Efficient Error Correction for Deep Sequencing of Viral Amplicons

Pavel Skums[1] , Zoya Dimitrova[1], David Campo[1], Gilberto Vaughan[1], Livia Rossi[1], Joseph Forbi[1], Jonny Yokosawa[2], Alex Zelikovsky[3] and Yury Khudyakov[1]

[1] Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta, USA {kki8, izd7, fyv6, jiv9, fld5,gzf7,yek0}@cdc.gov. [2]Universidade Federal de Uberlândia, Brazil, jyokosawa@icbim.ufu.br. [3] Georgia State University, Atlanta, GA {AlexZ@cs.gsu.edu}

Abstract. We present two new highly efficient error correction algorithms: (i) k-mer - based error correction (KEC); and (ii) empirical frequency threshold (ET). Both were compared to the recently published algorithm SHORAH to evaluate the relative performance using 24 experimental datasets obtained by 454-sequencing of amplicons with known sequences. We found that all three algorithms showed similar performance in terms of finding true sequences, but KEC and ET methods significantly outperformed SHORAH both in terms of their ability to remove false sequences and to estimate the frequency of true ones.

Keywords: HCV, quasispecies, pyrosequencing, error correction

1 Introduction

Hepatitis C virus (HCV) shows a very high level of sequence heterogeneity, which is responsible for its escape from neutralizing host immune responses and rapid development of drug resistance. Recent advances in high-throughput (HT) sequencing methods allow for analysis of the unprecedented number of HCV-genomic sequence variants from infected patients and present a novel opportunity for understanding HCV evolution, drug resistance and immune escape. However, owing to the massive scale of sequencing, sequence errors generated during HT sequencing require extensive computational processing with error correction algorithms in order to obtain high quality reads for genetic analysis. The key purpose of such algorithms is to discriminate between artifacts and actual sequences. This task becomes especially challenging for recognizing and preserving low-frequency natural variants in viral population.

SHORAH [5][6] is currently one of the best error correction algorithms available. It uses probabilistic clustering approach based on the Dirichlet process mixture. Another approach to error correction is based on the use of k-mers, or substrings of reads of a fixed length k [2][3][4]. These algorithms have good performance but high time- and memory-consumption needs, together with the possibility of errors introduced during the correction phase [5]. To overcome these disadvantages, the authors of EDAR algorithm [1] developed an approach for the detection and deletion of sequence regions containing errors. This error deletion works well for shotgun experiments, but is unacceptable for the small amplicon reads commonly analyzed in viral samples.

In this paper, we present two new efficient error correction algorithms: (i) k-mer-based error correction (KEC); and (ii) empirical frequency threshold (ET). KEC uses the EDAR algorithm optimized for amplicon sequencing for the detection of error regions and a novel algorithm for correction of errors associated with homopolymers. KEC does not require a reference sequence and is, therefore, suitable for *de-novo* sequencing. The ET algorithm uses estimation of a frequency threshold for indels and haplotypes calculated from experimentally obtained clonal sequences, also correcting homopolymers. Both algorithms were compared to SHORAH to evaluate their relative performance using 25 experimental amplicon datasets with known sequences obtained using 454 sequencing.

2 Algorithms description

2.1. KEC algorithm

The scheme of KEC includes 4 steps: (1) Calculate k-mers and their frequencies (k-counts). We assume that k-mers with high k-counts (“solid” k-mers) are correct, while k-mers with low k-counts (“weak” k-mers) contain errors. (2) Determine the threshold k-count (error threshold), which distinguishes solid k-mers from weak k-mers. (3) Find error regions. The error region is the segment [i,j] of read such that for every $p \in [i,j]$ the k-mer starting at the position p is considered weak. (4) Correct the errors in error regions.

