*k*GEM: An Expectation Maximization Error Correction Algorithm for Next Generation Sequencing of Amplicon-based Data

Alexander Artyomenko[1*], Nicholas Mancuso[1*], Pavel Skums[2], Ion Măndoiu[3*], and Alex Zelikovsky[1*]

[1] Department of Computer Science
Georgia State University
Atlanta, Georgia 30302-3994
email: {aartyomenko, nmancuso, alexz}@cs.gsu.edu
[2] Centers for Disease Control and Prevention
Atlanta, Georgia 30333
email: kki8@cdc.gov
[3] Department of Computer Science & Engineering
University of Connecticut
Storrs, CT 06269
email: ion@engr.uconn.edu

Abstract. The ability of high-throughput sequencing to generate large quantities of reads has allowed virologists to study the structure of viral populations from an infected host in detail. RNA-based viral populations, due to point mutations and recombination events, exist as heterogeneous "swarm" and are known as *quasispecies*. Discerning rare, i.e., low-frequency, variants within the population is made difficult due to errors introduced by the sequencing technology. In this paper we address the problem of error correction for high-throughput sequencing reads generated from a viral quasispecies sample. We propose an EM-based algorithm that performs competitively with previously reported methods.

Keywords: Error correction. Viral quasispecies. High-throughput sequencing. Next-generate sequencing.

1 Introduction

Mutation and recombination during replication drive the heterogeneity of RNA-based viral populations. The rapid rate of replication throughout infection produces a highly diverse yet closely related viral variants known as quasispecies. The variability of the quasispecies structure can have effects in a host's immune escape response and cell tropism[4].

* This work has been partially supported by NSF award IIS-0916401, NSF award IIS-0916948, Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture

Next-generation sequencing offers unprecedented level of sequencing depth and coverage. By exploiting this massive availability of genetic data, the quasispecies structure may be inspected directly. While the technology is capable of generating large amounts of data, it is not without flaws. Reads (i.e., examined sequence fragments) may contain base-call errors as well as homopolymer errors. A base-call error is the result of a nucleotide being miscalled by the machine. That is, instead of the actual nucleotide being reported a substitute nucleotide was called. This includes the possibility of insertions and deletions. Homopolymer errors are the result of the sequencing machine's inability to accurately interpret the signals generated from nucleotide incorporation during synthesis. As longer identical nucleotide subsequences are synthesized by the machine, the signal (typically either light or voltage-based) increases in a non-linear fashion. This results in either over or under-estimating the abundance of a specific nucleotide. A naive error correction method may drop all low frequency reads as it may be suspected they are the end product of errors in the sequencing technology. However, there exist low-frequency viral variants in the population. Distinguishing rare variants from errors remains a difficult and challenging aspect of viral quasispecies reconstruction.

Several approaches for error correction already exist. KEC is k-mer (i.e., substrings of fixed length k) based error correction algorithm[5]. Its approach is based on analyzing frequencies and compositions of k-mers for detecting regions with errors. ShoRAH uses a mixture method approach. A non-parametric Bayesian model is used to cluster reads[8]. QuasiRecomb is another approach utilizing a "jumping" Markov model that incorporates the possibility for recombinant viral progeny[9].

The paper is organized as follows. Section 2 describes the proposed methods and model for correcting errors. Section 3 discusses the experimental results and validation of the method. Finally in section 4 we conclude the paper with possible future improvements and implementation details.

2 Methods

We refer to an *extended* reference as the final consensus from an iterative sequence alignment. The extended reference may contain insertions due to deletions in reads, which we denote as `d`. Let a genotype G_i be a matrix where each column corresponds to a position on the extended reference and each entry of the column stores frequency of the corresponding nucleotide $\{\mathtt{a}, \mathtt{c}, \mathtt{t}, \mathtt{g}, \mathtt{d}\}$.

Formally, given a set of reads R emitted by a population P and $k \in \mathbb{N}$, a *k-genotype* $G^k = G^k(P)$ of the population P is a set $G^k = \{G_1, \ldots, G_k\}$ of k distinct genotypes that most likely emitted R:

$$G^k = \underset{H^k, |H^k| \leq k}{\arg\max} \Pr(R|H^k).$$

Note that if $|P| = k$, then finding k-genotype is equivalent to reconstruction of the original haplotypes in P. Thus the problem of finding correct

viral amplicons can be reduced to the following.

Population k-Genotype Reconstruction Problem. Given a set R of single amplicon reads emitted by haplotype population P, find a k-genotype G^k for the population P.

