Reconstruction of Infectious Bronchitis Virus Quasispecies from NGS Data

Bassam Tork[1], Ekaterina Nenastyeva[1], Alexander Artyomenko[1], Nicholas Mancuso[1], Mazhar I.Khan[3], Rachel O'Neill[4], Ion Mandoiu[2], and Alex Zelikovsky[1]

[1] Department of Computer Science, Georgia State University, Atlanta, GA 30303 Email: {btork,enenastyeva1,aartyomenko,nmancuso,alexz}@cs.gsu.edu
[2] Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269 Email : ion@engr.uconn.edu
[3] Department of Pathobiolgy and Veterinary Science, University of Connecticut, Storrs, CT 06269 Email : mazhar.khan@uconn.edu
[4] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269 Email : rachel.oneill@uconn.edu

Abstract.
Motivation: Poultry farms are susceptible to viral infections which cause significant economic losses worldwide in terms of impaired growth, reduced egg production and quality, and even mortality. In the United States where infections with virulent strains of Newcastle disease and highly pathogenic avian influenza are not common, the infectious bronchitis virus (IBV) is the biggest single cause of economic loss. Viral quasispecies sequences reconstruction by analyzing high-throughput sequencing data will contribute to understanding the roles and interactions of host animal (chicken) and viral genomes for improving animal health, well-being, and production efficiency which is one of the goals included in Blueprint for USDA Efforts in Agricultural Animal Genomics(2008-2017).
Methods: We propose a computational pipeline for quasispecies (closely related variants to ancestral genome) reconstruction consisting of 3 phases: (1) Read Error Correction (2) Read Alignment and (3) Reconstruction of Viral Quasispecies. We vary different parameter values of the pipeline, i.e parameter tuning, to get better results.
Results: Our experiments show that varying the parameter settings: (1) number of mismatches between superreads and subreads, (2) number of mismatches in the overlap between two superreads, (3) mutation rate, to reconstruct IBV quasispecies. We get better results in terms of Average Distance to Clones(ADC), and Average Prediction Error(APE).

Keywords: viral quasispecies, high-throughput sequencing, next-generation sequencing.

Introduction. Viral infections cause a significant burden on animal health, reducing yields and increasing production costs due to expensive control programs. Vaccination is a vital part of control programs; however, its effectiveness

is reduced by the quick evolution of escape viral quasispecies in animal hosts. Existing techniques for studying quasispecies evolution and response to vaccines have severely limited sensitivity and often require prior knowledge of sequence polymorphisms. By generating millions of short reads per run with no need for culture or cloning, next-generation sequencing (NGS) technologies enable comprehensive identification of viral quasispecies infecting an animal. However, analysis of NGS data is challenging due to the huge amount of data on one hand, and to the short read lengths and high error rates on another.

Methods. Many tools developed for Sanger reads do not work at all or have impractical runtimes when applied to NGS data. Even newly developed algorithms for de novo genome assembly from NGS data are tuned for reconstruction of haploid genomes, and work poorly when the sequenced sample contains a large number of closely related sequences, as is the case in viral quasispecies. To address these shortcomings we introduce a computational pipeline for accurate reconstruction of viral quasispecies sequences and estimate their frequencies from HTS data where it incorporates different NGS reads error correction methods, aligners, and genome assemblers (ViSpA [1], and ShoRAH [2, 3]), using different tuning parameters. We apply experiments on IBV 454 shotgun reads, collected from commercial poultry farms. For method validation we use IBV sanger clones (as ground truth).

The proposed computational pipeline for quasispecies reconstruction, consists of the following stages:

1. **Read Error Correction**. 454 Life Sciences can erroneously sequence one base pair per 1000 [4]. The error rate is strongly related to the presence and size of homopolymers [5], i.e., genome regions consisting of consecutive repetition of a single base (for example, TTTTT). We use KEC, SAET, and ShoRAH programs to do error correction prior to assembly which involve clustering of reads. ShoRAH clusters the reads in Bayesian fashion using the Dirichlet process mixture [2]. KEC clusters reads based on kmers [6], while Saet uses reads quality scores for error correction.
2. **Read Alignment**. In this step, we use an independent alignment program to map reads against a reference viral sequence[7]. This aligner can be easily replaced with another one.
3. **Reconstruction of Viral Quasispecies**. In this step, we use two assembly programs ViSpA [1] and ShoRAH [2] to reconstruct quasispecies from aligned reads and estimate their relative frequecies.