Methods proposed in EDAR were used for steps 1 and 3. However, they were optimized using efficient data structures based on hash maps. The error threshold estimation from [1] is not applicable to the amplicon data. It was replaced by an algorithm based on the detection of local minima in smoothed distributions. We call error region $x=[b,e]$ of a read r a tail, if either $b = 1$ or $e = n-k+1$ (n is the length of r). Let $l(x)$ be the length of x, and $h_i(w)$ be a homopolymer of length i composed of nucleotide $w \in \{A,T,G,C\}$.

Claim 1. *Suppose, that the non-tail error region x was caused by a one-nucleotide error E. Let w be the last nucleotide of x. If E is a replacement, then $l(x) = k$. If E is an insertion in the homopolymer of length r ($0 \leq r \leq k$), then $l(x)= k-r+1$, x is followed by a homopolymer $h_{l-1}(w)$. If E is a deletion in the homopolymer of length m, then $l(x) = k-m-1$ and if $m \geq 1$, then x is followed by a homopolymer $h_m(c)$, where $c \neq w$.*

Errors were identified and corrected in non-tail error regions using Claim 1, and then the corresponding prefixes or suffixes were deleted from reads for tails. The procedure was repeated until the dataset had no errors or the specified number of iterations was reached. Claim 1 considers only error regions with $l(x) \leq k$. The longer error regions correspond to the occurrence of >1 errors separated by ≤ k nucleotides. We found that this type of error is much less frequent and we correct it by a heuristics based on Claim 1.

2.2. ET algorithm

The key idea of the procedure is to calculate the frequency of erroneous sequences in amplicon samples where only a single sequence was expected. Each single-clone sample was processed in the following way: First, each sequence is aligned against a set of external references of all known genotypes. For each sequence the best match of the external set is chosen. The aligned sequence is clipped to the size of the chosen external reference. The 20 most frequent sequences that do not create insertions or deletions are selected, constituting the internal reference set. Each sequence is aligned against each member of the internal references set and its best match is chosen.

The frequency of erroneous indels and its standard deviation (s.d.) was calculated over all nucleotide positions for 15 single-clone samples. An indel threshold was defined as the average frequency of erroneous indels + 5 s.d. If a sequence contained an indel with a frequency lower than the threshold, the sequence was removed. Then all homopolymers of at least 4 nucleotides were identified, followed by removal of the insertions and replacement of the deletions by the repeated nucleotide. The frequency of erroneous sequences and its s.d. were calculated over the 15 single-clone samples. A sequence threshold was defined as the average frequency of erroneous sequences + 5 s.d. All sequences with a frequency lower than the threshold were removed. This procedure was applied to each mixture sample.

3 Algorithms comparison

Individual plasmid clones (n=10) containing different HCV hypervariable region 1 sequences were purified and sequenced using dye-terminator sequencing. A set of plasmid samples was generated. 14 samples contained a single clone. 10 samples contained 8 clones mixed together in different proportions (from 1% to 93%). The E1/E2 region (309 nt) was amplified from each sample and sequenced using GS FLX Titanium Series Amplicon kits. Low quality reads were removed using the GS Run Processor (Roche, 2010). Each sequence file was then analyzed using ET, KEC or SHORAH error correction algorithms. SHORAH was applied several times under different parameters and the best attained results are reported here. All results are summarized in Table.

Table. Test results of the single-clone (S) and mixture (M; n=8) samples. MT: Missing true sequences; FS: False sequences; MSE: root mean square error; HD: Average Hamming distance, averaged over all false sequences.

	ET				KEC				SHORAH			
	MT	FS	MSE	HD	MT	FS	MSE	HD	MT	FS	MSE	HD
S1	0	0	0.00	0	0	1	4.67	1	0	351	29.02	4.84
S2	0	0	0.00	0	0	0	0.00	0	0	269	30.12	4.44
S3	0	1	1.09	1	0	2	4.93	1.5	0	292	23.44	5.31
S4	0	1	0.98	2	0	1	2.84	1	0	271	44.68	5.39
S5	0	0	0.00	0	0	0	0.00	0	0	319	9.63	4.47