Population k-genotype EM (kGEM). We initialize the algorithm by the following procedure. Let $A = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}, \mathtt{d}\}$ and select uniformly and independently at random n reads from R. For each sampled read s denote its mth allele by s_m and let $f_{i,m}(e)$ denote the frequency of allele e in the mth position of G_i. We assume that enumeration of alleles in s is the same as in the extended reference. Compute the initial allele frequencies given by,

$$f_{i,m}(e) = \begin{cases} 1 - 4\varepsilon & \text{if } s_m = e \\ \varepsilon & \text{otherwise,} \end{cases}$$

where $\varepsilon > 0$ is the probability of error. We denote genotype G_i in iteration t as $G_i^{(t)}$ and similarly for some frequency f_i. The algorithm then performs the following four steps in each iteration:

1. **Estimate Read Emission Probability.**
   Estimate the probability $h_{i,r}$ that the ith genotype G_i has emitted aligned read r given by,

$$h_{i,r} = \prod_{m=\mathrm{start}(r)}^{\mathrm{end}(r)} \frac{f_{i,m}(r_m)}{\sum_{e \in A} f_{i,m}^2(e)}, \tag{1}$$

   where $\mathrm{start}(r)$ and $\mathrm{end}(r)$ refer to the beginning and ending positions of read r in the extended reference.
2. **Estimate k-Genotype Frequencies via EM**[2].
   We initialize frequencies of every G_i uniformly as $\frac{1}{k}$. Each iteration τ of the EM algorithm consists of the following two steps:
   (a) *E-Step:* Compute the expected number of reads $e_{i,r}$ emitted from the ith genotype G_i that match read r given by,

$$e_{i,r} = o_r \cdot p_{i,r} \tag{2}$$

$$p_{i,r} = \frac{f_i^{(\tau)} \cdot h_{i,r}}{\sum_{i'=1}^{k} f_{i'}^{(\tau)} \cdot h_{i',r}}, \tag{3}$$

   where f_i is the frequency of G_i and o_r is the observed frequency of r.
   (b) *M-Step:* Estimate the frequency $f_i^{(\tau+1)}$ of each $G_i^{(t)}$ as the portion of all reads emitted by $G_i^{(t)}$ as follows:

$$f_i^{(\tau+1)} = \frac{\sum_{r \in R} e_{i,r}}{\sum_{i'=1}^{k} \sum_{r \in R} e_{i',r}} \tag{4}$$

   Repeat steps (2)-(4) until the squared deviation

$$\sum_{i=1}^{k} (f_i^{(\tau)} - f_i^{(\tau+1)})^2$$

   falls below the pre-specified accuracy $\delta > 0$ for each $i = 1, \ldots, k$.

3. **Estimate Allele Frequencies.**
   Compute the normalized frequency $f_{i,m}(e)$ of each allele $e \in A$ in the mth position of $G_i^{(t+1)}$ as follows:

$$f_{i,m}(e) = \frac{\sum_{r \in R: r_m = e} p_{i,r}}{\sum_{r \in R: begin(r) \leq m \leq end(r)} p_{i,r}} \tag{5}$$

4. **Round Allele Frequencies.**
   Round allele frequencies according to the following rules:

$$f_{i,m}(e) = \begin{cases} 1 - 4\varepsilon & \text{if } e = \arg\max_{e' \in A} f_{i,m}(e') \\ \varepsilon & \text{otherwise} \end{cases} \tag{6}$$

Repeat steps (1)-(6) until the distance $||G_i^{(t)}, G_i^{(t+1)}||$ falls below the pre-specified accuracy $\delta > 0$ for each $i = 1, \ldots, k$. Collapse duplicates and drop rare genotypes (i.e., frequency f_i below specified threshold) upon completion. If the total number of genotypes in the population has changed repeat the entire procedure; otherwise, report the current set G^k.

3 Results

Using a sample of 44 HCV clones from [7], 20 simulated data sets were generated with Grinder version 0.5[1]. Each dataset consisted of 100,000 total reads from a random sample of 10 variants and was categorized by its error model and generated population distribution. All datasets contained mutation-based errors (i.e., substitution, insertion, and deletion) which were distributed uniformly throughout a given read at a rate of 0.1 percent. In addition, 10 datasets contained homopolymer errors distributed according to the model of Balzer et al[3]. The population distribution adhered to either a uniform or power-law model with parameter $\alpha = 2.0$. kGEM was compared against KEC and QuasiRecomb using sensitivity and positive predicted value (PPV) as a measure of the quality of the error-corrected data sets (Figure 1).[4] Reads were aligned using the tool InDelFixer[6]. Results shown are the mean and standard error over 5 datasets of the same "configuration". kGEM outperforms QuasiRecomb in sensitivity in all 5 datasets. Further, kGEM has comparable PPV for the homopolymer-inclusive datasets, and higher PPV for the non-homopolymer datasets. KEC was excluded since its clustering stage would not finish in a reasonable amount of time given the parameters. When re-run without clustering, KEC resulted in 100% sensitivity but extremely low PPV (e.g., 0.07%) and was dropped from consideration.