VispA [1] executes the following steps and outputs the quasispecies spectrum(i.e. quasispecies sequences and their relative frequencies):

- **Preprocess Aligned Reads**. ViSpA uses placeholders I and D for aligned reads containing insertions and deletions, in this process it do a simplistic error correction. Deletions supported by a single read are replaced either with the allele present in all the other reads in the same position if they are the same, or with N(unknown base pair), and removes insertions supported by a single read.

- **Construct the Read Graph**. In the read graph each vertex corresponds to a read and each directed edge connects two overlapping reads. ViSpA differentiates between two types of reads, super-read and sub-read which is a substring of the super-read. The read graph consists only of super-reads.
- **Assemble Candidate Quasispecies Sequences**. Each candidate quasispecies corresponds to a path in the read graph. ViSpA uses what is so called max-bandwidth paths for assembly.
- **Estimate Frequency of Haplotype Sequences**. In this step, ViSpA uses Expectation Maximaization algorithm to estimate the frequency of each reconstructed sequence using both super-reads and sub-reads.

ShoRAH [2, 3] executes the following steps and outputs the quasispecies spectrum:

- **Align Reads**. The first step for ShoRAH is producing a Multiple Sequence Alignment (MSA) of reads, it use its own aligner to align all reads to the reference and from the set of pairwise alignments it builds a MSA.
- **Correct Reads from Genotyping Errors (Local Haplotype Reconstruction)**. While ViSpA uses independent error correction programs, ShoRAH uses its own error correction method. Sequencing errors are corrected by a Bayesian inference algorithm which estimates the quality of the reconstruction, although only the maximum likelihood estimate is passed on to subsequent steps. ShoRAH implements a specific probabilistic clustering method based on the Dirichlet process mixture for correcting technical errors in deep sequencing reads and for highlighting the biological variation in a genetically heterogeneous sample.
- **Reconstruct Global Haplotype**. This step is similar to assembly of candidate quasispecies sequences in ViSpA.
- **Estimate Frequency**. In this step, ShoRAH estimates the frequency of each candidate sequence.

Compared Methods. We vary different parameter values, what we call parameter tuning. The following tuning parameters are used for ViSpA quasispecies reconstruction:

- n : number of mismatches between superreads and subreads,
- m : number of mismatches in the overlap between two superreads,
- t : mutation rate.

For ShoRAH, we use the default parameters for quasispecies reconstruction.

Gold Standards. Reads samples were collected from infected chickens, and quasispecies variants (1610 base pair length) were sequenced using life sciences 454 shotgun sequencing, followed by sanger sequencing of individual variants. We got 10 sanger clones (c1,...,c10) of average length 546 base pairs recovering a fraction of the full spectrum of quasispecies variants. These clones are considered

as the gold standards or the ground truth for parameter calibration and comparison with different methods. We measured the pairwise edit distance between all clones, clone(c8) have a large edit distance with others (between 40 and 45) and is considered as an outlier.

Comparison Measures. Before defining methods validation, we need to define the following parameters:

Symbol Description

c_i:	sanger clone i, $1 \leq i \leq 10$;
q_j :	reconstructed quasispecies j ;
f_{c_i}:	frequency of sanger clone i;
f_{q_j}:	frequency of reconstructed quasispecies j;
m_i:	how far is clone i from closest reconstructed quasispecies;
m_j:	how far is reconstructed quasispecies j from closest clone ;

m_i and m_j are defined as follows:

$$m_i = \min_j(c_i, q_j)$$

$$m_j = \min_i(c_i, q_j)$$

To validate different methods, we use the following two measures:

- **Average Distance to Clones(ADC)**=$\sum_{c_i} m_i \cdot f_{c_i}$
- **Average Prediction Error (APE)**=$\sum_{q_j} m_j \cdot f_{q_j}$

ADC and APE are respectively analogous to sensitivity, and positive predictive value (PPV). But they are different in sense that ADC and APE have better quality whenever they are close to 0, while sensitivity and ppv have better quality when they are close to 1. We disregard one of the clones ($c8$) in ADC calculation, since it is an outlier.

