 271–273.

Jojic,V. *et al.* (2008) Population sequencing using short reads: HIV as a case study. *Pac. Symp. Biocomput.*, **13**, 114–125.

Kingsmore,S.F. and Saunders,C.J. (2011) Deep sequencing of patient genomes for disease diagnosis: when will it become routine? *Sci. Transl. Med.*, **3**, 87ps23.

Más,A. *et al.* (2010) Unfinished stories on viral quasispecies and Darwinian views of evolution. *J. Mol. Biol.*, **397**, 865–877.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Miller,J.R. *et al.* (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315–327.

Pareek,C.S. *et al*. (2011) Sequencing technologies and genome sequencing. *J. Appl. Genet.*, **52**, 413–435.

Prabhakaran,S. *et al*. (2010) HIV-haplotype inference using a constraint-based dirichlet process mixture model. In *Extended abstract at the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS), Machine Learning in Computational Biology (MLCB) workshop*, Whistler, BC, Canada.

Prosperi,M.C. *et al.* (2011) Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics*, **12**, 5.

Wang,C. *et al.* (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res.*, **17**, 1195–1201.

Westbrooks,K. *et al.* (2008) HCV Quasispecies Assembly using Network Flows. *Lect. Notes Comput. Sci.*, **4983**, 159–170.

Zagordi,O. *et al.* (2010a) Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J. Comput. Biol.*, **17**, 417–428.

Zagordi,O. *et al.* (2010b) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.*, **38**, 7400–7409.

Zagordi,O. *et al.* (2011) ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, **12**, 119.