S6	0	1	5.26	2	0	1	6.10	1	0	194	18.70	3.94
S7	0	0	0.00	0	0	1	5.80	1	0	496	21.52	6.70
S8	0	0	0.00	0	0	0	0.00	0	0	262	14.37	4.58
S9	0	0	0.00	0	0	0	0.00	0	0	183	6.23	6.97
S10	0	0	0.00	0	0	0	0.00	0	0	288	7.77	5.11
S11	0	1	0.53	2	0	0	0.00	0	0	717	24.71	5.03
S12	0	0	0.00	0	0	0	0.00	0	0	611	25.94	5.52
S13	0	0	0.00	0	0	0	0.00	0	0	156	5.53	4.93
S14	0	0	0.00	0	0	0	0.00	0	0	161	6.83	6.60
Mean	0.00	0.29	0.56	0.50	0.00	0.43	1.74	0.39	0.00	326.43	19.18	5.27
M1	0	0	1.17	0	0	1	0.87	1	0	320	1.23	4.51
M2	0	0	1.50	0	0	0	1.75	0	0	738	3.70	4.44
M3	0	0	2.92	0	0	0	3.55	0	0	638	3.65	4.25
M4	0	0	2.18	0	0	0	2.30	0	0	577	2.88	5.20
M5	0	0	0.34	0	7	0	7.00	0	0	214	0.91	7.37
M6	1	0	2.20	0	1	0	1.97	0	1	394	2.48	4.54
M7	0	0	1.20	0	0	0	1.97	0	0	499	2.04	5.00
M8	1	0	0.89	0	1	0	2.31	0	1	336	3.09	5.54
M9	0	0	2.23	0	6	0	9.25	0	0	643	6.56	4.49
M10	1	0	3.53	0	1	0	4.21	0	2	637	5.88	5.32
Mean	0.30	0.00	1.82	0.00	1.60	0.10	3.52	0.10	0.40	499.60	3.24	5.07

All methods found the correct sequence in each single-clone sample. However, ET and KEC retained the lower number of false sequences. Similarly, ET and KEC showed lower number of false sequences than SHORAH in each mixed samples. All three algorithms were successful in identifying most of true sequences, with ET being the most accurate. KEC did not detect true sequences representing ~1% in mixtures M5 and M9. The low root mean square error between observed and expected frequencies of true sequences indicates a high accuracy of ET and KEC, whereas SHORAH has much higher MSE, owing to the detection of a greater number of false sequences. Analysis of the Hamming distance between false sequences and their closest match shows that false sequences retained by KEC and ET are genetically closer to true sequences than sequences retained by SHORAH.

4 Conclusions

SHORAH, ET and KEC perform equally efficient in finding true sequences. However, KEC and ET outperform SHORAH in removing false sequences and estimating the sequence frequency. At the same time, in contrast to SHORAH and ET, KEC does not require a reference sequence. Both algorithms, KEC and ET, are highly suitable for rapid recovery of high quality sequences from reads obtained by deep sequencing of genomic regions from heterogeneous viruses such as HCV and HIV.

References

1. Zhao, X., Palmer, L., Bolanos, R., Mircean, C., Fasulo, D., Wittenberg, D.: EDAR: An efficient error detection and removal algorithm for next generation sequencing data. Journal of computational biology, 17(11), 1549 — 1560 (2010)
2. Chaisson, M.J., Brinza, D., Pevzner, P.A.: De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res. 19, 336–346 (2009).
3. Chaisson, M.J., Pevzner, P.A.: Short read fragment assembly of bacterial genomes. Genome Res. 18, 324–330 (2008).
4. Pevzner, P., Tang, H., Waterman M.: An Eulerian path approach to DNA fragment assembly: Proc. Natl. Acad. Sci. USA, 9748—9753 (2001).
5. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N.: Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. Journal of Computational Biology, 17, 417—428 (2009).
6. Zagordi O, Klein R, Däumer M, Beerenwinkel N.: Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Research, 38 (21), 7400—7409 (2010).