Results and Discussion. In our experiments we use different setting values to get the best values of ADC and APE for different methods (Figure 1).

We say that $method_A$ dominates $method_B$ if both ADC & APE values of $method_A$ are less than or equal to the corresponding ADC & APE values of $method_B$. By looking on Figure 1, we see that methods v125KEC(v:vispa assembler, 1:n, 2:m, 5:t, KEC:correction method), v2210KEC, and v120SAET dominate all other methods, i.e. have the best values in terms of ADC & APE. Our results suggest that using different methods with different parameter calibration and parameter settings can improve the solution and predictive power of quasispecies inference problem in terms of recall, precision, and frequency.

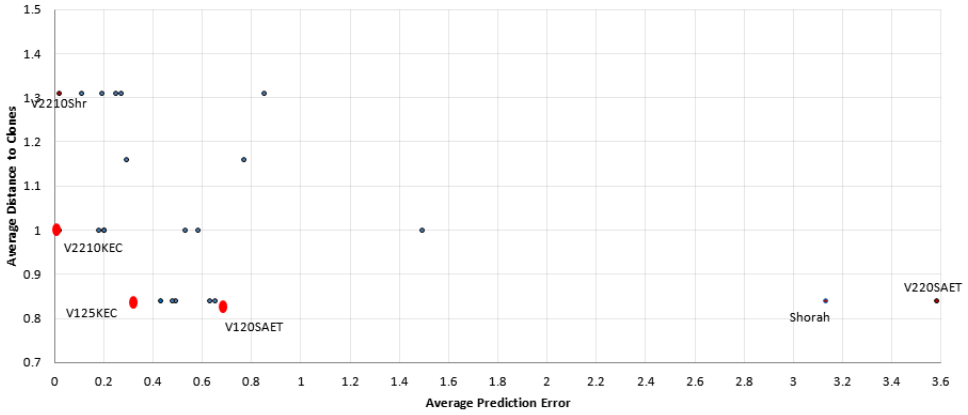

Fig. 1. Evaluation Diagram for Average Prediction Error and Average Dintance to Clones.

Acknowledgments. 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 GSU Molecular Basis of Disease Fellowship.

References

1. I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Mndoiu, P. Balfe, and A. Zelikovsky, "Inferring viral quasispecies spectra from 454 pyrosequencing reads," *BMC Bioinformatics*, vol. 12, no. Suppl 6, p. S1, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/S6/S1
2. O. Zagordi, L. Geyrhofer, V. Roth, and N. Beerenwinkel, "Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 17, no. 3, pp. 417–428, March 2010. [Online]. Available: http://dx.doi.org/10.1089/cmb.2009.0164
3. O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel, "Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data," *BMC Bioinformatics*, vol. 12, p. 119, 2011.
4. "454 life sciences, *http://www.454.com*."
5. W. Brockman, P. Alvarez, S. Young, M. Garber, G. Giannoukos, W. L. Lee, C. Russ, E. S. Lander, C. Nusbaum, and D. B. Jaffe, "Quality scores and snp detection in sequencing-by-synthesis systems," *Genome Res.*, pp. gr.070 227.107+, January 2008. [Online]. Available: http://dx.doi.org/10.1101/gr.070227.107
6. P. Skums, Z. Dimitrova, D. Campo, G. Vaughan, L. Rossi, J. Forbi, J. Yokosawa, A. Zelikovsky, and Y. Khudyakov, "Efficient error correction for next-generation sequencing of viral amplicons." *BMC Bioinformatics*, vol. 13 Suppl 10, 2012.

7. “Mosaik aligner,” http://bioinformatics.bc.edu/marthlab/Mosaik.