[4] Parameters for methods:
indelfixer -i data_set -g consensus -refine 3
kec -k 25 -i 5 data_set
quasirecomb -K 1:7 -global -i data_set
quasirecomb -K 7 -global -i data_set -refine
kgem -k 50 -tr 5 data_set

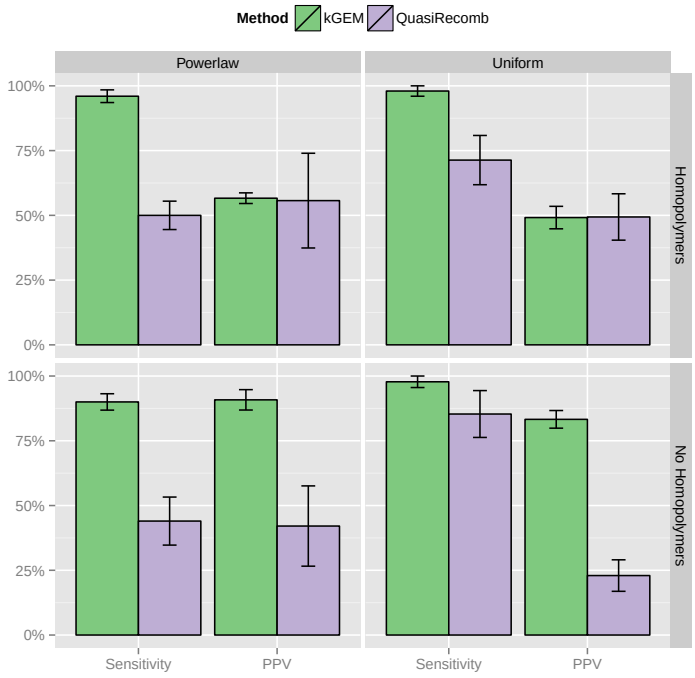

Fig. 1. Sensitivity and PPV for the results on simulated data sets

4 Conclusion

In this paper we propose a new expectation maximization-based method for error correction of amplicon NGS reads that performs reliably and quickly. We compared our method with existing analogs such as KEC and QuasiRecomb. Test results show kGEM is better in sensitivity and positive predicted value than QuasiRecomb. Possible future work could incorporate read quality scores.

References

1. Florent E. Angly, Dana Willner, Forest Rohwer, Philip Hugenholtz, and Gene W. Tyson. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, 2012.
2. I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I.I. Mandoiu, P. Balfe, and A. Zelikovsky. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, 12(Suppl 6):S1, 2011.

3. Susanne Balzer, Ketil Malde, and Inge Jonassen. Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics*, 27(13):i304–i309, 2011.
4. Aldo Manzin, Laura Solforosi, Enzo Petrelli, Giampiero Macarri, Grazia Tosone, Marcello Piazza, and Massimo Clementi. Evolution of hypervariable region 1 of hepatitis c virus in primary infection. *Journal of Virology*, 72(7):6271–6276, 1998.
5. P. Skums, Z. Dimitrova, D.S. Campo, G. Vaughan, L. Rossi, J.C. Forbi, J. Yokosawa, A. Zelikovsky, and Y. Khudyakov. Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics*, 13(Suppl 10):S6, 2012.
6. Armin Töpfer. Indelfixer. http://www.bsse.ethz.ch/cbg/software/InDelFixer.
7. T von Hahn, JC Yoon, H Alter, CM Rice, B Rehermann, P Balfe, and JA McKeating. Hepatitis c virus continuously escapes from neutralizing antibody and t-cell responses during chronic infection in vivo. *Gastroenterology*, 132:667–678, 2007.
8. Osvaldo Zagordi, Arnab Bhattacharya, Nicholas Eriksson, and Niko Beerenwinkel. Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, 12(1):119, 2011.
9. Osvaldo Zagordi, Armin Töpfer, Sandhya Prabhakaran, Volker Roth, Eran Halperin, and Niko Beerenwinkel. Probabilistic inference of viral quasispecies subject to recombination. In *Proceedings of the 16th Annual international conference on Research in Computational Molecular Biology*, RECOMB'12, pages 342–354, Berlin, Heidelberg, 2012. Springer-Verlag.